主要 AI モデル比較分析：ChatGPT Pro (o3/GPT-4o)、Gemini Ultra (Gemini 2.5 Pro with Deep Think) 、Claude Max (Claude 4 Opus) – パフォーマンス、ベンチマーク、機能 (2025 年 5 月時点)

Gemini Deep Research

1. エグゼクティブサマリー

本レポートは、2025 年 5 月時点における主要 AI プロバイダーのプレミアムサービス、すなわち OpenAI の「ChatGPT Pro」(主要モデル o3 および GPT-4o)、Google の「Gemini Ultra」(Gemini 2.5 Pro with Deep Think 搭載)、Anthropic の「Claude Max」(Claude 4 Opus 搭載)について、詳細な比較分析を提供するものです。分析の焦点は、コアパフォーマンス、各種ベンチマーク結果、技術仕様、およびプランの機能に置かれています。

これらの AI サービスは、それぞれ独自の特徴と強みを示しています。OpenAI は、汎用性の高いマルチモーダルモデルである GPT-4o と、より深い推論能力に特化した o シリーズ(特に o3)という 2 つの戦略的モデルを提供しています。GPT-4o は高速な API 応答と幅広い能力を特徴とし、o3 はより複雑な問題解決に優れる一方で、レイテンシが長くなる可能性があります。Google の Gemini 2.5 Pro with Deep Think は、巨大なコンテキストウィンドウ、ビデオを含む強力なマルチモーダル機能、そして複雑な問題解決のための明示的な「Deep Think」モードを前面に押し出しています。Anthropic の Claude 4 Opus は、コーディングと長時間タスクの実行におけるリーダーとしての地位を確立しており、「ハイブリッド推論」アプローチとエンタープライズ向けの堅牢なパフォーマンスを特徴としています。

AI の分野は急速な進歩を特徴としており、本レポートは主に 2025 年 4 月から 5 月にかけて入手可能であった情報およびデータに基づいています。モデルのバージョンやその性能は頻繁に更新される可能性がある点にご留意ください。

表 1：主要モデルおよびプラン概要

特徴項目	ChatGPT Pro (OpenAI)	Gemini Ultra (Google)	Claude Max (Anthropic)

主要モデル	o3, GPT-4o	Gemini 2.5 Pro with Deep Think	Claude 4 Opus
アーキテクチャ上の主要特徴	o3: 推論特化, GPT-4o: ネイティブマルチモーダル	Deep Think, 超巨大コンテキスト	ハイブリッド/拡張推論
月額料金 (USD)	$200 [1]	$249.99 [3]	$100 (5x Pro) / $200 (20x Pro) [4]
特筆すべきベンチマーク分野	o3: 数学/科学推論, GPT-4o: 汎用性, マルチモーダル	長文コンテキスト, ビデオ理解, 複雑な数学 (USAMO)	コーディング (SWE-bench), 長時間タスク実行
最大コンテキストウィンドウ (API)	o3: 200K [6], GPT-4o: 1M (限定) / 128K[7]	1M (2M 予定) [9]	200K [11]
主要モダリティ	テキスト, 画像, 音声 (GPT-4o) [7]	テキスト, 画像, 音声, ビデオ [9]	テキスト, 画像 [11]
知識カットオフ	GPT-4o: 2023 年 10 月 [16], o3: Web 検索可 [17]	2025 年 1 月 [9]	2025 年 3 月 [12]

この表は、各 AI サービスの核心的な違いと強みを一目で把握できるようにするためのものです。詳細な分析に入る前に、この概要が読者の理解を助け、レポート全体のロードマップとして機能することを意図しています。

2. 各社 AI モデルの紹介

2.1. OpenAI: ChatGPT Pro (o3 および GPT-4o)

OpenAI の「ChatGPT Pro」プランは月額$200 で提供され、同社の最も先進的なモデル群へのアクセスを可能にします [1]。これには、汎用性の高い GPT-4o と、専門的な推論モデルである o3 が含まれます。

GPT-4o ("omni") は 2024 年 5 月にリリースされ [16]、テキスト、画像、音声処理に優

れたネイティブマルチモーダルモデルとして設計されています。従来の GPT-4 世代と比較して、応答時間が速く、コスト効率も向上しており [7]、人間のようなインタラクションを目指しています。

一方、**o3 ("thinking model")** は 2025 年 4 月にリリースされた OpenAI の最も強力な「推論」モデルです [5]。このモデルは、応答を生成する前に「より長く考える」ように設計されており、特にコーディング、数学、科学といった分野で複雑な問題を解決する能力が高められています。また、画像を推論プロセスに統合する能力も特徴です [17]。ChatGPT インターフェース内では、エージェントのようにツールを自律的に使用することも可能です [20]。

知識のカットオフ時点は、GPT-4o が 2023 年 10 月です [16]。o3 については明確なカットオフ日は示されていませんが、ツールとして Web 検索機能を利用できます [17]。

2.2. Google: Gemini Ultra (Gemini 2.5 Pro with Deep Think)

Google の「AI Ultra」プランは月額$249.99 で提供され [3]、最上位モデルである Gemini 2.5 Pro、特にその「Deep Think」強化推論モードへのアクセスを提供します。

Gemini 2.5 Pro with Deep Think は 2025 年 5 月頃に発表され [3]、非常に複雑なタスク向けに設計されています。このモデルは、巨大なコンテキストウィンドウと、テキスト、コード、画像、音声、ビデオ入力に対応するネイティブなマルチモーダル性を活用します [9]。「Deep Think」モードは、応答前に複数の仮説を検討することを可能にし、より高度な推論を実現します [21]。

知識のカットオフ時点は 2025 年 1 月です [9]。

2.3. Anthropic: Claude Max (Claude 4 Opus)

Anthropic の「Claude Max」プランは、例えば Pro プランの 20 倍の使用量で月額 $200 といった価格設定があり [5]、同社の最も知的なモデルである Claude 4 Opus へのアクセスを提供します。

Claude 4 Opus は 2025 年 5 月に発表され [11]、「ハイブリッド推論」モデルとして、複雑なコーディング、エージェント的な検索、長時間のタスク実行に優れています。「拡張思考 (extended thinking) 」モードにより、より深い分析が可能です [11]。ローカルファイルへのアクセス権を与えられた場合、強力な記憶能力を発揮します [24]。

知識のカットオフ時点は 2025 年 3 月です [12]。

これら 3 社のフラッグシップモデルは、「思考」や「推論」のステップを明示的、あるいは制御可能にする方向へと進化している点が注目されます。OpenAI の o シリーズが「立ち止まって考える」能力を特徴とし[17]、Gemini が「Deep Think」モードを搭載し[21]、Claude が「拡張思考」を提供する[11]のは、この潮流の現れです。初期の LLM はパターンマッチングや次トークン予測に長けていましたが、複雑な多段階の問題解決には課題がありました。より信頼性が高く堅牢な問題解決能力への需要と研究の進展が、モデルが問題を分解するためにより多くの「時間」や計算ステップを取ることを可能にするアーキテクチャや技術（内部的な思考連鎖プロンプティングや専用の推論モードなど）の開発を促しました。これにより、o3 の AIME や GPQA での成績[27]、Gemini 2.5 Pro Deep Think の USAMO での成果[21]、Claude 4 Opus の SWE-bench での優位性[24]など、複雑なベンチマークでの性能向上が見られます。しかし、この「思考」プロセスは追加の計算コストとレイテンシを伴います[10]。この傾向は、AI がより高度な認知タスクを実行できるようになる一方で、速度、コスト、そしてより詳細なプロンプティング（例：「慎重に考えて」[17]）や「思考バジェット」の制御[14]に関する新たなトレードオフをユーザーにもたらすことを意味します。これらの思考ステップの可視化（例：Claude のサマリー[22]、Gemini の思考サマリー[21]）は、透明性と信頼性の向上も目的としています。

3. コアパフォーマンス詳細分析：ベンチマーク比較

このセクションでは、各モデルの性能を主要なベンチマークスコアに基づいて詳細に比較します。ベンチマークスコアは特定時点のスナップショットであり、テスト方法やモデルのバージョンによって変動する可能性があること、またモデルは急速に進化することに留意が必要です。可能な限り、ベンチマーク実施日やモデルのバージョン、設定を明記します。

3.1. 一般認知能力・推論・知識ベンチマーク

MMLU (Massive Multitask Language Understanding):
広範な専門知識と問題解決能力を測定するこのベンチマークでは、各社フラッグシップモデルが高いレベルで競合しています。

- **ChatGPT Pro (GPT-4o):** 88.7%[29]。旧 GPT-4 は 86.4% (5-shot)[32]。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 88.6% (Global MMLU Lite)[21]。Gemini 2.5 Pro (Mar '25 Preview) は 0.858 (85.8%相当)[35]。旧 Gemini Ultra (1.0) は 90.0%[36]。
- **Claude Max (Claude 4 Opus):** 88.8%[37]。拡張思考なしでは 87.4%[24]。旧 Claude 3 Opus は 86.8%[32]。
    - **分析:** 3 モデルとも非常に高いスコアを示しており、広範な知識と多タスク言

語理解能力を有しています。最新の比較では GPT-4o と Claude 4 Opus がわずかに優位にあるように見えます。

GPQA (Graduate-Level Google-Proof Q&A):
大学院レベルの複雑でニュアンスのある質問への対応能力を評価します。

- **ChatGPT Pro (GPT -4o):** 53.6% [29]。OpenAI o3 は GPQA Diamond で 83.3% [27]。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 83.0% (single attempt) [21]。Gemini 2.5 Pro (Mar '25 Preview) は 84.0% (pass@1)でリード [18]。
- **Claude Max (Claude 4 Opus):** 79.6% (拡張思考ありで 83.3%) [24]。
  - **分析:** Gemini 2.5 Pro と Claude 4 Opus (拡張思考あり)、そして OpenAI の o3 は、この難易度の高い大学院レベルの推論問題で非常に強力な性能を示し、GPT-4o を大幅に上回っています。これは専門的な推論能力の影響を浮き彫りにしています。

HellaSwag (Commonsense Reasoning):
日常的な出来事に関する常識的推論能力を評価します。

- **ChatGPT Pro (GPT -4 旧モデル):** 95.3% (10-shot) [30]。 (o3/GPT-4o の特定スコアは見当たらず)
- **Gemini Ultra (Gemini 1.0):** 87.8% [36]。 (Gemini 2.5 Pro のスコアは見当たらず)
- **Claude Max (Claude 4 Opus):** 具体的なスコアは見当たらず。HellaSwag-Pro という新しいベンチマークが導入され、GPT-4o はそのバリアントで失敗するとの報告あり [41]。
  - **分析:** 最新モデル間での直接比較データは限定的です。旧 GPT-4 が高い基準を設定しました。HellaSwag-Pro の知見 [41] は、GPT-4o のようなトップモデルでさえ、常識的推論の堅牢性にはまだ改善の余地があることを示唆しています。

DROP (Discrete Reasoning Over Paragraphs):
文章からの情報抽出と離散的推論能力を評価します。

- **ChatGPT Pro (GPT -4o):** 83.4% (F1 スコア) [29]。旧 GPT-4 は 80.9 (3-shot) [30]。
- **Gemini Ultra (Gemini 1.0):** 82.4 [36]。 (Gemini 2.5 Pro のスコアは見当たらず)
- **Claude Max (Claude 4 Opus):** 具体的なスコアは見当たらず。
  - **分析:** GPT-4o は強力な性能を示しています。Gemini 2.5 Pro および Claude 4 Opus との DROP に関する直接比較データは、提供された情報からは欠落しています。

ARC (AI2 Reasoning Challenge):
AI の推論能力を測るベンチマークです。特に ARC-AGI はより困難なセットです。

- **ChatGPT Pro (GPT -4 旧モデル):** ARC-Challenge で 96.3% (25-shot) [32]。GPT-4o

(Base LLM) は ARC-AGI-1 で 4.5%、ARC-AGI-2 で 0.0% [42]。o3 (Medium) は ARC-AGI-1 で 53.0%、ARC-AGI-2 で 3.0% [42]。o3 の高計算量設定では ARC AGI で 88%という報告もあるが [27]、後の分析では o3 (Medium) が ARC-AGI-1 で 53%、ARC-AGI-2 で 3%とされています [43]。

- **Gemini Ultra (Gemini 2.5 Flash Preview, Thinking 24K):** ARC-AGI-1 で 32.3%、ARC-AGI-2 で 2.5% [42]。 (Gemini 2.5 Pro の ARC-AGI スコアは見当たらず)
- **Claude Max (Claude 4 Opus, Thinking 16K):** ARC-AGI-1 で 35.7%、ARC-AGI-2 で 8.6% [42]。
    - **分析:** ARC-AGI ベンチマーク、特に ARC-AGI-2 は全モデルにとって大きな課題であり、人間の能力には遠く及ばないことが示されています [43]。Claude Opus 4 (Thinking 16K) が 3 モデル中では ARC-AGI-2 で最高のスコアを示していますが、ARC-AGI-1 では o3 (Medium) がリードしています。o3 の ARC スコアの不一致 [27] は、ベンチマークのバージョンや設定の違いを浮き彫りにしています。GPT-4o の ARC-AGI での低いスコア [42] は、他の強力な一般ベンチマークと比較して注目されます。

SimpleQA (Factuality):
事実に基づいた回答の正確性を評価します。

- **ChatGPT Pro (GPT -4.5):** 精度 62.5%、ハルシネーション率 37.1% [45]。o3: 精度 49%、ハルシネーション率 51% [46]。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 50.8% [21]。Gemini 2.5 Pro (Mar '25 Preview): 52.9% [18]。
- **Claude Max (Claude 4 Opus):** 具体的なスコアは見当たらず。
    - **分析:** 事実性とハルシネーションの低減は依然として主要な課題です。Gemini 2.5 Pro と GPT-4.5 は競争力のあるスコアを示しています。o3 の精度に対するハルシネーション率の高さは懸念材料として指摘されています [46]。

表 2.1：一般認知能力・推論ベンチマーク

ベンチマーク	ChatGPT Pro (o3)	ChatGPT Pro (GPT-4o)	Gemini Ultra (Gemini 2.5 Pro + Deep Think)	Claude Max (Claude 4 Opus)	備考

MMLU	N/A (GPT-4o が代表)	88.7% [29]	88.6% (Global MMLU Lite, May '25) [21]	88.8% / 87.4% (拡張思考なし) [25]	各社トップレベルで僅差
GPQA Diamond	83.3% [27]	53.6% [29]	83.0% (May '25, single) / 84.0% (Mar '25, p@1) [18]	79.6% / 83.3% (拡張思考あり) [25]	o3, Gemini, Claude (拡張思考) が GPT-4o を上回る
HellaSwag	N/A (GPT-4 旧版: 95.3% [39])	N/A	N/A (Gemini 1.0: 87.8% [36])	N/A	最新モデルでの直接比較データ不足
DROP (F1)	N/A (GPT-4 旧版: 80.9 [32])	83.4% [29]	N/A (Gemini 1.0: 82.4 [36])	N/A	GPT-4o が良好なスコア、他はデータ不足
ARC-AGI-1	53.0% (Medium) [42]	4.5% (Base) [42]	N/A (Flash: 32.3% [42])	35.7% (Thinking 16K) [42]	o3 がリード、全体的に低スコア
ARC-AGI-2	3.0% (Medium) [42]	0.0% (Base) [42]	N/A (Flash: 2.5% [42])	8.6% (Thinking 16K) [42]	Claude がリードするも、全モデル苦戦。人間レベルには程遠い [43]
SimpleQA (Accuracy)	49% (ハルシネーション 51%) [46]	N/A (GPT-4.5: 62.5% [45])	50.8% (May '25) / 52.9% (Mar '25) [21]	N/A	Gemini と GPT-4.5 が競争力あり、o3 はハルシネーションに課題

3.2. 数学的・科学的推論ベンチマーク

GSM8K (Grade School Math):
小学校レベルの算数文章問題を評価します。

- **ChatGPT Pro (GPT-4o):** 90.5% [31]。 (旧 GPT-4 は 92.0% 5-shot CoT [33])。o3 (high): 96.7% / 96.9% [47]。
- **Gemini Ultra (Gemini 2.5 Pro):** 直接的なスコアは見当たらず。
- **Claude Max (Claude 4 Opus):** 約 95% (推定) [48]。
    - **分析:** o3 と Claude 4 Opus が非常に高い性能を示し、GPT-4o を上回る可能性があります。このレベルでの強力な算術および多段階推論能力を示唆しています。

MATH (Mathematical Problem Solving):
より難易度の高い数学コンテストレベルの問題を評価します。

- **ChatGPT Pro (GPT-4o):** 76.6% [29]。o3: AIME スコアは高いが MATH スコア直接記載なし。
- **Gemini Ultra (Gemini 2.5 Pro):** MATH スコア直接記載なし。AIME/USAMO スコアは高い。
- **Claude Max (Claude 4 Opus):** AIME 2025: 75.5% (高計算量設定で 90.0%) [37]。旧 Claude 3 Opus: 60.1% (0-shot CoT) [32]。
    - **分析:** GPT-4o は強力な MATH スコアを持っています。Claude 4 Opus は、特に高計算量設定で非常に高い AIME スコアを示します。o3 も AIME で優れています (AIME 2024 で 91.6%、AIME 2025 で 88.9% [27])。MATH ベンチマークでの直接比較は不完全です。

AIME (American Invitational Mathematics Examination):
高校生向けの数学コンテストです。

- **ChatGPT Pro (o3):** 91.6% (AIME 2024), 88.9% (AIME 2025)[27]。Python ツール使用時: 98.4% (AIME 2025)[17]。o4-mini (Python ツール使用時): 99.5% (AIME 2025) [5]。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 83.0% (AIME 2025 single attempt) [21]。Gemini 2.5 Pro (Mar '25 Preview): 92.0% (AIME 2024), 86.7% (AIME 2025) [18]。
- **Claude Max (Claude 4 Opus):** 75.5% (AIME 2025), 90.0% (AIME 2025 高計算量設定) [37]。
    - **分析:** 全てのモデルが強力な数学コンテストの成績を示しています。OpenAI の o シリーズは、特にツールを使用した場合、ほぼ完璧なスコアに達します。Claude 4 Opus (高計算量設定) と Gemini 2.5 Pro も非常に競争力があります。

USAMO (United States of America Mathematical Olympiad):
非常に難易度の高い数学オリンピックです。

- **Gemini Ultra (Gemini 2.5 Pro with Deep Think):** 2025 USAMO で「印象的なスコア」[21]、49.4% [50]。
  - **分析:** USAMO は非常に挑戦的なベンチマークです。Gemini の Deep Think がここで特に強調されています。

GPQA Diamond (PhD-level science questions):
博士課程レベルの科学に関する質問で、前述の一般推論 GPQA と同じものです。

- **ChatGPT Pro (o3):** 83.3% [27]。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 83.0% [21]。Gemini 2.5 Pro (Mar '25 Preview): 84.0% [18]。
- **Claude Max (Claude 4 Opus):** 79.6% (拡張思考ありで 83.3%) [24]。
  - **分析:** 3 モデルとも非常に高いスコアを達成しており、高度な科学的推論能力を示しています。Gemini 2.5 Pro と o3/Claude 4 Opus (拡張思考) は僅差です。

表 2.2：数学的・科学的推論ベンチマーク

ベンチマーク	ChatGPT Pro (o3)	ChatGPT Pro (GPT-4o)	Gemini Ultra (Gemini 2.5 Pro + Deep Think)	Claude Max (Claude 4 Opus)	備考
GSM8K	96.7%-96.9% (high) [47]	90.5% [31]	N/A	~95% (est.) [48]	o3 と Claude Opus が GPT-4o を上回る可能性
MATH	N/A (AIME スコアは高い)	76.6% [29]	N/A (AIME/USAMO スコアは高い)	N/A (AIME スコアは高い, 旧 Opus 3: 60.1% [32])	GPT-4o が強力。他モデルは AIME 等で代替評価
AIME 2025	88.9% (ツール使用時	N/A	83.0% (May '25 single) /	75.5% (高計算量時	各モデル高スコア。o3 は

	98.4%) [17]		86.7% (Mar '25) [18]	90.0%) [37]	ツール使用でほぼ満点
USAMO 2025	N/A	N/A	49.4% (Deep Think) [50]	N/A	Gemini Deep Think が特筆される超難関ベンチマーク
GPQA Diamond	83.3% [27]	N/A	83.0% (May '25) / 84.0% (Mar '25) [18]	79.6% (拡張思考時 83.3%) [25]	3 モデルとも僅差で高スコア

3.3. コーディングおよびソフトウェアエンジニアリング能力

HumanEval (Python coding tasks):
Python のコーディングタスクの基本的な能力を評価します。

- **ChatGPT Pro (GPT-4o):** 90.2% [29]。旧 GPT-4 は 67.0% (0-shot) [32]。o3 の直接スコアはないが、SWE-bench は高い。[62] では GPT-4 が HumanEval で 86.6%、o3 が SWE-bench で 71.7%と記載。[54] では o3-mini が KotlinHumanEval で 86.96%。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 75.6% [51]。
- **Claude Max (Claude 4 Opus):** 直接的な HumanEval スコアはこれらの情報源にはない。旧 Claude 3 Opus は 84.9% (0-shot) [32]。[24] 等は Claude 4 Opus を「世界最高のコーディングモデル」とし、SWE-bench でリードしていると記述。
    - **分析:** GPT-4o が非常に高い HumanEval スコアを示しています。Claude 4 Opus はコーディング能力（特に SWE-bench）で称賛されていますが、「Max」プランの文脈での直接的な HumanEval スコアはこれらの情報源にはありません。Gemini 2.5 Pro も良好な成績です。

SWE-bench (Software Engineering Benchmark):
実際のソフトウェアエンジニアリングタスクにおける問題解決能力を評価します。

- **ChatGPT Pro (o3):** 69.1% [5]。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 63.2% [10]。
- **Claude Max (Claude 4 Opus):** 72.5% (基本)、79.4% (高計算量設定) [11]。Claude 4 Sonnet: 72.7% (基本)、80.2% (高計算量設定) [4]。
    - **分析:** Claude 4 Opus と Sonnet 4 が SWE-bench で大幅にリードしており、特に高計算量設定では、実際のソフトウェアエンジニアリングタスクにおけるトップ候補としての地位を確立しています。o3 も非常に強力です。

Codeforces (Competitive Coding):
競技プログラミングの能力を ELO レーティングで評価します。

- **ChatGPT Pro (o3):** ELO 2706[5]。o4-mini: ELO 2719[5]。
  - **分析:** OpenAI の o シリーズ、小型の o4-mini を含め、卓越した競技プログラミング能力を示しています。

LiveCodeBench (Competition-level Coding):
競技レベルのコーディング能力を評価します。

- **Gemini Ultra (Gemini 2.5 Pro with Deep Think):** リード[21]。スコア 80.4%[50]。Gemini 2.5 Pro (Mar '25 Preview): 70.4%[18]。
  - **分析:** Gemini 2.5 Pro、特に Deep Think モードがこの要求の厳しいコーディングベンチマークで優れています。

Terminal-bench (CLI-based coding):
コマンドラインインターフェースベースのエージェント的コーディングタスクを評価します。

- **Claude Max (Claude 4 Opus):** 43.2% (基本)、50.0% (高計算量設定)[23]。
- **Gemini Ultra (Gemini 2.5 Pro):** 25.3%[51]。
- **ChatGPT Pro (o3):** 30.2%[50]。
  - **分析:** Claude 4 Opus がエージェント的ターミナルコーディングタスクで強力なリードを示しています。

Aider Polyglot (Code Editing):
複数言語にまたがるコード編集能力を評価します。

- **ChatGPT Pro (o3):** o1 を上回る[27]。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 76.5% (全体) / 72.7% (差分)[21]。Gemini 2.5 Pro (Mar '25 Preview): 74.0%[18]。
- **Claude Max (Claude 4 Opus):** 強力な性能が言及されている[22]。
  - **分析:** Gemini 2.5 Pro と o3 が強力なコード編集能力を示しています。

表 2.3：コーディング・ソフトウェアエンジニアリングベンチマーク

ベンチマーク	ChatGPT Pro (o3)	ChatGPT Pro (GPT-4o)	Gemini Ultra (Gemini 2.5 Pro + Deep Think)	Claude Max (Claude 4 Opus)	備考

HumanEval	N/A (KotlinHumanEval o3-mini: 86.96% [54])	90.2% [29]	75.6% (May '25) [51]	N/A (旧 Opus 3: 84.9% [32])	GPT-4o が高スコア。Claude 4 Opus は SWE-bench で高評価
SWE-bench	69.1% [17]	N/A	63.2% (May '25) [21]	72.5% (基本) / 79.4% (高計算量) [24]	Claude 4 Opus/Sonnet がリード。o3 も強力
Codeforces ELO	2706 [17]	N/A	N/A	N/A	OpenAI o シリーズが非常に高い ELO
LiveCodeBench	N/A	N/A	80.4% (Deep Think) [50]	N/A	Gemini (Deep Think) がリード
Terminal-bench	30.2% [50]	N/A	25.3% [51]	43.2% (基本) / 50.0% (高計算量) [24]	Claude 4 Opus がリード
Aider Polyglot (whole/diff)	o1 を上回る [27]	N/A	76.5%/72.7% (May '25) [21]	強力な性能 [22]	Gemini と o3 が強力

3.4. マルチモーダル能力ベンチマーク

MMMU (Multimodal Understanding):
テキスト、画像など複数のモダリティを統合して理解する能力を評価します。

- **ChatGPT Pro (o3):** 82.9% [5]。
- **Gemini Ultra (Gemini 2.5 Pro with Deep Think):** 84.0% [21]。Gemini 2.5 Pro (May '25 Preview): 79.6% [21]。Gemini 2.5 Pro (Mar '25 Preview): 81.7%[18]。
- **Claude Max (Claude 4 Opus):** 76.5% (validation) [23]。拡張思考なしでは 73.7% [24]。
    - **分析:** Gemini 2.5 Pro (特に Deep Think モード) と OpenAI の o3 が一般的なマ

ルチモーダル推論でリードしています。Claude 4 Opus も競争力がありますが、この特定のベンチマークではわずかに後れを取っています。

Video-MME (Video Understanding):
ビデオコンテンツの理解能力を評価します。

- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 84.8% [21]。
  - **分析:** Gemini 2.5 Pro が強力なビデオ理解能力を示しています。他のモデルに関するデータは不足しています。

Vibe-Eval (Image Understanding):
画像理解能力を評価します (Gemini をジャッジとして使用)。

- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** 65.6% [21]。
  - **分析:** Gemini に特化した評価であり、良好な画像理解を示しています。

CharXiv-Reasoning (Scientific Diagram Comprehension):
科学的な図やグラフの理解能力を評価します。

- **ChatGPT Pro (o3):** 78.6% [27]。o4-mini: 72.0% [5]。
  - **分析:** OpenAI の o シリーズモデルが科学図の理解において強力な性能を示しています。

MathVista (Visual Math Reasoning):
視覚的に提示された数学問題の推論能力を評価します。

- **ChatGPT Pro (o3):** 86.8% [27]。
  - **分析:** o3 が視覚的な数学問題に優れています。

表 2.4：マルチモーダル能力ベンチマーク

ベンチマーク	ChatGPT Pro (o3)	ChatGPT Pro (GPT-4o)	Gemini Ultra (Gemini 2.5 Pro + Deep Think)	Claude Max (Claude 4 Opus)	備考
MMMU	82.9% [17]	N/A	84.0% (Deep Think) [21]	76.5% (validation) [37]	Gemini (Deep Think) と o3 がリード

Video-MME	N/A	N/A	84.8% (May '25) [21]	N/A	Gemini が強力なビデオ理解能力を示す
Vibe-Eval (Reka)	N/A	N/A	65.6% (May '25, Gemini judge) [21]	N/A	Gemini の画像理解指標
CharXiv-Reasoning	78.6% [27]	N/A	N/A	N/A	OpenAIo シリーズが科学図理解に強い
MathVista	86.8% [27]	N/A	N/A	N/A	o3 が視覚的数学推論に優れる
MMNeedle (long context images)	N/A	GPT-4o が他を凌駕 [56]	N/A	N/A	GPT-4o は強力だが、陰性サンプルでハルシネーションの課題あり [56]

3.5. 長文コンテキスト理解

Needle-in-a-Haystack (NIAH) / MRCR (Multi-Document Reading Comprehension):
大量のテキストの中から特定の情報を探し出す能力を評価します。

- **ChatGPT Pro (GPT -4o):** MMNeedle ベンチマークでは、GPT-4o が長文コンテキストの画像シナリオで他モデルを上回るものの、陰性サンプル（情報が存在しない場合）でのハルシネーション問題が指摘されています [56]。
- **Gemini Ultra (Gemini 2.5 Pro May '25 Preview):** MRCR で 93.0% (128k 平均)、82.9% (1M ポイントワイズ) [21]。旧 Gemini 1.5 Pro は 1M トークンまで 99.7%以上のリコール率 [10]。MRCR (128Kコンテキスト) で 91.5% [18]。
- **Claude Max (Claude 4 Opus):** 大規模文書分析における強力な NIAH 能力 (180 ページのレポートでのテスト [57])。旧 Claude 3 Opus は 200K トークンまで 99%以上のリコール率 [58]。
    - **分析:** Gemini 2.5 Pro と Claude モデル (Opus 3 および 4) は、長文コンテキストの想起タスクで優れた性能を発揮し、その大きなコンテキストウィンドウを

活用しています。GPT-4o の MMNeedle での性能は強力ですが、注意点もあります。

定性的文書分析:

- **Gemini Ultra (Gemini 2.5 Pro):** 約 1,500 ページまたは 3 万行のコードを処理可能[59]。
- **Claude Max (Claude 4 Opus):** 数百ページを処理可能[60]。180 ページの Nvidia レポートでストレステスト実施[57]。
- **ChatGPT Pro (o3):** Python を使用してアップロードされたファイルを分析可能[17]。
    - **分析:** 全モデルが大規模文書を扱えますが、Gemini の 1M トークンウィンドウ (2M 予定) は、非常に大きな入力に対して理論的な優位性をもたらします。

巨大なコンテキストウィンドウ（Gemini は最大 1M-2M トークン 9、Claude/o3 は 200K トークン 6、GPT-4o API は 1M トークン 7）を各社が競って搭載していますが、「Needle-in-a-Haystack」(NIAH)や MRCR のようなベンチマークでの性能は、必ずしもコンテキスト全体にわたる完璧な想起を保証するものではありません 10。
ユーザーがより大きな入力（長文ドキュメント、コードベース全体、長いチャット履歴）の処理を求めるにつれて、モデル開発者はコンテキストウィンドウサイズを拡大しています。理論的には、より大きなコンテキストウィンドウにより、モデルは一貫性を保ち、特定のタスクに対してより多くの情報にアクセスできます。しかし、極端に長いシーケンス全体で注意を維持し、正確な想起を行うことは、計算上およびアーキテクチャ上困難です（「中間での迷子」問題）。NIAH/MRCR のようなベンチマークは、この長文コンテキスト検索能力を具体的にテストするために開発されました。結果として、主要モデルは良好な性能を示すものの、常に完璧ではなく、コンテキストの極端な部分や「妨害」ドキュメントの数が増えると性能が低下する可能性があります。例えば、GPT-4o は MMNeedle で陰性サンプルに苦戦しています 56。さらに、非常に大きなコンテキストの処理は、計算コストとレイテンシの増加を伴います 51。
したがって、ユーザーは、より大きなコンテキストウィンドウが自動的にそのコンテキスト全体での完璧なパフォーマンスを意味するとは限らないことを理解すべきです。コンテキストの「効果的な」使用が鍵となります。大規模コンテキスト処理のコストとレイテンシも重要な考慮事項です。「Claude Max」プランで大きなコンテキストを使用するとメッセージ制限にすぐに達するという事実は 61、このトレードオフを示しています。

表 2.5：長文コンテキスト理解ベンチマーク

ベンチマーク/テスト	ChatGPT Pro (o3/GPT-4o)	Gemini Ultra (Gemini 2.5	Claude Max (Claude 4	備考 (コンテキスト長)

		Pro)	Opus)	
NIAH/MRCR (定量的スコア)	GPT-4o: MMNeedle で他を凌駕するが課題あり [56]	MRCR 93.0% (128k avg), 82.9% (1M pointwise) [21]	NIAH で強力な性能 [57], 旧 Opus 3: >99% (200K) [58]	Gemini: 1M トークン, Claude: 200K トークン
定性的長文読解性能 (例: 最大処理ページ数)	o3: アップロードファイル分析可 [17]	~1,500 ページ / 3 万行コード [59]	数百ページ [60], 180 ページレポート処理 [57]	Gemini が最大の処理量

4. 主要技術仕様と機能

4.1. コンテキストウィンドウ

- **ChatGPT Pro (o3):** API および ChatGPT で 200K トークン [6]。最大出力 100K トークン [6]。
- **ChatGPT Pro (GPT-4o):** ChatGPT インターフェースで 128K トークン、API 経由では 1M トークン（限定リリース）[7]。最大出力 16,384 トークン [8]。一部情報源では API も 128K と記載 [16]。*補足: GPT-4o は 128K コンテキストウィンドウ、最大 16K 出力トークン [8]。GPT-4.1（別モデル）は 1M コンテキスト、32K 出力 [63]。ChatGPT Pro プラン自体は 128K コンテキストウィンドウと記載されているが [64]、これはチャットインターフェースを指す可能性があり、o3 や GPT-4o API のような特定モデルへの API アクセスとは異なる場合がある。[61] では ChatGPT Plus（Pro ではない）の UI が 32K コンテキストウィンドウであると指摘。*
- **Gemini Ultra (Gemini 2.5 Pro):** API (gemini-2.5-pro-preview-05-06) で 1M トークン入力（2M 予定）[9]。最大出力 65,535 トークン [9]。Google AI Ultra プランでは 1,500 ページまたは 3 万行のコードのアップロードに言及 [59]。
- **Claude Max (Claude 4 Opus):** 200K トークン [11]。最大出力 32K トークン [12]。
  - **分析:** 現状、API 経由での一般利用可能な入力コンテキストウィンドウは Gemini 2.5 Pro が最大です。ChatGPT Pro は、o3 (200K) と GPT-4o (API で 128K/1M) という異なるコンテキストサイズのモデルへのアクセスを提供します。Claude 4 Opus は 200K と十分大きなウィンドウを備えています。最大出力トークン数もモデルによって異なり、長文生成タスクには重要な要素です。

4.2. 対応モダリティ

- **ChatGPT Pro (o3):** 入力: テキスト、画像 (「画像で思考する」機能 [5])。出力: テキ

スト。o3 自体のネイティブな音声/ビデオ処理については詳細不明だが、新しい音声モデルは GPT-4o アーキテクチャ上に構築[68]。

- **ChatGPT Pro (GPT -4o):** 入力: テキスト、画像、音声[7]。ビデオ入力も言及あり[51]。出力: テキスト、音声[7]。画像生成は GPT-4o にネイティブ搭載[13]。
- **Gemini Ultra (Gemini 2.5 Pro):** 入力: テキスト、コード、画像、音声、ビデオ[9]。出力: テキスト[9]。ネイティブ音声出力が計画/利用可能[14]。Ultra プランにはビデオ生成用 Veo 3 が含まれる[3]。
- **Claude Max (Claude 4 Opus):** 入力: テキスト、画像[11]。出力: テキスト[15]。Opus 4 のネイティブな音声/ビデオ入出力処理に関する明示的な言及はこれらの情報源にはないが、「マルチモーダル能力」という言葉は一般的に使用されている[24]。システムカードでは「視覚分析」に言及[19]。
    - **分析:** Gemini 2.5 Pro と GPT-4o が最も包括的なマルチモーダル対応を示し、音声やビデオを含む幅広い入力タイプをサポートしています。o3 と Claude 4 Opus は画像入力に強いものの、これらの情報源に基づくと主にテキスト出力に焦点が当てられているようです。

4.3. API パフォーマンス (レイテンシとスループット)

- **ChatGPT Pro (o3 API):**
    - TTFT (最初のトークンまでの時間): 約 13.27 秒～20.80 秒[74] (Artificial Analysis による o3 のデータ)。
    - TPS (1秒あたりのトークン数): 約 136.8～209.6 トークン/秒[74] (Artificial Analysis による o3 のデータ)。
    - *注: o3-mini ははるかに高速: TTFT 85 ミリ秒、TPS 102.7*[75]。
- **ChatGPT Pro (GPT -4o API):**
    - TTFT: 約 0.55 秒～0.58 秒[76] (Artificial Analysis、Nov '24 版)。Replicate.com: 特定の実行で 56 ミリ秒[7]。
    - TPS: 約 99.4～108.5 トークン/秒[76] (Artificial Analysis、Nov '24 版)。Replicate.com: 特定の実行で 15.10 TPS[7]。
- **Gemini Ultra (Gemini 2.5 Pro API):**
    - TTFT: 約 26.8 秒～42.01 秒[35] (Artificial Analysis による Mar '25 & May '25 プレビュー版のデータ)。ユーザー報告では 14～30 TPS[78]。
    - TPS: 約 147.7～212.7 トークン/秒[35] (Artificial Analysis による Mar '25 & May '25 プレビュー版のデータ)。
- **Claude Max (Claude 4 Opus API):**
    - TTFT: 約 2.26 秒～2.99 秒[66] (Artificial Analysis による Opus & Opus 拡張思考のデータ)。Anthropic プロバイダー: TTFT 2.99 秒[79]。

- TPS: 約 54.1〜57 トークン/秒 [66] (Artificial Analysis による Opus & Opus 拡張思考のデータ)。Anthropic プロバイダー: 57 TPS [79]。[48] では Sonnet 4 が 54.84 TPS、Opus 4 が 38.93 TPS と報告。
- **分析:** GPT-4o が最も低い TTFT を誇り、初期出力の応答性に優れています。Gemini 2.5 Pro と o3 は、「思考」プロセスにより TTFT が大幅に長くなりますが、生成開始後の TPS は高くなる可能性があります。Claude 4 Opus は TTFT では中間に位置しますが、TPS は o3 や Gemini 2.5 Pro と比較して低めです。

API パフォーマンスにおけるレイテンシ、スループット、そして「思考」時間のトレードオフは、モデル選択における重要な考慮事項です。o3、Gemini 2.5 Pro with Deep Think、Claude 4 Opus with Extended Thinking のように、明示的または集中的な推論ステップを持つモデルは、一般的に GPT-4o のような速度に最適化されたモデルと比較して、最初のトークンまでの時間（TTFT）が長くなる傾向があります [35]。この初期計算フェーズは TTFT に直接影響し、モデルの即時応答性を低下させます（例：o3 の TTFT は約 13〜20 秒 [74]、Gemini 2.5 Pro の TTFT は約 26〜42 秒 [35]）。しかし、推論が完了すると、その後のトークン生成速度（TPS）は依然として高い場合があります（例：o3 の TPS は約 130〜200 [74]、Gemini 2.5 Pro の TPS は約 140〜210 [35]）。GPT-4o のような、より広範で高速なインタラクションに最適化されたモデルは、はるかに低い TTFT（約 0.5 秒 [76]）を示しますが、非常に複雑なタスクにおいては、「思考型」モデルの推論の深さに必ずしも匹敵するわけではありません。これにより、明確なトレードオフが生じます。即時フィードバックを必要とするアプリケーションには低い TTFT が不可欠ですが、最終的な出力品質が最優先される複雑な問題解決では、より高い TTFT が許容される場合があります。開発者は、アプリケーションのレイテンシ許容度と必要な推論の深さに基づいて API を選択する必要があります。OpenAI の o3/o4-mini 向け「Flex プロセッシング」[52] や Gemini 2.5 Flash の制御可能な「思考バジェット」[14] のような機能は、このトレードオフにおける柔軟性を提供しようとする試みです。一部の「思考型」モデルの TTFT が長いことは、インタラクティブなリアルタイムアプリケーションにとっては、「思考」のストリーミング（o3 の場合 [52]）が知覚レイテンシを軽減しない限り、障壁となる可能性があります。

4.4. 知識カットオフ日

- **ChatGPT Pro (GPT -4o):** 2023 年 10 月 [16]。 (o3: 特定されていませんが、Web 検索機能あり [17])。
- **Gemini Ultra (Gemini 2.5 Pro):** 2025 年 1 月 [9]。
- **Claude Max (Claude 4 Opus):** 2025 年 3 月 [12]。

- **分析:** Claude 4 Opus が最も新しい知識カットオフ日を持ち、次いで Gemini 2.5 Pro です。しかし、全てのモデルが最新情報を提供するために統合された Web 検索/ツール利用（o3 [17]、Gemini [21]、Claude 4 [23]）にますます依存するようになっており、静的なカットオフ日の制約は以前ほどではなくなっています。このツール利用の有効性が新たな差別化要因となっています。

表 3：技術仕様概要

仕様	ChatGPT Pro (o3)	ChatGPT Pro (GPT-4o)	Gemini Ultra (Gemini 2.5 Pro + Deep Think)	Claude Max (Claude 4 Opus)
最大 API コンテキストウィンドウ (入力トークン)	200K [6]	1M (限定) / 128K [7]	1M (2M 予定) [9]	200K [11]
最大 API 出力トークン	100K [6]	16,384 [8]	65,535 [9]	32K [12]
対応入力モダリティ (API/Chat)	テキスト, 画像 [17]	テキスト, 画像, 音声, (ビデオ) [7]	テキスト, コード, 画像, 音声, ビデオ [9]	テキスト, 画像 [11]
対応出力モダリティ (API/Chat)	テキスト	テキスト, 音声, 画像 [7]	テキスト, (音声予定) [9]	テキスト [15]
知識カットオフ日	Web 検索可 [17]	2023 年 10 月 [16]	2025 年 1 月 [9]	2025 年 3 月 [12]

表 4：API パフォーマンス指標 (Artificial Analysis　調べ)

指標	ChatGPT Pro (o3)	ChatGPT Pro (GPT-	Gemini Ultra (Gemini 2.5	Claude Max (Claude 4	備考

		4o Nov '24)	Pro May '25 Preview)	Opus)	
TTFT (秒)	13.27 - 20.80 [74]	0.55 - 0.58 [76]	36.24 - 42.01 [35]	2.26 - 2.99 [66]	GPT-4o が最速、Gemini が最も遅い
TPS (出力)	136.8 - 209.6 [74]	99.4 - 108.5 [76]	147.7 - 212.7 [35]	38.93 - 57 [48]	o3 と Gemini が高速、Claude が遅め

5. プレミアムプラン概要：「Pro」「Ultra」「Max」

5.1. OpenAI ChatGPT Pro

- **料金:** 月額$200 [1]。
- **モデルアクセス:** チャットインターフェースで o3、o4-mini、o4-mini-high、GPT-4o、GPT-4、GPT-4.5 Preview モデルへの「ほぼ無制限」アクセス [64]。o1 pro モードへのアクセス [2]。
- **利用制限 (チャットインターフェース):** Pro ユーザーは「ほぼ無制限」とされていますが、公正利用および乱用防止策の対象となります [82]。比較として、Plus ユーザーには具体的な制限があります：GPT-4o (80 メッセージ/3 時間)、o3 (50 メッセージ/週)、o4-mini (150 メッセージ/日) [64]。Pro ユーザーはこれより大幅に高いか、より緩やかな上限が設定されていると考えられます [83]。
- **API 料金 (主要モデル):**
  - o3: 入力$10.00/1M トークン、出力$40.00/1M トークン [5]。Flex プロセッシングにより、緊急性の低いワークロードに対してより安価なレートが提供されます [52]。
  - GPT-4o: 入力$2.50/1M トークン、出力$10.00/1M トークン [7] (Nov '24 版)。
- **主な機能:** 最も強力なモデルへのアクセス、高い利用制限、優先アクセス、DALL-E 3 や Deep Research (120 クエリ/月 [1])、Operator リサーチプレビュー [2] などの高度なツール。

5.2. Google AI Ultra

- **料金:** 月額$249.99 (約¥36,400) [3]。
- **モデルアクセス:** Gemini 2.5 Pro with Deep Think および Veo 3 への「最高アクセス」[3]。

- **利用制限 (チャットインターフェース):** 「最高アクセス」は Pro プランよりも高い制限を示唆しています[65]。Deep Research の利用は ChatGPT Plus よりも寛大です[84]。Ultra プラン下での Gemini 2.5 Pro の具体的なメッセージ/クエリ数は詳細不明ですが、相当量であると予想されます。Gemini Apps には、プロンプトの複雑さやアップロードファイルサイズなどによって変動する一般的な利用制限があります[65]。
- **API 料金 (Gemini 2.5 Pro Preview 05 -06):**
  - 入力: $1.25/1M トークン (≤200k コンテキスト)、$2.50/1M トークン (>200k コンテキスト)[21]。
  - 出力 (思考トークン含む): $10.00/1M トークン (≤200k コンテキスト)、$15.00/1M トークン (>200k コンテキスト)[21]。
- **主な機能:** Gemini 2.5 Pro + Deep Think、Veo 3 (AI ビデオ)、Flow (AI 映画制作ツール)、Whisk Animate、NotebookLM 強化アクセス、Chrome/Gmail/Docs での Gemini 利用、Project Mariner (AI エージェント)、30 TB の Google One ストレージ、YouTube Premium[3]。

5.3. Anthropic Claude Max

- **料金:** 2 つのティア: 月額$100 (Pro の 5 倍利用量) または月額$200 (Pro の 20 倍利用量)[4]。
- **モデルアクセス:** Claude 4 Opus および Sonnet 4 へのアクセス (拡張思考モード含む)[15]。
- **利用制限 (チャットインターフェース):** Max 20x ($200/月) の場合: メッセージ長/コンテキストに応じて、5 時間 (1「セッション」) あたり少なくとも 900 メッセージ。月間 50 セッション (250 時間) を超えると制限の可能性あり[5]。[61] では、大きなコンテキストを使用するとメッセージクォータが急速に消費されることが指摘されています。
- **API 料金 (Claude 4 Opus):**
  - 入力: $15.00/1M トークン[11]。
  - 出力: $75.00/1M トークン[11]。
  - 拡張思考トークンは出力として課金[10]。バッチ処理とプロンプトキャッシングで割引あり[4]。
- **主な機能:** Claude 4 Opus/Sonnet の最大利用量、ターミナルでの Claude Code アクセス、高度なリサーチ機能、インテグレーション、新機能への早期アクセス、優先アクセス[4]。

これらのプレミアムティアは、単に最上位モデルへのアクセスを提供するだけでなく、より高

い利用上限、優先アクセス、そして多くの場合、補完的なサービスやツールをバンドルすることで価値を高めています（例：Gemini Ultra の 30 TB ストレージと YouTube Premium 3、ChatGPT Pro の DALL-E 3 と Deep Research 1、Claude Max の Claude Code 4）。
初期の AI サブスクリプションティアは、無料版よりも高性能なモデルへのアクセス提供に重点を置いていました。しかし競争が激化し、ユーザーのニーズが多様化するにつれて、単純なモデルアクセスだけではプレミアム価格を正当化するには不十分になっています。企業は、パワーユーザーやビジネスにとって不可欠な、より高いレート制限（または公正利用の注意書き付きの「ほぼ無制限」）を含む追加機能を提供することで、価値提案を強化しています 5。また、Google の Deep Research が Gemini のコンテキストを活用するように 3、OpenAI の DALL-E 1、Anthropic の Claude Code 4 のような専門ツールも、しばしばフラッグシップモデルの強みを活かしてプレミアムパッケージの一部となっています。さらに、Gemini Ultra の 30 TB ストレージのような非 AI 関連の特典もバンドルされることがあります 3。これらのトップティアへの加入決定には、パッケージ全体の評価が伴います。一部のユーザーにとってはコアモデルへのアクセスと利用上限の高さが鍵となりますが、他のユーザーにとってはバンドルされたサービスが決定要因となることもあります。「価値」は個人や組織のワークフローとニーズに大きく依存します。「ほぼ無制限」という性質は、多くの場合、乱用を防ぐための公正利用ポリシーを伴います 82。

表 5：プレミアムプラン比較

特徴	ChatGPT Pro	Gemini Ultra	Claude Max (20x)
月額料金 (USD)	$200 [1]	$249.99 [3]	$200 [5]
主要アクセスモデル	o3, GPT-4o, GPT-4.5 等 [80]	Gemini 2.5 Pro + Deep Think, Veo 3 [3]	Claude 4 Opus, Sonnet 4 [24]
チャットインターフェース主要機能	高度なツール (DALL-E 3, Deep Research) [1]	Deep Think, Veo 3, Flow, NotebookLM 強化 [3]	拡張思考, Claude Code [4]
主要モデル API アクセス	o3, GPT-4o	Gemini 2.5 Pro	Claude 4 Opus
利用制限 (チャット - 主要モデル)	「ほぼ無制限」(公正利用規約あり) [82]	「最高アクセス」(Pro より高い) [65]	900 メッセージ/5 時間 (目安), 月 50 セッ

			ション上限 [5]
主要モデル API 料金 (入力/出力 1M トークンあたり)	o3: $10/$40 [17] GPT-4o: $2.50/$10 [7]	$1.25-$2.50 / $10-$15 [21]	$15/$75 [11]
特筆すべき追加機能/バンドル	Operator プレビュー [2]	30 TB ストレージ, YouTube Premium, Project Mariner [3]	高度なリサーチ, 統合機能, 早期アクセス [4]

6. 総合比較表

以下に、本レポートで分析した主要な性能指標と特徴をまとめた総合比較表を示します。

項目	**ChatGPT Pro (o3 / GPT-4o)**	**Gemini Ultra (Gemini 2.5 Pro + Deep Think)**	**Claude Max (Claude 4 Opus)**
主要モデル	o3, GPT-4o	Gemini 2.5 Pro with Deep Think	Claude 4 Opus
特筆すべきアーキテクチャ特徴	o3: 推論特化, GPT-4o: ネイティブマルチモーダル	Deep Think, 超巨大コンテキスト	ハイブリッド/拡張推論
知識カットオフ	GPT-4o: 2023 年 10 月 [16], o3: Web 検索可 [17]	2025 年 1 月 [9]	2025 年 3 月 [12]
最大 API コンテキストウィンドウ (入力)	o3: 200K [6], GPT-4o: 1M (限定)/128K[7]	1M (2M 予定) [9]	200K [11]
最大 API 出力トーク	o3: 100K [6], GPT-4o:	65,535 [9]	32K [12]

ン	16K[8]		
主要入力モダリティ (API/Chat)	テキスト, 画像, (o3) 音声 (GPT-4o) [7]	テキスト, コード, 画像, 音声, ビデオ [9]	テキスト, 画像 [11]
主要出力モダリティ (API/Chat)	テキスト, 音声 (GPT-4o), 画像 (GPT-4o) [7]	テキスト, (音声予定) [9]	テキスト [15]
ベンチマークハイライト			
- MMLU (%)	GPT-4o: 88.7 [29]	88.6 (Global MMLU Lite, May '25) [21]	88.8 [37]
- GPQA Diamond (%)	o3: 83.3 [27], GPT-4o: 53.6 [29]	84.0 (Mar '25, p@1) [18]	83.3 (拡張思考) [37]
- SWE-bench (%)	o3: 69.1 [17]	63.2 (May '25) [21]	72.5 (基本) / 79.4 (高計算量) [24]
- MMMU (%)	o3: 82.9 [17]	84.0 (Deep Think) [21]	76.5 (validation) [37]
- AIME 2025 (%)	o3: 88.9 (ツール使用時 98.4%) [17]	83.0 (May '25 single) [21]	75.5 (高計算量時 90.0%) [37]
- NIAH/MRCR (%)	GPT-4o: MMNeedle で強力だが課題あり [56]	MRCR 93.0% (128k avg) [21]	NIAH で強力 [57] (旧 Opus 3: >99% @200K [58])
API パフォーマンス (目安)			
- TTFT (秒)	o3: 13-21 [74], GPT-4o:	27-42 [35]	2.2-3.0 [66]

	0.5-0.6 [76]		
- TPS (出力)	o3: 130-210 [74], GPT-4o: 100-110 [76]	140-210 [35]	39-57 [48]
サブスクリプションプラン名	ChatGPT Pro	Google AI Ultra	Claude Max (20x)
月額料金 (USD)	$200 [1]	$249.99 [3]	$200 [5]
プラン主要機能/利用制限 (Chat)	o3, GPT-4o 等への「ほぼ無制限」アクセス, 高度ツール [1]	Gemini 2.5 Pro + Deep Think への「最高アクセス」, Veo 3 等 [3]	Claude 4 Opus/Sonnet への高利用量アクセス (900 msg/5h 目安) [5]
主要モデル API 料金 (入力/出力 1M トークンあたり)	o3: $10/$40 [17], GPT-4o: $2.50/$10 [7]	$1.25-$2.50 / $10-$15 [21]	$15/$75 [11]

7. 定性的パフォーマンス分析：強みと弱み

7.1. ChatGPT Pro (o3 および GPT-4o)

- 強み:
  - **o3:** 数学、科学、コーディングにおける卓越した推論能力 (AIME, GPQA, Codeforces, SWE-bench で高スコア [17])。「画像で思考する」機能による独自の視覚分析 [5]。o1 と比較してエラー率が低い [17]。新規仮説の生成・評価能力が高い [17]。
  - **GPT-4o:** 全方位的に強力なパフォーマンス (MMLU, HumanEval で高スコア [29])。優れたネイティブマルチモーダル能力 (テキスト、画像、音声の入出力 [7])。高速な API 応答 (低い TTFT[76])。多言語タスクに強い [70]。画像内のテキストレンダリングが向上 [13]。
  - **プラン:** Pro ユーザー向けの「ほぼ無制限」アクセスは大きな魅力 [81]。多様な強力モデル群へのアクセス。
- 弱み・考慮事項:
  - **o3:** API の TTFT が比較的長い [74]。SimpleQA でのハルシネーション率が懸念材料 [46]。ARC-AGI-2 のスコアが低い [43]。o3 API のコストが比較的高め。

- **GPT-4o:** 一部の超高度な推論ベンチマークでは、o3 や競合の専門モードに劣る場合がある (例: GPQA[21])。MMNeedle では陰性サンプルでのハルシネーション問題が指摘されている[56]。
- **プラン:** 「ほぼ無制限」は公正利用の対象であり、一部の Plus ユーザーは依然として制限に達することがある[61]。

7.2. Gemini Ultra (Gemini 2.5 Pro with Deep Think)

- **強み:**
  - 最大のコンテキストウィンドウ (1M、2M 予定[9]) と、優れた長文コンテキスト想起能力 (MRCR/NIAH で 90-99%超[10])。
  - 「Deep Think」モードによる強化された複雑な推論能力。非常に難しい数学 (USAMO[21]) やコーディング (LiveCodeBench[21]) で優れる。
  - ビデオ入力や計画中の音声出力を含む強力なネイティブマルチモーダル性[9]。Video-MME[55] や MMMU (Deep Think 使用時[21]) でリード。
  - 一般的な推論 (GPQA[18]) やコーディング (WebDev Arena リーダー[21]) でも競争力あり。
  - Ultra プランには大きな非 AI 価値 (30TB ストレージ、YouTube Premium[3]) が含まれる。
- **弱み・考慮事項:**
  - API アクセスの TTFT が 3 社の中で最も長く[35]、特定の最適化なしでは高度にインタラクティブなリアルタイムアプリケーションには不向きな場合がある。
  - 「Deep Think」は実験的であり、当初は信頼できるテスターにのみ提供[21]。
  - 一部のユーザーレビューでは、Pro/Deep Think を使用しない場合、特定のタスクで Gemini が最新の競合他社と比較してコンテキスト認識が低い、または品質が「低下」する可能性があると示唆[84]。
  - サブスクリプション費用が最も高い[3]。

7.3. Claude Max (Claude 4 Opus)

- **強み:**
  - 市場をリードするコーディング性能、特に SWE-bench および Terminal-bench で優れる[23]。「コードのセンス」や複雑な複数ファイル変更の処理能力が高く評価されている[57]。
  - 「拡張思考」モードによる深い推論と長期間タスクの実行能力 (コーディングで 7 時間以上自律的に作業可能[11])。
  - 200K ウィンドウ内での優れた記憶力とコンテキスト管理、特にローカルファイルアクセス時[24]。強力な NIAH 性能[57]。

    - MMLU および GPQA (拡張思考あり) で良好な性能[37]。
    - 強力な指示追従性と「近道」行動の低減[11]。
    - クリエイティブライティングやニュアンスのあるコンテンツ生成に優れる[22]。
- **弱み・考慮事項:**
    - API の出力速度 (TPS) が o3 や Gemini 2.5 Pro より低い[48]。
    - コンテキストウィンドウ (200K) は大きいが、Gemini 2.5 Pro の 1M よりは小さい[12]。
    - API のトークンあたりコストが最も高く、特に出力コストが高い[11]。
    - MMMU スコアは競争力があるものの、o3 や Gemini 2.5 Pro にわずかに劣る[37]。
    - Max プランのメッセージ制限は、大きなコンテキストを使用するとすぐに達する可能性がある[61]。

各モデルは広範な能力を目指しつつも、明確な専門分野が形成されつつあります。Claude 4 Opus は高度なコーディングと長時間の自律的エージェントタスク[24]、OpenAI の o3 は集中的な数学的・科学的推論[27]、Gemini 2.5 Pro は巨大なコンテキスト処理と広範なマルチモーダル性[9]、そして GPT-4o は非常に汎用性が高く高速なマルチモーダルジェネラリスト[7]としての特性を強めています。AGI（汎用人工知能）の追求は、広範な認知タスクでの卓越を伴いますが、アーキテクチャの選択、訓練データの構成、ファインチューニングの優先順位が性能差を生んでいます。例えば、Anthropic の「Constitutional AI」と安全性への焦点は、Claude の構造化された出力と指示追従性に影響を与えている可能性があります[19]。OpenAI の o シリーズにおける強化学習を多用したアプローチは推論能力を高め[17]、Google の広大なデータリソースとインフラは巨大なコンテキストとマルチモーダル性をサポートしています。結果として、モデルは全般的に高性能でありながら、特定の分野で「突出した」能力を持つようになります。ユーザーや開発者はこれらの強みを認識しており（例：「Claude 4 はコーディングに優れる」[51]、「Gemini はコーディング、倫理的議論、翻訳で支配的」[90]、「GPT-4o は推論とマルチモーダルサポートに優れる」[51]）、この専門化はベンチマーク解釈をよりニュアンスのあるものにしています。あるベンチマーク（例：MMLU）での高スコアが、別のベンチマーク（例：SWE-bench）でのトップ性能を保証するものではありません。したがって、ユーザーは主要なユースケースに合わせてモデルを慎重に選択する必要があります。「万能」なトップモデルというよりは、タスク依存で「最適な」モデルが存在する状況です。これはまた、包括的で多様なベンチマークスイートの必要性を高めています。

8. 結論と専門家による推奨

2025 年 5 月現在、ChatGPT Pro (o3 および GPT-4o 搭載)、Gemini Ultra (Gemini 2.5 Pro + Deep Think 搭載)、および Claude Max (Claude 4 Opus 搭載) は、それぞれ独自の特徴と強みを持つ、非常に強力な AI 機能を提供しています。

全体評価:

- **OpenAI** は、GPT-4o の迅速なマルチモーダル性能と o3 の深い推論能力を組み合わせることで、多様なニーズに対応できる Pro プランを提供しており、広範な能力と深さを求めるユーザーにとって魅力的です。
- **Google の Gemini Ultra** は、比類のないコンテキストウィンドウとビデオを含む包括的なマルチモーダルサポートを特徴とし、膨大な量の多様なデータや Deep Think を介した複雑な仮説駆動型の研究を必要とするタスクに最適です。
- **Anthropic の Claude Max** は、Claude 4 Opus の優れたコーディング能力、長時間のタスク耐久性、堅牢な指示追従性により、開発およびエージェントワークフローにおいて強力な地位を確立しています。

ユースケース別推奨:

- **開発者向け (コーディング & ソフトウェアエンジニアリング):**
  - **最有力候補:** Claude Max (Claude 4 Opus)。SWE-bench/Terminal-bench でのトップスコアと長時間のコーディング能力が理由です [24]。
  - **強力な代替:** ChatGPT Pro (o3)。高い SWE-bench および Codeforces スコアを誇ります [17]。Gemini Ultra (Gemini 2.5 Pro with Deep Think) は、LiveCodeBench での性能と 1M コンテキストによる大規模コードベース分析に適しています [10]。
- **研究者向け (複雑な推論, 数学, 科学):**
  - **最有力候補 (数学/科学推論):** ChatGPT Pro (o3)。AIME, GPQA Diamond, FrontierMath での実績が理由です [27]。Gemini Ultra (Gemini 2.5 Pro with Deep Think) も USAMO や GPQA Diamond で強力です [21]。
  - **最有力候補 (長文研究統合):** Gemini Ultra。1M トークンのコンテキストと Deep Research ツールが理由です [3]。Claude Max (Claude 4 Opus) も 200K コンテキストと強力な要約・分析能力で適しています [57]。
- **クリエイティブプロフェッショナル向け (執筆, コンテンツ生成):**
  - **最有力候補 (汎用クリエイティブ & マルチモーダル):** ChatGPT Pro (GPT-4o)。ネイティブな画像生成、音声機能、強力な一般テキスト生成能力が理由です [13]。
  - **最有力候補 (ニュアンスのある/構造化された執筆):** Claude Max (Claude 4 Opus/Sonnet 4)。「人間らしい散文」、「視覚的センス」、ブランドボイスへ

の準拠能力が理由です[22]。

- **マルチモーダルアプリケーション向け (画像, 音声, ビデオ分析):**
  - **最有力候補:** Gemini Ultra (Gemini 2.5 Pro)。ビデオ入力を含む最も広範なモダリティサポートと、MMMU/Video-MME での高スコアが理由です[9]。
  - **強力な代替:** ChatGPT Pro (GPT-4o)。堅牢な画像・音声処理、ネイティブ画像生成能力が理由です[7]。o3 の「画像で思考する」機能も特筆されます[17]。
- **エンタープライズ & 長文ドキュメント処理向け:**
  - **最有力候補:** Gemini Ultra (Gemini 2.5 Pro)。1M トークンのコンテキストと高い NIAH/MRCR スコアが理由です[10]。
  - **強力な代替:** Claude Max (Claude 4 Opus)。200K コンテキストと実証済みの NIAH 性能が理由です[57]。

プレミアムティアの価値提案:

- **ChatGPT Pro:** OpenAI の多様な最先端モデル (速度/マルチモーダルの GPT-4o、深い推論の o3) への高い利用制限付きアクセスを提供し、汎用性を必要とするユーザーにアピールします。
- **Gemini Ultra:** より高価ですが、最大のコンテキストウィンドウ、広範なマルチモーダルサポート (Veo 3 のようなビデオツール含む)、そして大量のクラウドストレージ/サービスをバンドルすることで正当化され、Google エコシステムに深く関与している、または大規模なデータ処理を必要とするプロフェッショナルやクリエイターをターゲットとしています。
- **Claude Max:** おそらく最高のコーディングおよび長時間タスクエージェントモデル (Opus 4) への大量アクセスを提供し、AI 駆動のソフトウェア開発や複雑な自動化ワークフローに注力する開発者や企業にとって理想的です。

最終的な考察:
選択は、主要なユースケース、予算、および API 速度や特定のモダリティサポートのような必要な技術的機能に大きく依存します。これら 3 つのサービスはすべて 2025 年初頭から中盤にかけての AI の頂点を代表するものですが、その強みはますます専門化しています。新しいモデルやアップデートが頻繁に行われるため、継続的な評価が推奨されます。
「Pro」、「Ultra」、「Max」といったプレミアムティアは、単なるモデルの性能だけでなく、より高い利用クォータ[5]、専門ツールへのアクセス（Deep Research, Veo 3, Claude Code[3]）、優先アクセス、さらにはクラウドストレージのような非 AI 関連の特典[3]を含むサービスのバンドルとしての価値がますます高まっています。主要モデルの API コストは、これらのプラットフォーム上で開発を行う開発者にとって、依然として別途考慮すべき重要な要素です。初期の AI サブスクリプションティアは、無料版よりも高性能なモデルへのアクセス提供に焦点を当てていました。競争が激化し、ユーザー

のニーズが多様化するにつれて、単純なモデルアクセスだけではプレミアム価格を正当化するには不十分になっています。企業は、パワーユーザーやビジネスにとって不可欠な、より高いレート制限を含む追加機能を提供することで、価値提案を強化しています。Google の Deep Research が Gemini のコンテキストを活用するように[3]、OpenAI の DALL-E[1]、Anthropic の Claude Code[4] のような専門ツールも、しばしばフラッグシップモデルの強みを活かしてプレミアムパッケージの一部となっています。エコシステムへの囲い込みは、これらのバンドルサービスによって促進される可能性があります。

引用文献

1. ChatGPT Pricing Plans: Which Plan is Right for You?, 6月 1, 2025 にアクセス、 https://www.beingguru.com/chatgpt-pricing-plans-which-plan-is-right-for-you/
2. Is ChatGPT Plus worth your $20? Here's how it compares to Free and Pro plans | ZDNET, 6月 1, 2025 にアクセス、 https://www.zdnet.com/article/is-chatgpt-plus-worth-your-20-heres-how-it-compares-to-free-and-pro-plans/
3. Google AI Ultra とは？月額 3 万 6000 円の AI サブスクプランについて ..., 6 月 1, 2025 にアクセス、 https://romptn.com/article/58965
4. Pricing - Anthropic, 6 月 1, 2025 にアクセス、 https://www.anthropic.com/pricing
5. About Claude's Max Plan Usage | Anthropic Help Center, 6 月 1, 2025 にアクセス、 https://support.anthropic.com/en/articles/11014257-about-claude-s-max-plan-usage
6. OpenAI o3 and o4-mini models– FAQ [ChatGPT Enterprise & Edu], 6月 1, 2025 にアクセス、 https://help.openai.com/en/articles/9855712-openai-o3-and-o4-mini-models-faq-chatgpt-enterprise-edu
7. OpenAI GPT-4o API - Replicate, 6 月 1, 2025 にアクセス、 https://replicate.com/openai/gpt-4o
8. Model - OpenAI API, 6 月 1, 2025 にアクセス、 https://platform.openai.com/docs/models/gpt-4o
9. Gemini 2.5 Pro | Generative AI on Vertex AI- Google Cloud, 6 月 1, 2025 にアクセス、 https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro
10. Gemini 2.5 Pro: A Comparative Analysis Against Its AI Rivals (2025 Landscape) Dirox, 6 月 1, 2025 にアクセス、 https://dirox.com/post/gemini-2-5-pro-a-comparative-analysis-against-its-ai-rivals-2025-landscape
11. 「Claude 4」が登場！驚きの性能や料金、使い方を徹底解説 ..., 6 月 1, 2025 にアクセス、 https://romptn.com/article/59170
12. The Complete Guide to Claude Opus 4 and Claude Sonnet 4 - PromptHub, 6 月 1, 2025 にアクセス、 https://www.prompthub.us/blog/the-complete-guide-to-claude-opus-4-and-claude-sonnet-4

13. Introducing 4o Image Generation - OpenAI, 6 月 1, 2025 にアクセス、 https://openai.com/index/introducing-4o-image-generation/
14. Gemini 2.5 on Vertex AI: Pro, Flash & Model Optimizer Live | Google Cloud Blog, 6 月 1, 2025 にアクセス、 https://cloud.google.com/blog/products/ai-machine-learning/gemini-2-5-pro-flash-on-vertex-ai
15. Models overview - Anthropic API, 6 月 1, 2025 にアクセス、 https://docs.anthropic.com/en/docs/about-claude/models/overview
16. GPT-4o - Wikipedia, 6 月 1, 2025 にアクセス、 https://en.wikipedia.org/wiki/GPT-4o
17. OpenAI の新モデル「o3」と「o4-mini」が登場！性能・使い方 ..., 6 月 1, 2025 にアクセス、 https://romptn.com/article/56853
18. Gemini 2.5 Pro: Features, Tests, Access, Benchmarks & More | DataCamp, 6 月 1, 2025 にアクセス、 https://www.datacamp.com/blog/gemini-2-5-pro
19. www-cdn.anthropic.com, 6 月 1, 2025 にアクセス、 https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf
20. Introducing OpenAI o3 and o4-mini, 6 月 1, 2025 にアクセス、 https://openai.com/index/introducing-o3-and-o4-mini/
21. Gemini 2.5: Our most intelligent models are getting even better, 6 月 1, 2025 にアクセス、 https://blog.google/technology/google-deepmind/google-gemini-updates-io-2025/
22. Claude Opus 4 - Anthropic, 6 月 1, 2025 にアクセス、 https://www.anthropic.com/claude/opus
23. Everything to know about Claude 4 (Sonnet 4, Opus 4), from the Good, to the Bad, and the REAL weird... | The Neuron, 6 月 1, 2025 にアクセス、 https://www.theneuron.ai/explainer-articles/everything-to-know-about-claude-4-sonnet-4-opus-4-from-the-good-to-the-bad-and-the-mid
24. Introducing Claude 4 \ Anthropic, 6 月 1, 2025 にアクセス、 https://www.anthropic.com/news/claude-4
25. Introducing Claude 4 - Anthropic, 6 月 1, 2025 にアクセス、 https://www.anthropic.com/news/claude-4?utm_medium=direct&eloqua_id=4999
26. A Peek Behind the Claude Curtain - BankInfoSecurity, 6 月 1, 2025 にアクセス、 https://www.bankinfosecurity.com/peek-behind-claude-curtain-a-28530
27. OpenAI's O3: Features, O1 Comparison, Benchmarks & More | DataCamp, 6 月 1, 2025 にアクセス、 https://www.datacamp.com/blog/o3-openai
28. OpenAI O3 Review: Reasoning, Visuals & Tool Use Analysis - Monica, 6 月 1, 2025 にアクセス、 https://monica.im/blog/openai-o3-3/
29. Education level interpretation of Gpt-4o's benchmarks - OpenAI Developer Community, 6 月 1, 2025 にアクセス、 https://community.openai.com/t/education-level-interpretation-of-gpt-4os-benchmarks/763947
30. LLM Benchmarks in 2024: Overview, Limits and Model Comparison, 6 月 1, 2025

にアクセス、 https://www.vellum.ai/blog/llm-benchmarks-overview-limits-and-model-comparison
31. Claude Sonnet 3.5 VS ChatGPT 4o - AI/ML API, 6 月 1, 2025 にアクセス、 https://aimlapi.com/comparisons/claude-sonnet-3-5-vs-chatgpt-4o
32. Claude 3 Opus vs GPT-4: Which AI Model is Best? (2024) - PromptLayer, 6 月 1, 2025 にアクセス、 https://blog.promptlayer.com/comparing-frontier-models-claude-3-opus-vs-gpt-4/
33. arxiv.org, 6 月 1, 2025 にアクセス、 https://arxiv.org/abs/2303.08774
34. Gemini - Google DeepMind, 6 月 1, 2025 にアクセス、 https://deepmind.google/technologies/gemini/
35. Gemini 2.5 Pro - Intelligence, Performance & Price Analysis ..., 6 月 1, 2025 にアクセス、 https://artificialanalysis.ai/models/gemini-2-5-pro
36. NEW Google Gemini 2.5 Pro Unveiled: Smarter Tools for 2025 - Nitro Media Group, 6 月 1, 2025 にアクセス、 https://www.nitromediagroup.com/google-gemini-ai-advancements-2025-tools/
37. Claude 4: Tests, Features, Access, Benchmarks & More - DataCamp, 6 月 1, 2025 にアクセス、 https://www.datacamp.com/blog/claude-4
38. Claude 4 Review: Anthropic's New AI Models Shine in Coding and Beyond | MindPal Blog, 6 月 1, 2025 にアクセス、 https://mindpal.space/blog/claude-4-anthropic-ai-model-review-9k3h2g
39. GPT-4 Technical Report - Papers With Code, 6 月 1, 2025 にアクセス、 https://paperswithcode.com/paper/gpt-4-technical-report-1
40. Top LLM Benchmarks Explained: MMLU, HellaSwag, BBH, and Beyond - Confident AI, 6 月 1, 2025 にアクセス、 https://www.confident-ai.com/blog/llm-benchmarks-mmlu-hellaswag-and-beyond
41. HellaSwag-Pro: A Large-Scale Bilingual Benchmark for Evaluating the Robustness of LLMs in Commonsense Reasoning - arXiv, 6 月 1, 2025 にアクセス、 https://arxiv.org/html/2502.11393v1
42. Leaderboard - ARC Prize, 6 月 1, 2025 にアクセス、 https://arcprize.org/leaderboard
43. OpenAI's o3 model scores 3% on the ARC-AGI-2 benchmark, compared to 60% for the average human - Effective Altruism Forum, 6 月 1, 2025 にアクセス、 https://forum.effectivealtruism.org/posts/CoPNbwNqDai6orZhv/openai-s-o3-model-scores-3-on-the-arc-agi-2-benchmark
44. OpenAI's o3 is less AGI than originally measured - The Decoder, 6 月 1, 2025 にアクセス、 https://the-decoder.com/openais-o3-is-less-agi-than-originally-measured/
45. GPT 4.5 Released: Here Are the Benchmarks - Helicone, 6 月 1, 2025 にアクセス、 https://www.helicone.ai/blog/gpt-4.5-benchmarks
46. OpenAI o3 and o4-mini System Card, 6 月 1, 2025 にアクセス、 https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

47. Calibration of OpenAI o3 and o4-mini on Humanity's Last Exam - Scale AI, 6 月 1, 2025 にアクセス、 https://scale.com/blog/o3-o4-mini-calibration
48. Claude Sonnet 4 vs Claude Opus 4: A comprehensive comparison, 6 月 1, 2025 にアクセス、 https://www.keywordsai.co/blog/claude-sonnet-4-vs-claude-opus-4-a-comprehensive-comparison
49. Anthropic's Claude 4 is OUT and Its Amazing! - Analytics Vidhya, 6 月 1, 2025 にアクセス、 https://www.analyticsvidhya.com/blog/2025/05/anthropics-claude-4-is-out-and-its-amazing/
50. Introducing Claude 4 : r/ClaudeAI - Reddit, 6 月 1, 2025 にアクセス、 https://www.reddit.com/r/ClaudeAI/comments/1ksvebb/introducing_claude_4/
51. Claude 4 vs GPT-4o vs Gemini 2.5 Pro: Which AI Codes Best in 2025? - Analytics Vidhya, 6 月 1, 2025 にアクセス、 https://www.analyticsvidhya.com/blog/2025/05/best-ai-for-coding/
52. This week's launches: o3, o4-mini, GPT-4.1, and Codex CLI - OpenAI Developer Community, 6 月 1, 2025 にアクセス、 https://community.openai.com/t/this-weeks-launches-o3-o4-mini-gpt-4-1-and-codex-cli/1230312
53. Claude Opus 4 vs. Gemini 2.5 Pro vs. OpenAI o3 Coding Comparison - DEV Community, 6 月 1, 2025 にアクセス、 https://dev.to/composiodev/claude-opus-4-vs-gemini-25-pro-vs-openai-o3-coding-comparison-3jnp
54. OpenAI vs. DeepSeek: Which AI Understands Kotlin Better? - The JetBrains Blog, 6 月 1, 2025 にアクセス、 https://blog.jetbrains.com/kotlin/2025/02/openai-vs-deepseek-which-ai-understands-kotlin-better/
55. Gemini 2.5 Pro Preview: even better coding performance - Google Developers Blog, 6 月 1, 2025 にアクセス、 https://developers.googleblog.com/en/gemini-2-5-pro-io-improved-coding-performance/
56. Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models - ACL Anthology, 6 月 1, 2025 にアクセス、 https://aclanthology.org/2025.naacl-long.166.pdf
57. Claude Opus 4 and Sonnet 4 Technical Review: The Best Coding Models for Developers?, 6 月 1, 2025 にアクセス、 https://www.helicone.ai/blog/claude-opus-and-sonnet-4-full-developer-guide
58. Introducing the next generation of Claude - Anthropic, 6 月 1, 2025 にアクセス、 https://www.anthropic.com/news/claude-3-family
59. Google AI Pro &Ultra — get access to Gemini 2.5 Pro &more, 6 月 1, 2025 にアクセス、 https://gemini.google/subscriptions/
60. Claude 4: How to Access and Use It - Chatbase, 6 月 1, 2025 にアクセス、 https://www.chatbase.co/blog/claude-4
61. Why you are constantly hitting message limits with Pro plan, and why you don't get to have this problem with ChatGPT : r/ClaudeAI - Reddit, 6 月 1, 2025 にアクセス、 https://www.reddit.com/r/ClaudeAI/comments/1j08v73/why_you_are_constantly_hitting_message_limits/

62. Open AI O3 vs GPT-4: Top Differences That You Should Know in 2025 | YourGPT, 6 月 1, 2025 にアクセス、 https://yourgpt.ai/blog/updates/open-ai-o3-vs-gpt-4-top-differences-that-you-should-know-in-2025
63. Introducing GPT-4.1 in the API - OpenAI, 6 月 1, 2025 にアクセス、 https://openai.com/index/gpt-4-1/
64. ChatGPT Subscription Tiers: Comparative Analysis of Usage Limits - Wizbrand, 6 月 1, 2025 にアクセス、 https://www.wizbrand.com/tutorials/chatgpt-subscription-tiers-comparative-analysis-of-usage-limits/
65. Gemini Apps limits &upgrades for Google AI subscribers, 6 月 1, 2025 にアクセス、 https://support.google.com/gemini/answer/16275805?hl=en
66. Claude 4 Opus - Intelligence, Performance &Price Analysis, 6 月 1, 2025 にアクセス、 https://artificialanalysis.ai/models/claude-4-opus
67. Claude 4 benchmarks show improvements, but context is still 200K - Bleeping Computer, 6 月 1, 2025 にアクセス、 https://www.bleepingcomputer.com/news/artificial-intelligence/claude-4-benchmarks-show-improvements-but-context-is-still-200k/
68. Introducing next-generation audio models in the API - OpenAI, 6 月 1, 2025 にアクセス、 https://openai.com/index/introducing-our-next-generation-audio-models/
69. Audio and speech - OpenAI API, 6 月 1, 2025 にアクセス、 https://platform.openai.com/docs/guides/audio
70. Choosing the right AI model for your task - GitHub Docs, 6 月 1, 2025 にアクセス、 https://docs.github.com/en/copilot/using-github-copilot/ai-models/choosing-the-right-ai-model-for-your-task
71. Top LLMs in 2025: Comparing Claude, Gemini, and GPT-4 LLaMA - FastBots.ai, 6 月 1, 2025 にアクセス、 https://fastbots.ai/blog/top-llms-in-2025-comparing-claude-gemini-and-gpt-4-llama
72. Anthropic's Claude in Amazon Bedrock - AWS, 6 月 1, 2025 にアクセス、 https://aws.amazon.com/bedrock/anthropic/
73. Building with Claude - Anthropic API, 6 月 1, 2025 にアクセス、 https://docs.anthropic.com/en/docs/overview
74. o3 - Intelligence, Performance &Price Analysis, 6 月 1, 2025 にアクセス、 https://artificialanalysis.ai/models/o3
75. OpenAI o3-mini API - Replicate, 6 月 1, 2025 にアクセス、 https://replicate.com/openai/o3-mini
76. GPT-4o (Nov '24) - Intelligence, Performance &Price Analysis ..., 6 月 1, 2025 にアクセス、 https://artificialanalysis.ai/models/gpt-4o
77. Google's Gemini 2.5 Pro Experimental Outperforms Top AI Models - DeepLearning.AI, 6 月 1, 2025 にアクセス、 https://www.deeplearning.ai/the-batch/googles-gemini-2-5-pro-experimental-outperforms-top-ai-models/
78. Gemini 2.5 Pro takes huge lead in new MathArena USAMO benchmark - Reddit, 6 月 1, 2025 にアクセス、

https://www.reddit.com/r/singularity/comments/1jpqjez/gemini_25_pro_takes_huge_lead_in_new_matharena/
79. Claude 4 Opus Thinking: API Provider Performance Benchmarking & Price Analysis, 6 月 1, 2025 にアクセス、 https://artificialanalysis.ai/models/claude-4-opus-thinking/providers
80. Limitations on the OpenAI o-series reasoning models on ChatGPT, 6 月 1, 2025 にアクセス、 https://community.openai.com/t/limitations-on-the-openai-o-series-reasoning-models-on-chatgpt/1230183
81. ChatGPT Plus o3, o4-mini, o4-mini-high Usage Limits 2025: Complete Guide - 拼账号, 6 月 1, 2025 にアクセス、 https://pinzhanghao.com/openai/chatgpt-plus-o3-limits-2025-complete-guide/
82. OpenAI o3 and o4-mini Usage Limits on ChatGPT and the API, 6 月 1, 2025 にアクセス、 https://help.openai.com/en/articles/9824962-openai-o3-and-o4-mini-usage-limits-on-chatgpt-and-the-api
83. Users who are on the pro subscription and feel they are getting their money's worth out of the $200/mo - what do you use ChatGPT for? : r/OpenAI - Reddit, 6 月 1, 2025 にアクセス、 https://www.reddit.com/r/OpenAI/comments/1jwug4q/users_who_are_on_the_pro_subscription_and_feel/
84. What's the usage limit on Gemini 2.5 Pro for Pro subscribers? : r/Bard - Reddit, 6 月 1, 2025 にアクセス、 https://www.reddit.com/r/Bard/comments/1kujluq/whats_the_usage_limit_on_gemini_25_pro_for_pro/
85. Gemini Developer API Pricing | Gemini API | Google AI for Developers, 6 月 1, 2025 にアクセス、 https://ai.google.dev/gemini-api/docs/pricing
86. BREAKING : Anthropic introduces Claude MAX : r/ClaudeAI - Reddit, 6 月 1, 2025 にアクセス、 https://www.reddit.com/r/ClaudeAI/comments/1jvbpek/breaking_anthropic_introduces_claude_max/
87. Gemini Deep Research — your personal research assistant - Google Gemini, 6 月 1, 2025 にアクセス、 https://gemini.google/overview/deep-research/
88. Claude 4 Opus vs. Gemini 2.5 pro vs. OpenAI o3: Coding comparison - Composio, 6 月 1, 2025 にアクセス、 https://composio.dev/blog/claude-4-opus-vs-gemini-2-5-pro-vs-openai-o3/
89. Write beautiful code, ship powerful products | Claude by ... - Anthropic, 6 月 1, 2025 にアクセス、 https://www.anthropic.com/solutions/coding
90. I tested Claude vs ChatGPT vs Gemini with 10 prompts — Here's what won, 6 月 1, 2025 にアクセス、 https://techpoint.africa/guide/claude-vs-chatgpt-vs-gemini/
91. I tested Gemini 2.5 Pro vs Claude 4 Sonnet with the same 7 prompts — here's who came out on top | Tom's Guide, 6 月 1, 2025 にアクセス、 https://www.tomsguide.com/ai/i-tested-gemini-2-5-pro-vs-claude-4-sonnet-

with-the-same-7-prompts-heres-who-came-out-on-top