生成 AI の創造性と推論力の進化：OpenAI o1、o3 および Google Gemini 2.5 Pro の包括的分析

Gemini Deep Research

要約

本レポートは、OpenAI の o1 および o3 シリーズ、ならびに Google の Gemini 2.5 Pro という最新の推論型生成 AI モデルの創造性と推論力について、研究的観点と実用的観点の両方から詳細な分析を提供する。これらのモデルは、従来の単純なパターンマッチングを超え、より深い「思考」プロセスを導入することで、複雑な問題解決やニュアンスのある創造的タスクへの対応能力を飛躍的に向上させている。

OpenAI の o シリーズ、特に o1 と o3 は、「思考時間を長くする」という新しいパラダイムを導入し、特に STEM（科学・技術・工学・数学）分野での高度な推論能力を追求している[1]。o1 は初期のプレビュー版で高い期待を集めたものの、正式リリース版では性能の一貫性やコスト面での課題が指摘された[3]。後継の o3 は、ツール連携やマルチモーダルな「画像による思考」といった機能を強化し、多くのベンチマークで最高水準の性能を示したが、特にコーディングタスクにおける実用的な信頼性や幻覚（ハルシネーション）の頻度に関して、ユーザーから賛否両論の声が上がっている[2]。

一方、Google の Gemini 2.5 Pro は、100 万トークンという広大なコンテキストウィンドウ、人間による評価での高い選好度、「Deep Think」と呼ばれる高度な推論モードを特徴とし、特に反復的なコーディング作業や創造的な文章作成において実用的な強みを発揮している[7]。しかし、API の安定性や、一部の非常に複雑な学術的問いに対する応答の単純さ、安全フィルターによる創造的表現の制約といった課題も報告されている[10]。

学術的な評価においては、AIME（米国数学招待試験）や GPQA（大学院レベル科学問題）などの標準化されたベンチマークが重要な指標となるが[13]、これらのスコアと実際の多様な応用場面での実用性との間にはしばしば乖離が見られる。ベンチマークの飽和や汚染、特定スキルへの偏りは、モデルの真の能力を完全には反映しない可能性がある[15]。

創造性の評価はさらに複雑であり、TTCT（トーランス創造的思考検査）の適応や、新規性、有効性、意外性といった多面的な基準が用いられるが、LLM が生み出す「独創性」が真に新しいものか、あるいは高度な訓練データの再構成なのかという根本的な問いは残る[17]。

結論として、これらの最先端 AI モデルは推論と創造性の領域で目覚ましい進歩を遂げているが、その能力を安定的かつ倫理的に実社会へ実装するには、幻覚の抑制、透明性の向上、コスト効率の改善など、多くの課題が残されている。今後の発展は、マルチモーダル性、エージェント能力の深化、そして人間とのより自然な協調関係の構築に向けられると予想される。

I. はじめに：進化する生成 AI の推論力と創造性

A. 高度推論型生成 AI の黎明期

近年の生成 AI 技術の発展は目覚ましく、初期の大規模言語モデル（LLM）から、より洗練された推論能力、複雑な問題解決能力、そしてニュアンスに富んだ創造的生成能力を持つ現世代へと急速な進化を遂げている。この進化の核心には、単なるパターンマッチングや統計的予測を超え、モデルがより深い「思考」プロセスを経て応答を生成するという設計思想の転換がある[1]。この動きは、AI がより人間的な認知能力に近づく上での重要な一歩と見なすことができる。特に、応答生成前に追加の計算時間を費やすことで、より複雑な論理的連鎖（Chain of Thought: CoT）を内部的に構築し、その結果として質の高い出力を目指すアプローチが顕著になっている。

B. 主要モデルの紹介

本レポートで分析対象とするのは、この新しい潮流を代表する最先端の推論型生成 AI モデルである。

- **OpenAI の o シリーズ (o1, o3):** OpenAI が開発した o シリーズは、従来の GPT アーキテクチャを基盤としつつも、特に推論能力の強化に特化したモデル群である。「思考時間」の確保と CoT プロセスの精緻化に重点を置いて設計されており、一連の「推論モデル」の系譜を形成している[1]。o1 がその先駆けとなり、o3 はその能力をさらに発展させた後継モデルとして位置づけられる。
- **Google の Gemini 2.5 Pro:** Google のフラッグシップモデルである Gemini 2.5 Pro は、そのマルチモーダルな処理能力、100 万トークンという広大なコンテキストウィンドウ、そして「Deep Think」と呼ばれる複雑な推論を目的とした機能を特徴としている[7]。これらの特徴により、多様な入力形式を統合的に理解し、長大な文脈における深い洞察や創造的な出力を生成することが期待されている。

これらのモデルは、AI 研究開発の最前線に位置し、その能力と限界を理解することは、今後の技術動向を占う上で不可欠である。

C. レポートの目的と構成

本レポートの目的は、OpenAI の o1 および o3 シリーズ、ならびに Google の Gemini 2.5 Pro について、その推論力と創造性を学術的観点と実用的観点の両面から批判的に評価し、比較分析することである。具体的には、各モデルのアーキテクチャ上の特徴、主要なベンチマークにおける性能、実際の応用事例、ユーザーによる評価、そして現時点での限界点などを網羅的に調査する。

レポートの構成は以下の通りである。まず、AI の推論力と創造性を評価するための方法論と主要なベンチマークについて概観する。次に、OpenAI の o シリーズ（o1、o3）と Google の Gemini 2.5 Pro それぞれについて、その能力を詳細に分析する。その後、これらのモデル間の比較分析を行い、それぞれの長所と短所を明らかにする。さらに、これらの AI が実世界の様々な分野でどのように応用され、どのような影響を与えつつあるかを探る。最後に、現行モデルが抱える課題や限界点を整理し、今後の技術的展望と人間社会への示唆について考察する。

この分析を通じて、現行の最先端 AI モデルが達成した到達点と、今後の発展に向けた課題を明確にすることを試みる。特に、主要な AI 開発機関が、モデルサイズや訓練データ量の単純なスケールアップだけでなく、推論メカニズムそのものの質的向上に戦略的に注力している点は注目に値する。o1 が「推論」シリーズの第一弾として「思考」に時間を費やすと明示されたこと[1]や、Mira Murati 氏がこれを従来のモデルスケーリングパラダイムと対比させたこと[1]、Google が Gemini 2.5 Pro に「強化された推論モード」として「Deep Think」を導入したこと[8]などは、この戦略的転換を示唆している。これは、次世代の AI 能力、特に複雑なタスク処理能力の向上において、単にモデルを大規模化するだけでは不十分であるという認識が共有されつつあることを物語っている。

II. AI の推論力と創造性の評価：方法論とベンチマーク

AI の推論力と創造性を客観的に評価することは、その能力を理解し、さらなる発展を促す上で極めて重要である。しかし、これらの高度な認知能力を測定するための万能な方法論は未だ確立されておらず、様々なアプローチが試みられている。

A. AI の推論力の定義と測定

1. 学術的視点

AI における推論力は、単なるパターン認識や情報検索を超え、複数の情報を統合し、論理的な帰結を導き出したり、未知の状況に対して適応的な判断を下したりする能力を指す。学術的には、記号論理に基づく演繹・帰納推論、類推、因果関係の理解、複数ステップにわたる問題解決、戦略的思考などが含まれる。近年の大規模言語モデル

（LLM）においては、プロンプトに対して応答を生成する際に、内部的に一連の思考ステップ（Chain of Thought: CoT）を明示的または暗示的に経由する能力が、推論力の一つの指標として注目されている[1]。CoT は、モデルがどのように結論に至ったかの透明性を高める可能性を持つと同時に、より複雑な推論タスクの遂行を可能にするメカニズムとして研究されている。

2. 主要な推論ベンチマーク

AI の推論力を定量的に評価するために、様々なベンチマークが開発・利用されている。

- **AIME (American Invitational Mathematics Examination):** 高度な数学的推論力と問題解決能力を測定するために設計された試験であり、LLM の論理的思考力や数理的洞察力を評価する上で重要な指標となる[1]。
- **GPQA (Graduate-Level Physics, Chemistry, Biology Questions):** 物理学、化学、生物学の大学院レベルの専門知識と、それに基づく科学的推論能力を評価する。専門性の高い領域での AI の理解度と応用力を測る上で有用である[1]。
- **MMLU (Massive Multitask Language Understanding):** 人文科学、社会科学、STEM（科学・技術・工学・数学）など 57 の多様な分野にわたる知識と推論能力を総合的に測定するベンチマークである[13]。幅広い知識とそれを活用する能力を評価する。
- **コーディングベンチマーク (Codeforces, SWE-bench, HumanEval など):** プログラミング、アルゴリズム設計、コード生成・デバッグといった文脈における論理的推論能力を評価する。これらのベンチマークは、AI が具体的な問題を解決するための論理構造を構築し、それを実行可能なコードに変換する能力を測る[1]。
- **MMMU (Multimodal Multi-task Multilingual Understanding):** テキスト情報だけでなく、画像などの視覚情報を統合して推論する能力を評価する。現実世界の多くの問題はマルチモーダルな情報を伴うため、この種のベンチマークの重要性が高まっている[2]。

B. AI の創造性の定義と測定

1. 学術的視点

AI における創造性とは、一般的に、新規性（novelty）、有効性（effectiveness）または有用性（usefulness）、そして意外性（surprise）を兼ね備えたアイデアや成果物を生み出す能力として定義されることが多い[18]。人間の場合、創造性は既存の知識や経験を基にしながらも、それらを予期せぬ形で組み合わせたり、新たな視点から解釈したりするプロセスを含む。

LLM の創造性を評価する際には、トーランス創造的思考検査（TTCT）のような既存の心理学的測定法を AI 向けに改変する試みが見られる。TTCT では、流暢性（アイデアの数）、柔軟性（アイデアのカテゴリの多様性）、独創性（アイデアのユニークさ）、精巧性（アイデアの詳細さや複雑さ）といった基準が用いられる[17]。しかし、LLM が生み出す「独創性」が、真に新しい概念の創出なのか、それとも膨大な訓練データに含まれるパターンの高度な再構成や模倣に過ぎないのかという点は、依然として議論の的である[17]。また、単に独創的であるだけでなく、生成された内容がタスクの要求を満たし、質の高いものであるか（新規性と品質のバランス）も重要な評価軸となる[29]。

2. 創造的成果物の評価方法

AI による創造的な成果物、特に物語生成のようなタスクの評価は複雑である。評価基準には、物語の一貫性、新規性、意外性、多様性などが含まれる。評価方法としては、専門家や非専門家による人間評価、あるいは他の LLM を評価者として用いる「LLM-as-a-judge」といった手法が用いられる[18]。

C. 現行評価方法論の限界

現在の AI 評価方法論、特に標準化されたベンチマークにはいくつかの限界が存在する。

- **ベンチマークの飽和とゲーミフィケーション:** 最先端モデルが一部のベンチマークで人間を超える、あるいは満点に近いスコアを達成するようになると、そのベンチマークはモデル間の能力差を識別する力を失う（飽和）。また、モデルが特定のベンチマークで高得点を取るように過剰に最適化され、実世界での汎用的な能力向上に繋がらない「ゲーミフィケーション」のリスクも指摘されている[4]。
- **データ汚染:** ベンチマークのテストデータが、意図的か否かにかかわらずモデルの訓練データに混入してしまうと、スコアが人為的につり上がり、モデルの真の能力を過大評価する結果となる[15]。
- **限定的な焦点:** 多くのベンチマークは、特定の隔離されたスキルやタスクを測定するものであり、実世界の複雑で多面的な問題解決能力や、長期的なプロジェクト遂行能力を十分に反映しているとは言えない[15]。
- **創造性評価の主観性:** 創造性のような主観的な質を客観的に測定することは本質的に困難である。専門家の合意に基づく評価手法（Consensual Assessment Technique: CAT）などが用いられるが、評価者間のばらつきやバイアスの影響を完全に排除することは難しい[18]。

これらの限界点を踏まえると、ベンチマークスコアは AI の能力の一側面を示すものに過ぎず、それらを絶対的な指標として捉えることには注意が必要である。多様な実世界

のシナリオにおける実際のタスク遂行能力や、質的な分析を含む多角的な評価アプローチが、AI の真の推論力と創造性を理解するためには不可欠と言える。この視点は、本レポート全体を通じて、各モデルの評価において繰り返し強調されるべき重要な前提となる。例えば、o3 が AIME で高いスコアを達成したという事実[13]と、一部のユーザーから実用的なコーディングタスクにおける「怠慢さ」が指摘されているという報告[5]との間のギャップは、まさにこのベンチマークと実用性の乖離を示唆している。

以下に、本レポートで言及される主要なベンチマークをまとめる。

表 II.1: 推論力と創造性に関する主要ベンチマーク

ベンチマーク名	説明	主要評価能力	o1/o3/Gemini 2.5 Pro との関連性
AIME	米国数学招待試験。高度な数学的問題解決能力を測定。	数学的推論、論理思考	OpenAI o1, o3, o3-mini, o4-mini, Gemini 2.5 Pro (Deep Think 含む) が挑戦し、高いスコアを記録。[1]
GPQA	大学院レベルの物理学、化学、生物学の質問。専門的科学知識と推論力を評価。	科学的推論、専門知識の応用	OpenAI o1, o3, o3-mini, Gemini 2.5 Pro が挑戦し、博士号レベルまたはそれを超える性能を示す。[1]
MMLU	57 の多様なタスクにおけるマルチタスク言語理解能力を測定。	広範な知識、問題解決能力	OpenAI o1, o3 などが GPT-4o を上回る性能を示す。[13]
Codeforces	競技プログラミングのプラットフォーム。アルゴリズム設計と実装能力を評価。	コーディング、論理的推論、アルゴリズム的思考	OpenAI o1, o3 が高いパーセンタイルを記録。[1]

SWE-bench	実世界のソフトウェアエンジニアリングタスクにおけるコード編集・生成能力を評価。	コーディング、問題解決、実用的ソフトウェア開発	OpenAI o3, Gemini 2.5 Pro が挑戦。o3 はカスタムスキャフォールドなしで SOTA を達成。[2]
HumanEval	Python コーディングタスク。基本的なアルゴリズムとデータ構造の理解を評価。	コーディング、基本的なプログラミング能力	多くの LLM のコーディング能力評価の標準。[16]
MMMU	マルチモーダル（テキスト、画像など）なマルチタスク・多言語理解能力を評価。	マルチモーダル推論、視覚情報と言語情報の統合	OpenAI o1, o3, Gemini 2.5 Pro (Deep Think 含む) が挑戦し、人間専門家レベルの性能を示す。[2]
LMArena	人間の嗜好に基づいてモデルの応答品質を評価するプラットフォーム。	対話品質、自然さ、有用性	Gemini 2.5 Pro がリード。[7]
WebDevArena	ウェブ開発タスクに特化したコーディングリーダーボード。	ウェブ開発関連のコーディング能力	Gemini 2.5 Pro がリード。[7]
ARC-AGI	知能測定を目的としたベンチマーク。抽象化と推論能力を評価。	一般的知能、抽象化、推論	OpenAI o3 が高スコアを達成したが、その解釈については議論がある。[37]
TTCT (改変版)	トーランス創造的思考検査の LLM 向け改変版。	流暢性、柔軟性、独創性、精巧性	LLM の創造性評価研究で利用。[17]
物語生成評価	物語の一貫性、新規性、意外性、多様性	創造的記述、物語構	LLM の創造的ライティング能力評価で利

	などを評価。	成	用。人間評価や LLM-as-a-judge が用いられる。[18]

この表は、読者が本レポートで議論される様々なベンチマークの概要を把握し、各モデルの性能評価の文脈を理解するための一助となることを意図している。

III. OpenAI o シリーズ：推論力と創造性の詳細分析

OpenAI の o シリーズは、同社が推論能力の飛躍的向上を目指して開発した一連のモデル群である。GPT シリーズの成功を基盤としつつも、特に複雑な問題解決や論理的思考に特化した設計が施されている。本章では、o1 およびその後継である o3 に焦点を当て、そのアーキテクチャ、推論・創造能力に関するベンチマーク性能、実用的な側面、そして限界について詳述する。

A. OpenAI o1 (o1-preview, o1-mini, o1-pro を含む)

1. アーキテクチャと推論の進化

OpenAI o1 は、2024 年 9 月 12 日にプレビュー版がリリースされ、同年 12 月 5 日にフルバージョンが公開された、反射型事前学習トランスフォーマー（GPT）である[1]。その中核的な設計思想は「応答前に思考する」という点にあり、複雑な推論タスク、科学、プログラミングにおいて GPT-4o を凌駕する性能を目指している[1]。このアプローチは、OpenAI の Mira Murati 氏によれば、モデルサイズや訓練データ、訓練計算量を増やす従来のモデルスケーリングパラダイムとは異なる、「新しい追加のパラダイム」であり、応答生成時により多くの計算能力を費やすことでモデルの出力を改善するものである[1]。

具体的には、o1 は応答を返す前に長い「思考の連鎖 (Chain of Thought: CoT)」を生成するように訓練されている[1]。この CoT プロセスは、人間が難問に取り組む際に時間をかけて考える様に似ており、強化学習（RL）を通じて、モデルが自身の思考プロセスを洗練させ、異なる戦略を試し、誤りを認識することを学習する[13]。

o1 シリーズにはいくつかのバリエーションが存在する。

- **o1-preview:** 初期にリリースされたプレビュー版。
- **o1-mini:** o1-preview よりも高速かつ 80%安価で、特にプログラミングや STEM 関連タスクに適しているが、「広範な世界の知識」は o1-preview ほどではないとされる[1]。
- **o1 (フルバージョン):** プレビュー版を経て正式リリースされたモデル。

- **o1-pro:** より多くの計算資源を使用し、より質の高い応答を提供することを目的とした高価格帯のバージョン。OpenAI の最も高価な AI モデルの一つである [1]。

2. 推論ベンチマークにおける性能

o1 シリーズは、特に複雑な推論を要するベンチマークにおいて、先行モデルからの大幅な性能向上を示した。

- **AIME (米国数学招待試験):** o1-preview は、問題の 83%（15 問中 12.5 問、コンセンサス評価）を解決し、GPT-4o の 13%（15 問中 1.8 問）を大きく上回った [1]。フルバージョンの o1 は、単一サンプルで平均 74%（11.1/15）、1000 サンプルを学習済みスコアリング関数で再ランク付けした場合に 93%（13.9/15）を達成した [13]。
- **GPQA Diamond (博士号レベルの科学問題):** o1-preview は物理学、化学、生物学関連のベンチマークテストで博士号レベルの性能を示し、o1 は人間の専門家の性能を上回ったとされる [1]。
- **Codeforces (競技プログラミング):** 89 パーセンタイルにランクインした [1]。
- **MMLU (大規模マルチタスク言語理解):** GPT-4o を 57 サブカテゴリ中 54 で上回った [13]。

これらの結果は、o1 が特定の推論タスクにおいて顕著な能力を持つことを示唆している。

3. 創造的能力と実用的性能

o1 シリーズは、その推論能力を活かして、創造的なタスクや実用的な問題解決にも応用が期待された。

- **コーディング:** o1-mini は特にプログラミングに適しているとされ [1]、o1-preview のユーザーからは、完全に動作するコードを提供してくれるとの肯定的なフィードバックがあった [3]。
- **STEM タスク:** 細胞シーケンシングデータのアノテーション支援や、量子光学に必要な複雑な数式の生成など、高度な科学技術タスクでの活用例が示された [19]。
- **ユーザーエクスペリエンス (プレビュー版 vs. 正式版):**
  - o1-preview は、複雑なタスクや詳細なコード生成において好意的に受け止められた [3]。
  - しかし、o1 のフルバージョンリリース後は、プレビュー版と比較して「性能が低下した」「怠慢になった」「応答が短く不完全なコードスニペットしか提供しない」といった批判が一部ユーザーから上がり、Claude のような代替モデルを検討する声も聞かれた [3]。一部には、小規模な LLM よりもコーディング

能力が劣るとの評価もあった[3]。
    - o1-pro に対する評価は分かれた。高価な割に o1 からの性能向上が僅かであり、プログラミングには時間がかかりすぎるとの意見がある一方で、高レベルで自由度の高い質問に対しては「プログラミングの怪物」であり、大規模なコードベースのリファクタリングを成功させるといった肯定的な報告も存在した[4]。

4. 強みと限界

o1 シリーズは、推論能力の向上という点で重要な一歩を示したが、いくつかの課題も抱えている。

- **強み (プレビュー版/コンセプト):** 数学、科学、コーディングといった複雑な推論分野において、既存モデルから大幅な能力向上を達成した[1]。
- **限界 (正式版/Pro 版で観測されたもの):**
    - **コストとアクセス:** o1-preview の API は GPT-4o より大幅に高価で、フル o1 API へのアクセスも当初は制限されていた[1]。o1-pro は OpenAI の最も高価なモデルの一つである[1]。
    - **性能低下/一貫性の欠如:** プレビュー版からフルバージョンの o1 への移行に際して、一部ユーザーから顕著な品質低下が指摘された[3]。
    - **非公開の思考の連鎖 (CoT):** OpenAI は、安全性と競争上の優位性を理由に、o1 の CoT の開示を禁止している。これは、大規模言語モデル（LLM）を扱う開発者から透明性の損失であると批判されている[1]。
    - **無関係な情報への感度:** 問題に無関係な情報を追加すると、o1-preview および o1-mini の性能が大幅に低下した（それぞれ-17.5%、-29.1%）[1]。
    - **安全性への懸念:** 高度な科学的理解力を持つため、生物兵器関連などでの悪用が懸念された[1]。また、o1 が潜在的に危険な助言に対して GPT-4o の一般的な拒否よりも詳細な応答を生成する傾向があり、これがより危険と評価されるケースがあった[22]。

o1 シリーズは、「思考時間」を確保することで推論能力を大幅に向上させるというコンセプトを実証し、特に STEM 分野での可能性を示した。しかし、プレビュー版から製品版への移行において、性能の一貫性の維持、ユーザーの期待管理、そして能力とコスト・アクセスのバランス取りに課題が生じ、初期の期待と一部ユーザーの実用体験との間に乖離が見られた。特に、CoT の非公開は透明性に関する重要な論点となっている。この背景には、強力な推論能力が、実環境で全てのユーザーに対して効果的に展開するには不安定であるか、あるいはリソース集約的すぎる可能性があることが示唆される。

B. OpenAI o3 (o3-mini, o3-pro (該当する場合) を含む)

1. o1 からの進化：強化された推論、ツール利用、マルチモーダリティ

OpenAI o3 は、o1 の後継として位置づけられ、同社の最も強力な推論モデルとして 2025 年 1 月 31 日に o3-mini が、その後フルバージョンがリリースされた[2]。o3 シリーズは、o1 の推論能力を基盤としつつ、ツール利用の統合とマルチモーダルな理解能力を大幅に強化している。

- **ツール連携の強化:** o3 は、ChatGPT 内で利用可能な様々なツール（ウェブ検索、Python によるファイル分析、画像分析、画像生成、メモリなど）をエージェント的に利用し、組み合わせて使用することができる[2]。これらのツールは、モデルの思考の連鎖（CoT）の中で活用され、問題解決能力を拡張する[39]。
- **マルチモーダル推論（「画像による思考」）:** o3 は、画像を単に見るだけでなく、直接的に思考の連鎖に統合することができる。ユーザーがアップロードした画像をトリミング、ズーム、回転などの処理を加えながら分析し、問題解決に役立てる[2]。
- **強化学習のさらなる活用:** 大規模な強化学習が継続されており、訓練計算量と推論時の思考時間を増やすことで性能が向上し続けることが確認されている[2]。
- **o3-mini:** o1-mini の後継として、コスト効率に優れながらも STEM 分野（科学、数学、コーディング）で高い能力を発揮する。関数呼び出しなどの開発者向け機能もサポートし、より複雑な課題に取り組む際には「より深く思考する」か、速度が求められる場合には速度を優先するといった柔軟性を持つ[23]。ただし、視覚的な推論タスクには対応していないため、その場合は o1（または o3 フルバージョン）が推奨される[23]。

2. 推論および創造性ベンチマークにおける性能

o3 シリーズは、多くの主要ベンチマークで最高水準（SOTA: State-of-the-Art）の性能を達成している。

- **主要ベンチマークでの SOTA:** Codeforces 、SWE-bench（カスタムスキャフォールドなし）、MMMU などで SOTA を記録[2]。
- **AIME:** o3-mini は、中程度の推論努力で o1 と同等、高程度の推論努力で o1 を上回る性能を示した[23]。o3 フルバージョンは AIME 2024 で 91.6%を達成[25]。
- **GPQA Diamond:** o3-mini は高程度の推論努力で o1 と同等レベル[23]。o3 フルバージョンは 83.3%を達成[25]。
- **MMMU:** o3 フルバージョンは 82.9%[25]。MathVista では 86.8%[25]。
- **ARC-AGI:** o3 は 87.5%という高いスコアを達成したが、一部の研究論文では、これは汎用的な知性によるものではなく、膨大な計算資源を用いた試行錯誤の結果で

あると指摘されている [37]。

- **FrontierMath:** OpenAI は当初 25%以上のスコアを主張したが、後に独立したテスト機関が公開版モデルで約 10%のスコアを報告し、透明性やテスト条件に関する議論を呼んだ [41]。

3. 創造的能力と実用的性能

o3 は、その高度な推論能力とツール連携を活かし、創造的なタスクにおいても高い性能を発揮することが期待されている。

- **創造的アイデア生成:** 特に生物学、数学、工学の文脈において、新しい仮説を生成し、批判的に評価する能力が注目されている [2]。
- **画像生成:** ネイティブな画像生成能力を持ち、単に画像を見るだけでなく、画像と共に思考することができる [2]。
- **コーディング:** Codeforces や SWE-bench で SOTA を達成するなど、ベンチマーク上では高いコーディング能力を示す [2]。しかし、実用的なコーディングタスクにおいては、o3 および o4-mini が「極めて悪い、怠慢だ」といった厳しいユーザー評価も多く、指示に従わない、不完全なコードを生成する、コード内の幻覚（ハルシネーション）率が高いといった問題が指摘されている [5]。一部ユーザーは、以前のモデルからの大幅な性能低下を感じている [5]。一方で、「ディープリサーチ・ライト」的なタスクやエージェント的なワークフローには適しているとの評価もある [43]。
- **視覚タスク:** 画像、チャート、グラフィックの分析に優れている [2]。
- **対話性/パーソナリティ:** 一部のユーザーは、o3 が以前の OpenAI モデルよりも優れたパーソナリティを持ち、より創造的で対話が楽しいと感じている [10]。

4. 指摘されている問題点と限界

o3 は高い能力を持つ一方で、いくつかの重要な課題も抱えている。

- **幻覚 (ハルシネーション):**
  - o3 は全体的により多くの主張をする傾向があり、その結果、正確な主張が増える一方で、不正確な主張や幻覚も増える、特に PersonQA においてその傾向が顕著である [39]。
  - OpenAI 自身も、o3 がコーディングにおける関数呼び出しの幻覚を起こしやすい可能性があることを認識しており、これを軽減するためのガイダンスを提供している [6]。
  - ユーザーからは、コード生成時に存在しない関数を呼び出すなど、高い幻覚率が報告されている [42]。

- **コスト:** o3、特にフルバージョンは高価である[9]。
- **コーディング性能の不一致:** ベンチマークでの SOTA 達成と、実用的なコーディングにおけるユーザーからの厳しい批判との間に大きな隔たりがある[2]。ChatGPT Plus ティア版でのコンテキストウィンドウの制限も、コーディング時の問題として指摘されている[42]。
- **ベンチマークスコアの不一致:** FrontierMath のスコアに関する論争は、透明性や、内部研究版と公開版モデルとの差異に関する疑問を提起している[41]。ARC-AGI のスコアについても、「真の知性」対「総当たり的探索」という点で議論がある[37]。

o3 は、高度な推論能力と実用的なツール利用、そしてマルチモーダルな理解（特に視覚）を統合する上で大きな進歩を示している。印象的なベンチマークスコアを達成している一方で、特にコーディングにおける実用性についてはユーザー間で評価が大きく分かれており、「怠慢さ」、幻覚の頻発、以前のモデルやより小規模な特化モデルと比較しての信頼性の低下などが頻繁に報告されている。これは、非常に複雑な推論メカニズムを大規模に、かつ安定して展開することの難しさ、あるいはトレードオフが存在する可能性を示唆している。モデルがより「エージェント的」になり、ツールを用いた複雑な複数ステップの推論を試みるようになると、潜在的な障害点や信頼性の低い挙動のモードも増加し、強力ではあるが、特定のタスクにおいては予測可能性や依存性が低下する可能性がある。その「推論」能力は強力であるものの、まだ一貫して制御可能であるか、あるいは堅牢であるとは言えない状況のようだ。

IV. Google Gemini 2.5 Pro ：推論力と創造的領域における能力

Google の Gemini 2.5 Pro は、同社の AI 開発における最先端モデルとして位置づけられ、その高度な推論能力、広大なコンテキスト処理能力、そしてマルチモーダルな対応力で注目を集めている。本章では、Gemini 2.5 Pro のアーキテクチャ上の特徴、推論および創造性に関するベンチマーク性能、実用的な応用例、ユーザーからのフィードバック、そして現時点での限界について詳細に分析する。

A. アーキテクチャのハイライトと推論アプローチ

Gemini 2.5 Pro は、応答生成前に十分な「思考」時間を持つ「思考モデル」として設計されている[7]。このアプローチは、OpenAI の o シリーズと同様に、より複雑でニュアンスのある問題解決を目指すものである。

- **「Deep Think 」機能:** 特に注目すべきは、「Deep Think」と呼ばれる実験的な強化推論モードである。これは、応答前に複数の仮説を検討することを可能にする新しい研究技術を用いており、非常に複雑な数学やコーディングタスクに対応することを目的としている[8]。

- **広大なコンテキストウィンドウ:** 最大 100 万トークンという非常に大きなコンテキストウィンドウをサポートしており、長文の文書理解やビデオ理解において最先端の性能を発揮する [8]。これにより、書籍全体や大量の法的記録といった大規模なデータセットを一度に処理することが可能となる。
- **マルチモーダル入力:** テキストだけでなく、音声、画像、ビデオといった多様な形式の入力を処理できる [7]。これにより、より現実に近い複雑な情報を統合的に理解し、応答を生成することが期待される。
- **「思考バジェット」:** 開発者がモデルの「思考」処理量を調整できる機能。これにより、コストとレイテンシのバランスを取りながら、タスクの要求に応じてモデルの計算資源の投入量を制御できる [7]。

B. 推論および創造性ベンチマークにおける性能

Gemini 2.5 Pro は、多くの主要なベンチマークで高い性能を示している。

- **LMArena:** 人間の嗜好に基づくモデル応答評価でトップクラスの評価を得ており、ユーザーにとって質の高い応答を生成することを示唆している [7]。
- **WebDevArena:** コーディングに関する主要なリーダーボードで ELO スコア 1415 を達成し、トップの座を維持している [7]。
- **数学ベンチマーク:**
  - 「Deep Think」機能を用いた場合、2025 年の USAMO（米国数学オリンピアード）で「印象的なスコア」を記録 [8]。
  - 多数決のようなテスト時計算量増加テクニックなしで、AIME 2025 でリード [14]。
- **コーディングベンチマーク:**
  - 「Deep Think」機能を用いた場合、LiveCodeBench でリード [8]。
  - SWE-Bench Verified では、カスタムエージェント設定で 63.8%を達成 [14]。
  - Aider Polyglot ベンチマークでは 72.90%（o3 の 79.60%と比較）[9]。
- **MMMU:** 「Deep Think」機能を用いた場合、84.0%を達成 [8]。
- **GPQA:** 多数決のようなテスト時計算量増加テクニックなしでリード [14]。Epoch.ai による独立した再現テストでも Gemini の GPQA スコア 84%が確認されている（同テストで o3 は 82%）[36]。
- **Humanity's Last Exam:** 18.8%のスコアを記録し、o3 mini や Claude 3.7 Sonnet を上回った [47]。

C. 実用的な創造性

Gemini 2.5 Pro は、その高度な能力を活かして、様々な創造的タスクで高い評価を得ている。

- **創造的ライティング:**
  - ユーザーからは、「信じられないほど素晴らしい」「以前のバージョンから飛躍的に向上した」「物語を本当に掴んでいる」といった非常に肯定的なフィードバックが寄せられている[48]。ニュアンスの捉え方や現実的なキャラクター描写において、他のモデルを「頭一つ抜けている」との評価もある[11]。
  - 応答のスタイルや構造が改善され、より創造的で整形された回答を提供できるようになった[7]。
- **コーディング（「Vibe Coding」）:**
  - コンテキスト認識能力とコードの反復改善能力の高さから、「Vibe Coding」（感覚的なコーディングスタイル）においてしばしば好まれる[9]。o3 と比較してデバッグの必要性が少なく、より最新のパッケージを使用する傾向があるとの指摘もある[10]。
  - 一部のユーザーは、エラーの少ない優れたコーディング能力を高く評価している[9]。
- **マルチモーダルコンテンツ生成:**
  - 自然な会話体験を実現するネイティブな音声出力をサポート[8]。
  - テキスト読み上げ（TTS）機能では、Gemini 2.5 Pro Preview TTSが Google の最も強力な TTS モデルとされ、ポッドキャストやオーディオブックなどの制作に活用できる[44]。ユーザーからは「危険なほど強力」との声も上がっている[50]。
  - 一行のプロンプトからビデオゲームのコードを生成する能力も示されている[14]。
- **学習・教育分野:** LearnLM（教育専門家と共に構築されたモデルファミリー）の統合以来、学習分野で主要なモデルとなっており、教育者からも高い評価を得ている[8]。

D. ユーザーからのフィードバックと認識されている限界

Gemini 2.5 Pro は多くの強みを持つ一方で、いくつかの課題も指摘されている。

- **強み（ユーザーが認識）:** コーディングの反復開発やコンテキスト認識に優れている[9]、創造的な文章作成が非常に得意[11]、多くのタスクで o3 よりも安価[10]、コンテキストウィンドウが広い[10]、幻覚が少ない[10]、日常的な利用において高速で実用的[10]。
- **限界・課題:**
  - **API の安定性・エラー:** 特にマルチモーダルプロンプティングにおいて、API から 5XX エラーや空の応答が返されることがあり、一部開発者からは「テストがほとんど不可能」との声も上がっている[12]。

- **ドキュメント処理:** Gemini アプリ（gemini.google.com）において、コンテキストやドキュメントの扱いに苦慮するケースが報告されており、モデルがファイルにアクセスできなかったり、誤った回答をしたりすることがある[51]。
- **複雑な問題に対する単純な解決策:** ある博士課程の学生は、Gemini が博士号レベルの問題に対しては単純すぎる解決策を提示する傾向があり、o3 のような想像力豊かな深みに欠けると感じている[10]。別のユーザーは、Gemini 2.5 Pro を幅広い知識を持つが実務経験に乏しい「ジュニアプログラマー」のようだと評し、o3 の方が「シニアレベルのプログラミング」に適していると感じている[10]。
- **安全フィルタリング:** 安全性への配慮から、成人向けフィクションで扱われるようなトピックに対しても過度に慎重になり、応答を拒否することがある。これが創造的なユースケースを妨げているとの指摘がある[11]。

Gemini 2.5 Pro は、多くのユーザーにとって実用性が高く、特に反復的なコーディング作業（「Vibe Coding」）や創造的な文章作成において優れた能力を発揮している。その広大なコンテキストウィンドウと人間による高い評価は、このモデルの強固な基盤を示している。「Deep Think」機能は、より深い推論へのコミットメントを明確にしている。しかしながら、API の安定性に関する問題や、複雑なドキュメント処理における時折の困難、あるいは非常にニュアンスの深い学術的な問いに対する応答が単純化されすぎるといった点は、さらなる改善の余地があることを示している。また、高度な能力と安全フィルタリングとの間の緊張関係も、一部の創造的な応用において影響を及ぼしている。このモデルは、一般的な高度タスク、特に長文コンテキストやユーザーの「感触」が重要なタスクにおいては非常に強力なジェネラリストであるが、専門的でリスクの高い、あるいは極めて複雑なシナリオにおいては、まだ荒削りな部分が残っているか、その推論が広範ではあっても、特定の専門家ユーザーにとっては o3 のような「想像力豊かな深み」には必ずしも達していない可能性がある。

V. 比較分析：OpenAI o シリーズ vs. Google Gemini 2.5 Pro

OpenAI の o シリーズ（特に o1 と o3）と Google の Gemini 2.5 Pro は、現行の推論型生成 AI の最前線を代表するモデルであり、それぞれが独自の強みと特徴を有している。本章では、これらのモデルを推論能力、創造性、実用性能、コストといった観点から直接比較し、その差異を明らかにする。

A. 推論能力の直接比較

1. ベンチマーク対決

ベンチマーク	OpenAI o1 (バリアント)	OpenAI o3 (バリアント)	Google Gemini 2.5 Pro (バリアント)	主要な観察/ニュアンス
AIME (数学)	o1-preview: 83% (コンセンサス)[1]。o1 (フル): 74% (単一), 93% (再ランク)[13]	o3-mini (高努力): o1 を凌駕[23]。o3 (フル): 91.6% (AIME 2024)[25]	Deep Think で高スコア (USAMO 2025)[8]。AIME 2025 でリード (テスト時テクニックなし)[14]	いずれも数学的推論に強い。o4-mini もツールアクセスありで AIME 2024/2025 で SOTA[2]。
GPQA (科学)	o1: 博士号レベル、人間専門家超え[1]	o3-mini (高努力): o1 と同等[23]。o3 (フル): 83.3%[25]	リード (テスト時テクニックなし)[14]。Epoch.ai による独立テストで 84% (o3 は 82%)[36]	非常に僅差。テスト条件により若干の変動あり。
コーディング (SWE-bench, Codeforces 等)	o1: Codeforces で 89 パーセンタイル[1]	o3: Codeforces, SWE-bench で SOTA[2]。Aider Polyglot で 79.60%[9]	Gemini 2.5 Pro: WebDevArena でリード[8]。SWE-Bench (カスタムエージェント)で 63.8%[14]。Aider Polyglot で 72.90%[9]	ベンチマークにより優劣が異なる。o3 は一部で SOTA だが、Gemini も特定分野で強い。
MMMU (マルチモーダル)	o1: 78.2% (視覚有効時)[13]	o3: 82.9%[25]	Gemini 2.5 Pro (Deep Think): 84.0%[8]	Gemini 2.5 Pro が僅かにリードする傾向。
LMArena (人間嗜好)	N/A (o シリーズは直接比較データ少)	N/A	Gemini 2.5 Pro: リード[7]	Gemini 2.5 Pro が人間にとって好ましい応答を生成する傾向。

2. 推論アプローチの質的差異

- **OpenAI o シリーズ:** 明示的で、より深い可能性のある「思考の連鎖 (CoT)」を強化学習で洗練させるアプローチを強調している[1]。一部の専門家ユーザーからは、より「想像力豊か」で、「シニアレベル」の複雑な問題に適していると評価されている[10]。これは、問題解決において、より多くの計算時間を費やして多様な思考経路を探求する設計思想の表れかもしれない。
- **Google Gemini 2.5 Pro:** 「Deep Think」機能は仮説探求型の推論を示唆し[8]、「思考バジェット」は推論の深さとコストの制御を可能にする[7]。一部ユーザーからは、より単純で直接的な解決策を提示する傾向があり、これが文脈によっては長所にも短所にもなり得ると評価されている[10]。これは、効率性と広範な適用性を両立させようとする設計思想を反映している可能性がある。

B. 創造性の直接比較

創造性の側面	OpenAI o1 (バリアント)	OpenAI o3 (バリアント)	Google Gemini 2.5 Pro (バリアント)	裏付け証拠/ユーザーフィードバック
物語生成 (一貫性, 独創性, 精巧性)	具体的なユーザー評価は少ない	「良い個性」「創造的」との評価[10]。仮説生成能力に注目[2]。	ユーザーから非常に高い評価。「物語を掴む」[48]。「ニュアンス、現実的なキャラクター描写で他を圧倒」[11]。	Gemini が物語生成の質で好まれる傾向。
コーディング (解決策の洗練度, 新規性)	o1-pro: 複雑なタスクで創造的との評価[4]。	ベンチマークは SOTA だが、実用では「怠慢」「幻覚」の指摘[2]。	反復的「Vibe Coding」、コンテキスト認識で好まれる[9]。	実用的なコーディングの創造性では Gemini が優位か。o シリーズはベンチマークと実用の乖離が課題。
アイデア生成	N/A	生物学、数学、	「Deep Think」	o3 が特定の学術

(仮説生成)		工学分野での新規仮説生成と批判的評価能力が強調[2]。	が複雑な問題での仮説探求を示唆[8]。	分野での高度なアイデア生成に強みを持つ可能性。
マルチモーダル創造性 (画像/音声)	N/A	ネイティブ画像生成、「画像による思考」[2]。	強力なネイティブ音声・TTS 機能[8]。	o3 は視覚、Gemini は聴覚の創造性でそれぞれ特徴的な強み。

一般的な LLM の創造性に関する研究[17]を踏まえると、両モデルファミリーとも精巧性（Elaboration）には優れている可能性が高い。真の独創性（Originality）については、依然として訓練データへの依存という制約があり、複雑な問題である。o3 の「新規な仮説を生成し批判的に評価する」能力[2]は、より高次の創造的推論を示唆している。

C. 実用シナリオにおける性能 (ユーザー視点)

ユースケース	OpenAI o シリーズ (総評)	Google Gemini 2.5 Pro (総評)	主な強み	主な弱点/懸念
反復的コーディング	o3 はベンチマーク SOTA だが実用で賛否[2]。o1-pro は一部で高評価[4]。	しばしば好まれる。コンテキスト処理、デバッグの少なさ、最新パッケージ使用[9]。	Gemini: 反復開発、コンテキスト認識。o1-pro: 複雑なタスク。	o3: 実用での信頼性、幻覚。o1(フル): 性能低下の指摘。
複雑な STEM 問題解決/研究	o3 は「ディープリサーチ・ライト」として分析力評価[2]。博士課程ユーザーは o3 の「想像力」を評価[10]。	広大なコンテキストで長文分析に強み[8]。「Deep Think」が複雑な数学/コーディングに対応[8]。	o3: 想像力豊かな深み、STEM 分析。Gemini: 長文コンテキスト、Deep Think。	Gemini: 一部で単純すぎる解決策の指摘。o シリーズ: コスト。
長文ドキュメン	o3 は 20 万トー	100 万トークン	Gemini: 圧倒的	o シリーズ: コン

ト分析	クンコンテキスト[25]。	コンテキストで圧倒的有利[8]。	なコンテキスト長。	テキスト長の限界。
創造的ライティング	o3 は「創造的」との評価はあるが、物語生成の具体的賞賛は Gemini より少ない[10]。	ユーザーから非常に高い評価。質、ニュアンス[7]。	Gemini: 物語の質、ニュアンス。	o シリーズ: Gemini ほどの特筆すべき評価は少ない。
マルチモーダルタスク	o3 は「画像による思考」、ネイティブ画像生成[2]。	強力なネイティブ音声・TTS、ビデオ理解[8]。	o3: 視覚統合推論。Gemini: 音声・ビデオ関連。	それぞれ得意分野が異なる。
全般的なユーザビリティ/信頼性	コスト、レート制限、性能の一貫性に懸念[1]。	一部ユーザーに高速、安価、実用的と評価[10]。API 安定性に課題[12]。	Gemini: 日常利用での実用性、コスト効率。	o シリーズ: コスト、一貫性。Gemini: API 安定性。

D. コスト、アクセシビリティ、エコシステム

- **OpenAI:** o シリーズ、特に Pro バージョンは高価である[1]。最新モデルの API アクセスは段階的または制限されることがある[1]。ChatGPT を通じて広範なユーザーベースを持つ。
- **Google:** Gemini 2.5 Pro は多くのタスクでよりコスト効率が良いと位置づけられている[10]。Google AI Studio、Vertex AI、Gemini アプリに統合されている[7]。レート制限付きで無料ユーザーにも広く提供されている[47]。

総合的に見ると、OpenAI の o3 は生の IQ や推論ベンチマークでしばしば SOTA を主張し、一部の専門家からは深く想像力豊かな問題解決において好まれる一方で、Google の Gemini 2.5 Pro は実用的なユーザビリティ、反復的なコーディング、創造的な文章作成の質（ユーザー嗜好）、そしてより広範なタスクにおけるコスト効率で頻繁に優位性を示している。両者ともマルチモーダルのフロンティアを押し広げているが、o3 が視覚的推論の統合に、Gemini が音声/TTS と広大なコンテキストにそれぞれ異なる方向で注力している。どちらのモデルが「優れている」かは、特定のユースケース、ユーザーの専門知識、そしてコスト対潜在的な信頼性の低さへの許容度に大きく依存す

る。この背景には、o3 のようなモデルが特定の推論領域で「生の能力」を発揮する一方で、実用上のハードルに直面しているのに対し、Gemini 2.5 Pro は多くの一般的な高度タスク、特に長文コンテキストが有利に働くタスクやユーザーの「感触」が重要なタスクにおいて、よりバランスの取れた実用的なツールとして機能しているという状況が考えられる。

VI. 実世界の応用と影響

OpenAI の o シリーズや Google の Gemini 2.5 Pro のような高度な推論型生成 AI は、既に様々な専門分野でその影響力を示し始めている。これらのモデルは、単なる研究対象から実用的なツールへと移行しつつあり、生産性の向上や新たな発見の促進に貢献している。しかし、その応用は常に順風満帆というわけではなく、実世界の複雑さや人間の期待との間に課題も存在する。

A. コーディングとソフトウェア開発

AI はソフトウェア開発のあり方を大きく変えつつある。

- **生産性向上支援:** GitHub Copilot のような AI 支援ツールを利用する開発者は、コーディング速度が 55%向上したという報告がある [34]。これらのモデルは、定型的なコードの生成、解決策の提案、デバッグ支援などを行うことができる [1]。
- **信頼性の課題:** 生成されたコードは多くの場合、人間の開発者によるレビューと修正を必要とする [35]。AI による「怠慢」なコード生成、幻覚、論理エラー、古い手法の利用といった懸念も指摘されている [3]。
- **コード理解支援:** AI アシスタントは、既存のコードを説明するという、開発者にとって重要な意味理解タスク（センスメイキング）に頻繁に利用されている [35]。
- **開発者ワークフローへの影響:** 単純なコード生成から、AI との共同作業によるアイデア創出や問題解決へと、開発者の役割が変化しつつある [35]。

B. 創造的コンテンツ生成

AI は文章作成、アート、音楽など、多岐にわたる創造的分野で活用されている。

- **文章作成:** 記事、物語、脚本などの草稿作成において強力な能力を発揮する。特に Gemini 2.5 Pro は、質の高い創造的ライティングで注目されている [7]。しかし、LLM は一般的に、真の創造性（新規性、意外性）という点では経験豊富な人間の作家には及ばないことが多いとされる [17]。
- **アート・画像生成:** OpenAI o3 は DALL-E のような画像生成機能を統合し、推論プロセスの中で画像を生成・活用できる [2]。Midjourney のような特化型ツールとの比較では、LLM 統合型ツールはプロンプトへの忠実度やテキスト生成に優れる一

方、Midjourney は美的品質や芸術的スタイルで勝るが、学習曲線が急であるとされる[54]。
- **音楽・音声生成:** Gemini 2.5 Pro は高度なテキスト読み上げ（TTS）機能やネイティブな音声処理能力を提供し、ポッドキャスト生成や表現力豊かな朗読などを可能にしている[8]。

C. 科学的発見と仮説生成

AI は科学研究のプロセスを加速させる可能性を秘めている。

- **AI 共同科学者 (Gemini 2.0 ベース):** 文献情報を統合し、長期的な計画立案を行うことで、新しい仮説、研究提案、実験プロトコルの生成を支援するシステムが開発されている[56]。
- **o3 の可能性:** 生物学、数学、工学分野における新規仮説の生成と評価能力が注目されている[2]。
- **HyperWrite の仮説メーカー (GPT-4 ベース):** 研究上の問いから仮説を生成するツールも登場している[58]。
- **影響:** 発見の加速、実験的検証の効率化、科学の民主化などが期待される[56]。

D. 法務および専門的文書分析

法律分野などの専門分野でも、AI による文書処理の効率化が進んでいる。

- **法務テックにおける LLM:** 文書レビュー、契約分析、法的調査、デューデリジェンスなどの自動化に活用されている[60]。
- **利点:** 効率と精度の向上、コスト削減、専門家がより価値の高い戦略的業務に集中できる環境の実現[60]。
- **Gemini 2.5 Pro の長文コンテキスト:** 大量の法的記録などの分析に適している[25]。
- **課題:** 生成内容の正確性担保、データプライバシーの保護、訓練データに存在するバイアスの緩和が不可欠である[61]。依然として人間の監督が重要となる[61]。

E. ベンチマーク性能と実用性のギャップ

繰り返しになるが、高いベンチマークスコアが必ずしも円滑な実世界での応用に直結するわけではない[4]。実用性は、使いやすさ、既存システムとの統合の容易さ、長時間の対話における信頼性、コスト、そしてベンチマークでは捉えきれない特定のタスクのニュアンスといった多くの要因に影響される[15]。

これらの高度な AI モデルの実用的な影響は、既に多様な専門分野で感じられ始めており、大幅な生産性向上や新たな発見の道筋を提供している。しかし、この影響はしばしば人間の監督の必要性、複雑またはニュアンスに富んだタスクにおける信頼性の課題、

そして標準化されたテスト性能と実世界のワークフローの複雑さとの間の持続的なギャップによって調整される。これらのツールを真に信頼できるパートナーとするための「最後の 1 マイル」は、まだ構築途上にあると言える。この状況は、AI が強力なアシスタントである一方で、自律的な専門家にはまだ至っていないことを示している。その有用性は、人間の専門知識を置き換えるのではなく、それを増強する「共同操縦士」や「共同科学者」として使用される場合に最大化される。創造性の民主化効果[28]は大きいものの、AI の生成物に対する批判的な評価の必要性も同様に重要である。

VII. 現在の課題、限界、および将来展望

最先端の推論型生成 AI は目覚ましい進歩を遂げているが、その能力を最大限に引き出し、社会に責任ある形で実装するためには、依然として多くの技術的・倫理的課題が存在する。本章では、これらの課題と限界を整理し、AI の推論力と創造性の将来的な展望について考察する。

A. 持続する技術的・倫理的課題

- **幻覚 (ハルシネーション) と事実の正確性:** モデルは依然として、もっともらしいが不正確または無意味な情報を生成することがある。OpenAI o3 は主張の量が増えることで、正確な情報と共に不正確な情報や幻覚も増加する傾向が指摘されている[5]。Gemini のユーザーからも、事実誤認や誤解が報告されている[51]。この問題は、AI の信頼性を損なう主要な要因の一つである。
- **バイアス:** 膨大な訓練データセットに存在するバイアスが、モデルの出力において永続化または増幅される可能性がある。これは、法務分析や採用といった応用分野における公平性に影響を及ぼしかねない[61]。
- **安全性とアラインメント:** モデルが安全ガイドラインを遵守し、有害なコンテンツを生成したり、悪用を助長したりしないようにすることが極めて重要である（例：o1 の生物兵器に関する懸念[1]）。OpenAI の o シリーズモデルは、ジェイルブレイク評価において堅牢性の向上を示しているが、o1 の簡潔すぎる拒否応答や、危険な助言に対してより深く関与してしまうといった課題は残っている[19]。
- **過度の依存:** ユーザーが AI の出力を批判的に評価することなく過度に依存するリスクがある。特に、リスクの高いタスクにおいては注意が必要である[22]。

B. 推論の透明性と説明可能性

- **「ブラックボックス」問題:** 深層学習モデル、特に LLM は、その内部動作の解釈が困難な「ブラックボックス」としての性質を持つことが多い。
- **OpenAI o1 の非公開の思考の連鎖:** 推論プロセスの意図的な秘匿は、モデルの挙動を理解し、デバッグや検証を行う必要がある開発者や研究者にとって懸念材料とな

っている[1]。思考プロセスを示すことが信頼構築に不可欠であるという考え方とは対照的である。
- **信頼と採用への影響:** 透明性の欠如は、説明責任が求められる重要な分野での AI 導入を妨げる可能性がある。

C. AI の推論力と創造性の未来

- **現行アーキテクチャの先へ:** トランスフォーマーのスケーリングや現在の CoT 手法を超える新しいアプローチの可能性が探求されている。ARC-AGI に関する論文は、現在の LLM の手法（大規模な試行錯誤）が AGI（汎用人工知能）への道ではない可能性を示唆している[37]。
- **マルチモーダリティの役割:** 視覚、聴覚、その他の感覚データをより深く統合することで、推論能力（複雑な図表や現実世界のシーンの理解など）と創造性（よりリッチなマルチメディア生成など）の両方が大幅に向上すると期待される[2]。
- **エージェント AI:** 自律的に計画を立て、ツールを使用し、複雑な複数ステップのワークフローを実行できるモデルがより普及し、複雑なオペレーションに影響を与えるだろう[2]。
- **パーソナライゼーションとコンテキスト:** モデルは、記憶や過去の対話履歴を活用して、より関連性の高い、パーソナライズされたインタラクションを提供する能力が向上すると考えられる[2]。

D. 人間の役割とスキルへの影響

- **変化する雇用環境:** 反復的なタスクの自動化が進む一方で、批判的思考、創造性、AI 管理スキルを必要とする役割への需要が高まる[28]。
- **拡張された創造性:** AI は、創造的な行き詰まりを克服し、創造活動を民主化し、より高レベルの概念化に集中することを可能にするツールとして機能する[28]。
- **新たなスキルの必要性:** プロンプトエンジニアリング、AI 倫理、データリテラシー、AI の出力を批判的に評価する能力などが重要になる[62]。

高度な生成 AI の進む先は、ますます有能で自律的なマルチモーダルエージェントの登場を示唆している。しかし、この未来を責任ある形で実現するためには、幻覚やバイアスといった根本的な限界に対処し、より大きな透明性を育み、そして人間がこれらの強力な技術と効果的に協調できるようにスキルや社会構造を適応させるという、重要なブレークスルーが必要となる。「真の」知能対洗練されたパターンマッチングという議論は今後も続くだろう。この道のりはエキサイティングであると同時に、多くの課題を伴う。進歩は、単なるスケーリングだけでなく、安全性、透明性、そして AI 認知の本質に関する基礎研究、さらには社会全体の適応にかかっている。

VIII. 結論

本レポートでは、OpenAI の o1 および o3 シリーズ、ならびに Google の Gemini 2.5 Pro という最新の推論型生成 AI モデルの推論力と創造性について、学術的および実用的観点から包括的な分析を行った。

A. 主要な調査結果の要約

- **OpenAI o シリーズの強み:** 特に STEM 分野における深い推論能力、o3 における高度なツール利用と「画像による思考」といった視覚的推論の統合が挙げられる。
- **Google Gemini 2.5 Pro の強み:** 広大なコンテキストウィンドウ、ユーザーに好まれる創造的な文章作成能力と反復的なコーディング支援、強力な音声関連機能、そして「Deep Think」による潜在的な高度推論能力が特徴である。
- **各モデルの弱点:**
  - **OpenAI o シリーズ:** 高コスト、o3 のコーディングにおけるユーザーからの信頼性や幻覚に関する指摘、o1 正式版の期待外れ感。
  - **Google Gemini 2.5 Pro:** API の安定性に関する問題、一部の複雑な問いに対する応答の過度な単純化、あるいは安全フィルターによる創造的表現の制約。

B. 現状の最先端技術に関する総合評価

AI の推論力と創造性は目覚ましい進歩を遂げており、現行モデルはかつて人間特有と考えられていた能力を発揮し始めている。特定のベンチマークにおいては「超人的」な性能を示す一方で、実用的で信頼性が高く、真に人間的な意味での「知的」あるいは「創造的」な AI は、依然として発展途上にあると言える。「思考する」という新しいパラダイムは有望であるが、その成熟にはさらなる時間と研究が必要である。

C. 生成 AI の将来展望に関する総括

生成 AI の分野は、急速なイテレーションと激しい競争によって特徴づけられており、これが技術革新を推進する一方で、徹底的な評価と責任ある展開という点では課題も生じさせている。将来的には、マルチモーダルな情報をより深く統合し、自律的なエージェントとして機能する AI システムが主流となり、人間と機械の能力の境界線をさらに曖昧にしていくと予想される。この進化の過程においては、継続的な批判的研究、倫理的警戒、そして社会全体の適応が不可欠となる。

現在の推論重視型 LLM である o1、o3、Gemini 2.5 Pro は、AI がますます複雑な認知的タスクに取り組めることを示す重要な転換点を示している。しかし、その高度な推論力と創造的可能性を社会の構造にシームレスに、かつ堅牢で信頼性が高く、倫理的に調和した形で統合する道のりは、まだ初期から中期段階にあり、今後も多大な研究開発努

力が求められる。これらのモデルは強力で変革的であり、新しい時代を象徴しているが、完成品ではない。「推論型」あるいは「思考重視型」モデルは明確な前進であるが、AGI への道、あるいは単に全ての領域で非常に信頼性の高い超人的アシスタントへの道はまだ遠く、スケーリングやモデルに「より長く考えさせる」ことだけでは解決できない困難な問題を克服する必要がある。

引用文献

1. OpenAI o1 - Wikipedia, 6 月 8, 2025 にアクセス、https://en.wikipedia.org/wiki/OpenAI_o1
2. Introducing OpenAI o3 and o4-mini, 6 月 8, 2025 にアクセス、https://openai.com/index/introducing-o3-and-o4-mini/
3. the o1 model is just strongly watered down version of o1-preview, and it sucks. - Reddit, 6 月 8, 2025 にアクセス、https://www.reddit.com/r/OpenAI/comments/1h8uf4k/the_o1_model_is_just_strongly_watered_down/
4. o1 pro mode is pathetic. : r/OpenAI - Reddit, 6 月 8, 2025 にアクセス、https://www.reddit.com/r/OpenAI/comments/1hmkrrf/o1_pro_mode_is_pathetic/
5. O3 and o4-mini are extremly bad lazy and not suitable for coding anymore - ChatGPT, 6 月 8, 2025 にアクセス、https://community.openai.com/t/o3-and-o4-mini-are-extremly-bad-lazy-and-not-suitable-for-coding-anymore/1236706
6. o3/o4-mini Function Calling Guide | OpenAI Cookbook, 6 月 8, 2025 にアクセス、https://cookbook.openai.com/examples/o-series/o3o4-mini_prompting_guide
7. Google's top AI, Gemini 2.5 Pro, just got a final, powerful upgrade - Chrome Unboxed, 6 月 8, 2025 にアクセス、https://chromeunboxed.com/googles-top-ai-gemini-2-5-pro-just-got-a-final-powerful-upgrade/
8. Gemini 2.5: Our most intelligent models are getting even better - Google Blog, 6 月 8, 2025 にアクセス、https://blog.google/technology/google-deepmind/google-gemini-updates-io-2025/
9. OpenAI o3 vs. Gemini 2.5 Pro vs. o4-mini - Composio, 6 月 8, 2025 にアクセス、https://composio.dev/blog/openai-o3-vs-gemini-2-5-pro-openai-o4-mini/
10. I compared o3 and o4-mini with Gemini 2.5 Pro: o3 is great but Gemini is better : r/OpenAI - Reddit, 6 月 8, 2025 にアクセス、https://www.reddit.com/r/OpenAI/comments/1k59orl/i_compared_o3_and_o4mini_with_gemini_25_pro_o3_is/
11. Gemini 2.5 Creative writing vs. filtering - Google AI Studio, 6 月 8, 2025 にアクセス、https://discuss.ai.google.dev/t/gemini-2-5-creative-writing-vs-filtering/75171
12. Gemini 2.5 Pro is great, it's just doesn't work, 6 月 8, 2025 にアクセス、https://discuss.ai.google.dev/t/gemini-2-5-pro-is-great-its-just-doesnt-

work/82051
13. Learning to reason with LLMs | OpenAI, 6 月 8, 2025 にアクセス、 https://openai.com/index/learning-to-reason-with-llms/
14. Gemini 2.5: Our most intelligent AI model - Google Blog, 6 月 8, 2025 にアクセス、 https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/
15. 20 LLM evaluation benchmarks and how they work - Evidently AI, 6 月 8, 2025 にアクセス、 https://www.evidentlyai.com/llm-guide/llm-benchmarks
16. Avoiding Common Pitfalls in LLM Evaluation - HoneyHive, 6 月 8, 2025 にアクセス、 https://www.honeyhive.ai/post/avoiding-common-pitfalls-in-llm-evaluation
17. Assessing and Understanding Creativity in Large Language Models - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/html/2401.12491v1
18. Evaluating Creative Short Story Generation in Humans and Large Language Models - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/html/2411.02316v4
19. Introducing OpenAI o1-preview, 6 月 8, 2025 にアクセス、 https://openai.com/index/introducing-openai-o1-preview/
20. Google's new Gemini 2.5 Pro update raises the bar again - The Rundown AI, 6 月 8, 2025 にアクセス、 https://www.therundown.ai/p/googles-gemini-update-raises-the-bar
21. Gemini Deep Research — your personal research assistant - Google Gemini, 6 月 8, 2025 にアクセス、 https://gemini.google/overview/deep-research/
22. OpenAI o1 System Card | OpenAI, 6 月 8, 2025 にアクセス、 https://openai.com/index/openai-o1-system-card/
23. OpenAI o3-mini, 6 月 8, 2025 にアクセス、 https://openai.com/index/openai-o3-mini/
24. LLM Benchmarks: Overview, Limits and Model Comparison - Vellum AI, 6 月 8, 2025 にアクセス、 https://www.vellum.ai/blog/llm-benchmarks-overview-limits-and-model-comparison
25. Claude 4 Opus vs Gemini 2.5 Pro vs OpenAI o3 | Full Comparison - Leanware, 6 月 8, 2025 にアクセス、 https://www.leanware.co/insights/claude-opus4-vs-gemini-2-5-pro-vs-openai-o3-comparison
26. [2401.12491] Assessing and Understanding Creativity in Large Language Models - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/abs/2401.12491
27. The Torrance Test of Creative Thinking: Exploring the Assessment of Innovative Thought, 6 月 8, 2025 にアクセス、 https://www.jamestaylor.me/the-torrance-test-of-creative-thinking-exploring-the-assessment-of-innovative-thought/
28. Creative professions at risk ? Impact and adoption of generative AI - Anthem Creation, 6 月 8, 2025 にアクセス、 https://anthemcreation.com/en/artificial-intelligence/creative-professions-in-danger-how-generative-ai-is-reshaping-the-future-of-creation/
29. Beyond Memorization: Mapping the Originality-Quality Frontier of Language Models - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/html/2504.09389v1

30. LLM-based NLG Evaluation: Current Status and Challenges - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/html/2402.01383v2
31. A LLM-Based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation - ACL Anthology, 6 月 8, 2025 にアクセス、 https://aclanthology.org/2024.findings-emnlp.559.pdf
32. Evaluating Large Language Models for Narrative Topic Labeling - ACL Anthology, 6 月 8, 2025 にアクセス、 https://aclanthology.org/2025.nlp4dh-1.25.pdf
33. Measuring Information Distortion in Hierarchical Ultra long Novel Generation: The Optimal Expansion Ratio - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/html/2505.12572v1
34. Software Development Life Cycle Perspective: A Survey of Benchmarks for Code Large Language Models and Agents - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/html/2505.05283v2
35. Examining the Use and Impact of an AI Code Assistant on Developer Productivity and Experience in the Enterprise - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/html/2412.06603v2
36. Is the GPQA score of o3 really worse than Gemini 2.5 Pro? : r/OpenAI - Reddit, 6 月 8, 2025 にアクセス、 https://www.reddit.com/r/OpenAI/comments/1k3oxco/is_the_gpqa_score_of_o3_really_worse_than_gemini/
37. Understanding and Benchmarking Artificial Intelligence: OpenAI's o3 Is Not AGI - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/abs/2501.07458
38. Understanding and Benchmarking Artificial Intelligence: OpenAI's o3 Is Not AGI - arXiv, 6 月 8, 2025 にアクセス、 https://arxiv.org/html/2501.07458v1
39. OpenAI o3 and o4-mini System Card, 6 月 8, 2025 にアクセス、 https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf
40. Thinking with images | OpenAI, 6 月 8, 2025 にアクセス、 https://openai.com/index/thinking-with-images/
41. OpenAI's o3 Model Scores Lower on FrontierMath Benchmark Than Initially Claimed, 6 月 8, 2025 にアクセス、 https://theoutpost.ai/news-story/open-ai-s-o3-model-scores-lower-on-frontier-math-benchmark-than-initially-claimed-14470/
42. O4-mini-high and o3 context window issue and high hallucination rate - Bugs, 6 月 8, 2025 にアクセス、 https://community.openai.com/t/o4-mini-high-and-o3-context-window-issue-and-high-hallucination-rate/1236872
43. Vibe Check: o3 Is Here—And It's Great - Every, 6 月 8, 2025 にアクセス、 https://every.to/chain-of-thought/vibe-check-o3-is-out-and-it-s-great
44. Gemini models | Gemini API | Google AI for Developers, 6 月 8, 2025 にアクセス、 https://ai.google.dev/gemini-api/docs/models
45. Gemini 2.5 Pro | Generative AI on Vertex AI - Google Cloud, 6 月 8, 2025 にアクセス、 https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-

5-pro
46. OpenAI o3 vs. Gemini 2.5 vs. OpenAI o4-Mini on Coding - DEV Community, 6 月 8, 2025 にアクセス、 https://dev.to/composiodev/openai-o3-vs-gemini-25-vs-openai-o4-mini-5ej4
47. Everyone can now try Gemini 2.5 Pro - for free - ZDNET, 6 月 8, 2025 にアクセス、 https://www.zdnet.com/article/everyone-can-now-try-gemini-2-5-pro-for-free/
48. Gemini 2.5 Pro 05-06 Creative writing is fantastic - Google AI Developers Forum, 6 月 8, 2025 にアクセス、 https://discuss.ai.google.dev/t/gemini-2-5-pro-05-06-creative-writing-is-fantastic/82620
49. Advanced audio dialog and generation with Gemini 2.5 - Google Blog, 6 月 8, 2025 にアクセス、 https://blog.google/technology/google-deepmind/gemini-2-5-native-audio/
50. Gemini 2.5 Pro TTS is... dangerously powerful. I wasn't ready : r/Bard - Reddit, 6 月 8, 2025 にアクセス、 https://www.reddit.com/r/Bard/comments/1ks7e9c/gemini_25_pro_tts_is_dangerously_powerful_i_wasnt/
51. Gemini 2.5 Pro/Flash performs extremely poorly with context and especially documents, 6 月 8, 2025 にアクセス、 https://support.google.com/gemini/thread/347613789/gemini-2-5-pro-flash-performs-extremely-poorly-with-context-and-especially-documents?hl=en
52. Google flexes AI muscle with Gemini 2.5 Pro updates - who doesn't love higher prompt limits? | ZDNET, 6 月 8, 2025 にアクセス、 https://www.zdnet.com/article/google-flexes-ai-muscle-with-gemini-2-5-pro-updates-who-doesnt-love-higher-prompt-limits/
53. o3 vs o4-mini vs Gemini 2.5 pro: The Ultimate Reasoning Battle - Analytics Vidhya, 6 月 8, 2025 にアクセス、 https://www.analyticsvidhya.com/blog/2025/04/o3-vs-o4-mini-vs-gemini-2-5/
54. I tested ChatGPT vs Midjourney V7 with 7 AI image prompts — it wasn't even close | Tom's Guide - Reddit, 6 月 8, 2025 にアクセス、 https://www.reddit.com/r/midjourney/comments/1k5ryc2/i_tested_chatgpt_vs_midjourney_v7_with_7_ai_image/
55. DALL-E 3 vs Midjourney 6/7: Side-by-Side Comparison - Aiarty Image Enhancer, 6 月 8, 2025 にアクセス、 https://www.aiarty.com/midjourney-guide/dalle-vs-midjourney.htm
56. Accelerating scientific breakthroughs with an AI co-scientist - Google Research, 6 月 8, 2025 にアクセス、 https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/
57. How we're using AI to drive scientific research with greater real-world benefit - Google Blog, 6 月 8, 2025 にアクセス、 https://blog.google/technology/research/google-research-scientific-discovery/
58. Hypothesis Maker | AI-powered research hypothesis generator | HyperWrite AI

Writing Assistant, 6 月 8, 2025 にアクセス、 https://www.hyperwriteai.com/aitools/hypothesis-maker
59. How Generative AI Will Impact the Future of Work - Workday Blog, 6 月 8, 2025 にアクセス、 https://blog.workday.com/en-gb/how-generative-ai-will-impact-the-future-work.html
60. How Large Language Models (LLMs) Are Revolutionizing the Legal Industry | Jan 31, 2025, 6 月 8, 2025 にアクセス、 https://ioni.ai/post/how-large-language-models-llms-are-revolutionizing-the-legal-industry
61. Understanding and Utilizing Legal Large Language Models - Clio, 6 月 8, 2025 にアクセス、 https://www.clio.com/resources/ai-for-lawyers/legal-large-language-models/
62. How Generative AI Is Shaping the Future of Law: Challenges and Trends in the Legal Profession - Thomson Reuters Institute, 6 月 8, 2025 にアクセス、 https://www.thomsonreuters.com/en-us/posts/innovation/how-generative-ai-is-shaping-the-future-of-law-challenges-and-trends-in-the-legal-profession/
63. AI is the future of law, but most legal pros aren't trained for it, new report says - ABA Journal, 6 月 8, 2025 にアクセス、 https://www.abajournal.com/web/article/ai-is-the-future-of-law-but-most-legal-pros-arent-trained-for-it-a-new-report-says