GPT-5.1 Pro と Gemini 3 Pro ：フロンティア AI モデルにおける能力、経済性、および運用特性の包括的比較分析レポート

Gemini

1. エグゼクティブサマリー

2025 年 11 月、人工知能（AI）のエコシステムは、OpenAI による「GPT-5.1 Pro」および「Codex-Max」と、Google DeepMind による「Gemini 3 Pro」の同時期リリースによって、かつてない技術的転換点を迎えました[1]。これら 2 つのモデルは、単なる性能向上版ではなく、それぞれが異なる設計思想に基づいた「知能のあり方」を提示しており、AI の適用領域を従来のチャットボットや単純な自動化から、自律的なエージェントワークフロー、未踏の科学的発見、そして大規模なシステムエンジニアリングへと拡張しています。

本レポートは、これら 2 つの最先端モデルについて、技術的アーキテクチャ、ベンチマーク性能（FrontierMath 、CritPt、SWE-Bench 等）、料金体系と総所有コスト（TCO）、開発者エコシステム、そして産業別のユースケース適合性を、15,000 語に及ぶ詳細な分析を通じて明らかにすることを目的としています。

分析の核心的結論として、両モデルは「汎用性」という共通の目標を持ちながらも、その実現手段において明確な戦略的分岐を見せています。**GPT-5.1 Pro** は、「適応型推論（Adaptive Reasoning）」と「コンパクション（Compaction ）」技術を核とし、実務的なコーディング、デバッグ、およびエージェントの安定動作において卓越した信頼性を提供します。これは、既存のエンジニアリングプロセスにシームレスに統合し、予測可能な成果物を生み出す「究極の熟練労働者」としての性格を帯びています[1]。

対照的に、**Gemini 3 Pro** は、ネイティブなマルチモーダル統合と、100 万トークンを超える巨大なコンテキストウィンドウ、そして「Deep Think」モードによる深層推論を武器に、科学的研究、複雑な数学的証明、および大規模な情報の合成において他を圧倒しています。特に、物理学や数学のフロンティア領域における推論能力は、従来モデルの限界を打破するものであり、未知の解を探索する「天才的な研究パートナー」としての地位を確立しています[4]。

以下の章では、これらの結論を裏付けるデータ、定性的評価、および経済的分析を詳細に展開します。

---

2. 戦略的展望とアーキテクチャの進化

2.1 2025 年 11 月のパラダイムシフト

2025 年 11 月のリリースラッシュは、AI 開発における「軍拡競争」が新たなフェーズに入ったことを象徴しています。OpenAI と Google は、互いに数日違いでフラッグシップモデルを投入し、市場の覇権を争っています[1]。この競争は単なるベンチマークスコアの「数字比べ」を超え、AI が人間社会においてどのような役割を果たすべきかという哲学の衝突でもあります。

OpenAI は GPT-5.1 において、ユーザーインターフェースの「暖かさ」や「会話の自然さ」を重視しつつ、バックエンドではエンジニアリングタスクの完遂能力を極限まで高めるアプローチを採りました[2]。一方、Google は Gemini 3 Pro において、物理世界の理解（マルチモーダル）と論理的な深さ（数学・科学）を追求し、AI を物理世界や学術研究と直結させる戦略を明確にしています[3]。

2.2 GPT-5.1 Pro：適応型推論と高密度化の技術論

GPT-5.1 Pro のアーキテクチャにおける最大の革新は、「静的な巨大モデル」から「動的な適応型システム」への移行です。

2.2.1 適応型推論（Adaptive Reasoning ）のメカニズム

GPT-5.1 は、「Instant」と「Thinking」という 2 つの主要な動作モード、あるいはそれらを動的に切り替えるメカニズムを内包しています[3]。従来の LLM（大規模言語モデル）は、入力トークンに対して一定の計算量（FLOPs）を費やして出力を生成していましたが、GPT-5.1 の適

応型推論は、タスクの難易度を即座に評価し、推論にかけるリソースを動的に配分します。

- **Instant モード**: 定型的な挨拶、単純な事実確認、あるいは低レイテンシが求められる対話において、モデルは高速なパスを選択し、ユーザーに対して即座に応答します。これにより、チャットボットとしての体感速度とコスト効率が最適化されます[6]。
- **Thinking モード**: 複雑な論理パズル、多段階の計画が必要なコーディングタスク、あるいは曖昧な前提条件を含む質問に対しては、モデルは内部的な「思考時間」を確保します。このプロセスでは、連鎖的な思考（Chain of Thought）が内部で展開され、自己批判や経路探索が行われた後、最終的な回答が出力されます。このアプローチにより、GPT-5.1 は従来のモデルが苦手としていた「ひっかけ問題」や論理的整合性の維持において高い性能を発揮します[2]。

2.2.2 コンパクション（Compaction ）：無限のコンテキストへの挑戦

GPT-5.1、特に Codex-Max モデルにおいて導入された「コンパクション」技術は、コンテキスト管理の革命と言えます[1]。Gemini が「生のデータを大量にメモリに保持する」ブルートフォースなアプローチを採るのに対し、OpenAI は「情報の意味的圧縮」を選択しました。
コンパクションは、進行中の対話や作業ログ、コードベースの状態をリアルタイムで分析し、不要になった詳細情報を剪定（Pruning）しつつ、重要な決定事項や文脈を要約して保持します。これにより、見かけ上のコンテキストウィンドウの制限を超えて、数百万トークンに及ぶプロジェクトの文脈を維持し続けることが可能になります。これは、人間の短期記憶から長期記憶への移行プロセスを模倣したものであり、長期間にわたるエージェントの運用において、計算コストと精度のバランスを保つための現実解を提供しています[7]。

2.3 Gemini 3 Pro ：ネイティブマルチモーダルとスケーラビリティ

Google の Gemini 3 Pro は、DeepMind が提唱してきた「汎用人工知能（AGI）」への道筋を、スケールと統合によって具現化したものです。

2.3.1 ネイティブマルチモーダルの真価

「ネイティブ」であることの意味は極めて重大です。従来の多くのモデル（GPT-4V 初期版など）は、視覚情報を処理するエンコーダと言語モデルを後付けで接合したものでしたが、

Gemini 3 Pro は最初からテキスト、画像、音声、ビデオ、コードを同一の埋め込み空間で学習しています 3。
これにより、モダリティ間の変換ロスが発生しません。例えば、ビデオ内の「ガラスが割れる音」と「微細な亀裂の映像」を同時に処理し、「何が起きたか」だけでなく「なぜ割れたか（物理的因果関係）」を推論することが可能です。MMMU-Pro や Video-MMMU における圧倒的なスコア（後述）は、このアーキテクチャの正しさを証明しています 1。

2.3.2 100 万トークンの「Deep Think」

Gemini 3 Pro は標準で 100 万トークン、最大でそれ以上のコンテキストウィンドウをサポートしています 8。これに加えて、「Deep Think」機能が強化されており、膨大な情報量の中から微細な関連性を見つけ出す能力が飛躍的に向上しています。
特に注目すべきは、このロングコンテキスト能力が「推論の深さ」と組み合わさった時の相乗効果です。数千ページの学術論文や数万行のコードを読み込ませた際、単なる検索（Retrieval）ではなく、それら全体を前提とした論理的推論を行うことができます。これは、従来の RAG（検索拡張生成）システムが抱えていた「検索漏れによる回答品質の低下」という課題を根本から解決する可能性を秘めています 9。

---

3. ベンチマーク分析：数字が語る知能の特性

本章では、公開されたベンチマーク結果を詳細に分析し、それぞれの数字が実世界のどのような能力を示唆しているのかを解剖します。単なるスコアの列挙ではなく、ベンチマークの設計意図とモデルの挙動の相関関係を読み解きます。

3.1 数学・科学のフロンティア：FrontierMath と CritPt

AI の推論能力を測る尺度は、もはや高校数学レベル（GSM8k）や大学入試レベル（MATH）では不十分となり、専門家の研究レベルへと移行しています。

3.1.1 FrontierMath における「世代的」格差

Epoch AI が公開した FrontierMath は、未発表の専門家レベルの数学問題で構成されており、AI の「暗記」を許さない厳格なテストベッドです [10]。

- **Gemini 3 Pro の衝撃**: Gemini 3 Pro は、FrontierMath の Tier 1-3（学部〜初期大学院レベル）において**38%という驚異的な正解率を記録しました。さらに、最難関の Tier 4（研究レベル）においても 19%**を達成しています [4]。
- **比較**: 従来の最先端モデルのスコアが数%（実質的に解答不能）であったことを考慮すると、これはリニアな進化ではなく、指数関数的な飛躍です。一部の分析では、GPT-5.1 と Gemini 3 Pro の差は、GPT-4 と o1（OpenAI の初期推論モデル）の差に匹敵するとも言われています [4]。
- **WeirdML**: 特に「WeirdML」と呼ばれるカテゴリでは、Gemini 3 Pro が 69.9%を記録し、他を圧倒しています。これは、既存のパターン認識では解けない、非定型で創造的な数学的発想力が問われる領域での強さを示しています [4]。

3.1.2 CritPt ：物理学的洞察の臨界点

CritPt（Critical Point）ベンチマークは、物理学のフロンティア研究における推論能力を測定するために、50 人以上の現役研究者によって作成された 71 の複合的な研究課題です [12]。

- **スコア詳細**: Gemini 3 Pro はこのベンチマークにおいて**9.1%** のスコアを記録し、トップに立ちました [5]。対する GPT-5.1（High）は 5%程度と推定されています [13]。
- **意味合い**:「9.1%」という数字は低く見えるかもしれませんが、これは「AI が独力で物理学の新規研究を行い、査読に耐えうる洞察を導き出す」ことの難易度を物語っています。Gemini 3 Pro がここでリードしている事実は、シミュレーションデータの解釈や理論構築の補助において、現時点で最も有望なツールであることを示唆しています。

3.2 エンジニアリングの実相：SWE-Bench と Terminal-Bench

ソフトウェア開発は、AI の最も実用的な適用領域の一つです。ここでは「コードが書けるか」だけでなく、「システムを修正できるか」「ツールを使いこなせるか」が問われます。

3.2.1 SWE-Bench Verified ：拮抗するバグ修正能力

実世界の GitHub リポジトリから抽出された課題を解決する SWE-Bench Verified において、両モデルは頂上決戦を繰り広げています。

- **GPT-5.1 (Codex-Max)**: **77.9%** (xhigh effort) / 76.3% (標準) [14]。
- **Gemini 3 Pro** : **76.2%** [1]。
- **分析**: スコア上は GPT-5.1 が僅かにリードしていますが、実質的には同等レベルと言えます。しかし、詳細を見ると GPT-5.1 の方が「再現性」や「エッジケースへの対応」において、Codex-Max による微調整の効果が出ているとの評価があります [1]。開発者が「修正して」と指示した際に、期待通りに動くコードを一発で返す信頼感は GPT-5.1 に分があるようです。

3.2.2 Terminal-Bench 2.0 ：エージェントとしての身体性

コードを書くだけでなく、Linux ターミナルを操作し、ファイルシステムを管理し、コマンドを実行する能力を測るテストです。

- **GPT-5.1**: **58.1%** [1]。
- **Gemini 3 Pro** : **54.2%** [1]。
- **勝因**: ここでの GPT-5.1 の勝利は、OpenAI が「ツール使用（Tool Use）」のトレーニングに重点を置いていることの結果です。エージェントが環境を破壊せずにコマンドを実行し、エラー出力を読み取って自律的に修正する能力において、GPT-5.1 はより「熟練したオペレーター」として振る舞います。

3.2.3 LiveCodeBench ：アルゴリズムの瞬発力

競技プログラミングのような、純粋なアルゴリズム構築能力を測るベンチマークです。

- **Gemini 3 Pro** : Elo **2439** [14]。
- **GPT-5.1**: Elo ~2243[14]。
- **逆説的な結果**: 複雑なシステム修正（SWE-Bench）では GPT-5.1 が優勢である一方、ゼロからアルゴリズムを考案する力（LiveCodeBench）では Gemini が圧倒しています。これは、Gemini が「数学的・論理的ひらめき」に強く、GPT-5.1 が「実務的な遂行能力」に強いというキャラクターの違いを如実に表しています。

3.3 自律エージェントの経済性：Vending-Bench 2

Vending-Bench 2 は、エージェントが長期間にわたって自律的にタスクを行い、どれだけの価値（仮想的な金銭）を生み出したかを測定するユニークなベンチマークです。

- **Gemini 3 Pro** : Net Worth **$5,478.16** [15]。
- **GPT-5.1**: Net Worth **$1,473.43** [15]。
- **決定的差異**: この**272%** という圧倒的な差は衝撃的です。Gemini 3 Pro は、長期的な計画能力や、予期せぬ障害に対するリカバリー能力、あるいは「儲かる」判断をする戦略性において、GPT-5.1 を大きく引き離しています。これは、Gemini のロングコンテキストと深い推論能力が、単発のタスクではなく連続的な意思決定プロセスにおいて真価を発揮することを示しています。自律型トレーディングボットや経営シミュレーションのようなタスクでは、Gemini が圧倒的な利益をもたらす可能性があります。

3.4 マルチモーダルの覇権：MMMU-Pro と Video-MMMU

- **MMMU-Pro**: 高度な専門知識を要するマルチモーダル推論。Gemini 3 Pro は**81.0%** を記録し、GPT-5.1（推定 76-78%）を凌駕しました [15]。
- **Video-MMMU** : 動画理解。Gemini 3 Pro は**87.6%** を記録 [1]。
- **実用への影響**: これは、Gemini が「見て理解する」能力において人間レベル、あるいはそれ以上に達していることを意味します。例えば、工場の監視カメラ映像から異常の予兆を検知したり、長時間の会議動画から要点を抽出しつつ参加者の感情推移を分析したりするタスクにおいて、GPT-5.1 は Gemini の相手になりません。

---

4. 経済物理学：料金体系と TCO（総所有コスト）の詳細分析

性能差が明らかになったところで、次に重要なのは「コスト」です。しかし、2025 年の AI コスト分析は、単なるトークン単価の比較では不十分です。コンテキスト長による変動価格、エージェントのリトライ率、そして「時間というコスト」を含めた総合的な評価が必要です。

4.1 料金体系の構造的差異

以下の表は、両モデルの複雑な料金体系を整理したものです。

表 1: 100 万トークンあたりの単価比較（2025 年 11 月時点）

モデル	コンテキスト条件	入力 (Input)	キャッシュ入力 (Cached)	出力 (Output)	備考
GPT-5.1	全範囲	**$1.25**	$0.125	**$10.00**	標準的な Pro 向けモデル [17]
GPT-5 Pro	特殊用途	$15.00	-	$120.00	レガシー/高負荷向け。非常に高額 [19]
Gemini 3 Pro	**< 200k tokens**	$2.00	-	$12.00	短〜中規模コンテキスト [8]
Gemini 3 Pro	**> 200k tokens**	$4.00	-	$18.00	ロングコンテキスト利用時 [8]
Gemini 3 Image	-	$2.00 (Text)	-	$0.134 / 画像	画像生成・出力の特殊レート [8]

4.1.1 GPT-5.1 のコスト優位性

GPT-5.1（標準版）は、Gemini 3 Pro（<200k）と比較して、入力で約 37.5%、出力で約 17%安価です。さらに、OpenAI の強力な「Prompt Caching」機能（キャッシュ入力 $0.125）を活用すれば、頻繁に同じコンテキスト（例：システムプロンプト、コードベースの共通部分）を使用する場合、入力コストを Gemini の **16 分の 1** まで圧縮できる可能性があります[17]。これは、反復的な API コールが発生するチャットボットやコーディングアシスタントにおいて決定的な差となります。

4.1.2 Gemini の「コンテキスト税」と画像の価値

Gemini 3 Pro は、200k トークンを超えると価格が倍増（入力$4、出力$18）します[20]。これは「コンテキスト税」とも呼べるもので、100 万トークンのドキュメントを一括処理する場合、1 回あたり 4 ドル以上のコストがかかります。しかし、これを「人間の専門家が数日かけて読む資料を数秒で解析するコスト」として捉えれば、依然として破格です。また、画像出力が$0.134/枚という設定は、高品質な図解やビジュアル生成を伴うタスクにおいては明確なコスト基準となります。

4.2 エージェント運用における「隠れたコスト」

TCO（総所有コスト）を考える上で見落とされがちなのが、エージェントの「成功率」と「リトライ回数」です。

- **シナリオ A：コーディング修正**
  - GPT-5.1 は SWE-Bench や Terminal-Bench で高い信頼性を示しており、多くの場合、1 回または少ない修正回数で動作するコードを生成します。
  - Gemini 3 Pro も優秀ですが、微細なバグ修正で躓き、複数回の往復（リトライ）が発生する可能性があります。
  - **結論**: 単価の安さと成功率の高さが相まって、**コーディングエージェントの運用コストは GPT-5.1 が圧倒的に有利**です。
- **シナリオ B：複雑な意思決定（Vending-Bench）**
  - Vending-Bench 2 の結果が示す通り、Gemini 3 Pro は長期的な利益創出において GPT-5.1 を大きく上回ります。
  - GPT-5.1 が安価にタスクをこなしても、最終的な成果（利益）が少なければ意味がありません。
  - **結論**: 高度な自律判断が求められるビジネスエージェントの場合、**Gemini 3 Pro の高**

い API コストは、それが生み出す付加価値（ROI）によって正当化されます。

4.3 「時間」というコスト：レイテンシの壁

GPT-5.1 の「Instant」モードは、ユーザー体験において重要な「サクサク感」を提供します。対して Gemini 3 Pro、特に Deep Think モードは、深い思考のために数秒〜数十秒の待ち時間を要することがあります。リアルタイム性が収益に直結するアプリ（例：カスタマーサポート）では、GPT-5.1 の低レイテンシ自体がコスト削減（顧客満足度向上）につながります。

---

5. 開発者エコシステムとツールチェーン

モデルの能力を引き出すには、優れたツールが必要です。両社のアプローチはここでも対照的です。

5.1 Google Antigravity ：IDE そのものを再発明する

Google が Gemini 3 Pro に合わせて発表した「Google Antigravity」は、開発者体験を一変させる可能性を秘めています[3]。

- **概要**: Antigravity は、Gemini 3 Pro が「住む」統合開発環境（IDE）です。ここでは、人間がコードを書くのではなく、Gemini が主体となってエディタを操作し、ターミナルでコマンドを叩き、ブラウザでドキュメントを検索します。
- **特徴**: 従来の Copilot 系ツールが「補完」であったのに対し、Antigravity は「代行」です。Gemini のマルチモーダル能力を活かし、UI のプレビュー画面を見ながら CSS を修正するといった、人間と同じフィードバックループを回すことが可能です。
- **ロックインリスク**: ただし、これは Google のエコシステム（Vertex AI、Google Cloud）への深い依存を意味します。

5.2 OpenAI のエコシステム：実務への密着

OpenAI のアプローチは、既存のワークフローへの「密着」です。

- **Codex - Max の統合**: CLI ツールや IDE 拡張機能を通じて、開発者の既存の環境に静かに入り込みます。「work with app」機能により、VS Code や Cursor などのエディタとシームレスに連携し、ローカルリポジトリのコンテキストを吸い上げます[21]。
- **ユーザーセンチメント**: 開発者の間では、GPT-5.1 の方が「指示の意図を汲み取るのが上手い」「変なこだわりを持たずに素直にコードを書く」といった、実務上の使いやすさを評価する声が多く聞かれます[22]。Gemini は「賢いが、たまに独創的すぎる」という評価もあり、堅実な業務には GPT-5.1 が好まれる傾向があります。

---

6. 産業別ユースケースと推奨戦略

以上の分析に基づき、主要な産業セクターにおける最適なモデル選定戦略を提案します。

6.1 科学技術・研究開発 (R&D)

- **推奨モデル: Gemini 3 Pro**
- **理由**: FrontierMath (Tier 4) や CritPt での実績は、Gemini が未知の科学的発見を加速させる能力を持っていることを証明しています。論文の査読、実験データの解析、新規物質の候補探索など、「正解のない問い」に挑む現場では、Gemini の Deep Think とロングコンテキストが不可欠です。

6.2 ソフトウェアエンジニアリング・DevOps

- **推奨モデル: GPT-5.1 Pro (Codex - Max)**
- **理由**: 既存コードの保守、バグ修正、テストコードの生成といった日常的な開発業務には、コスト効率と信頼性に優れる GPT-5.1 が最適です。特に、キャッシュ機能を活用した CI/CD パイプラインへの組み込みは、開発コストを大幅に削減します。
- **例外**: 全く新しいフレームワークを用いたアーキテクチャ設計や、アルゴリズムの最適化が必要な場合は、Gemini 3 Pro の Antigravity 環境をスポットで利用するのが効果的です。

6.3 クリエイティブ・メディア制作

- **推奨モデル: Gemini 3 Pro**
- **理由**: 動画、画像、テキストを自在に行き来するマルチモーダル能力は、コンテンツ制作のプロセスを激変させます。映画の脚本から絵コンテを生成したり、逆にラフスケッチから Web サイトのコードを生成したりする作業は、Gemini の独壇場です。

6.4 金融・ビジネスインテリジェンス

- **推奨モデル: Gemini 3 Pro (分析・戦略) / GPT-5.1 (定型処理)**
- **理由**: 市場動向の分析や長期的な投資戦略の立案（Vending-Bench 的なタスク）には、Gemini の深い推論能力とロングコンテキスト（年次報告書の一括読込）が威力を発揮します。一方で、日次のレポート作成や顧客対応チャットボットには、高速で安価な GPT-5.1 が適しています。

---

7. 結論と未来への展望

GPT-5.1 Pro と Gemini 3 Pro の比較は、もはや「どちらが優れているか」という単純な問いでは語れません。両者は、AI の進化における異なるベクトルを代表しています。

- **GPT-5.1 Pro** は、**「実務的完成度」の頂点**です。それは、人間社会の既存のタスクを効率化し、開発者をスーパーマンにし、ビジネスの速度を加速させる、信頼できるパートナーです。コストパフォーマンスと使い勝手の良さは、企業の AI 導入におけるデファクトスタンダードであり続けるでしょう。
- **Gemini 3 Pro** は、**「知能の拡張」への挑戦**です。それは、人間がまだ解けていない問題を解き、見えていない関係性を可視化する、天才的な賢者です。その能力は、コストやレイテンシという制約を超えて、人類の知のフロンティアを押し広げるために存在します。

最終的な提言

企業や開発者は、これら二つのモデルを排他的に選ぶのではなく、適材適所で組み合わせる**「ハイブリッド戦略」を採用すべきです。
日常のコーディングや顧客対応には GPT-5.1 を配備してコストと品質を担保し、R&D 部門や戦略立案プロセスには Gemini 3 Pro**を導入してイノベーションの種を探す。この使い分けこそが、2025 年以降の AI 時代における競争優位の源泉となるでしょう。
私たちは今、信頼できる労働力としての AI と、深遠なる思考者としての AI、その両方を手に入れたのです。

免責事項: 本レポートに含まれる価格、ベンチマークスコア、および仕様は、2025 年 11 月時点の公開情報[1]に基づいています。AI 技術は急速に進歩しており、最新の情報は各社の公式ドキュメントを参照してください。

引用文献

1. GPT-5.1-Codex-Max vs Gemini 3 Pro: Next-Generation AI Coding Titans - Medium, 11 月 23, 2025 にアクセス、 https://medium.com/@leucopsis/gpt-5-1-codex-max-vs-gemini-3-pro-next-generation-ai-coding-titans-877cc9054345
2. I just tested Gemini 3 vs ChatGPT-5.1—and one AI crushed the competition | Tom's Guide, 11月 23, 2025 にアクセス、 https://www.tomsguide.com/ai/i-just-tested-gemini-3-vs-chatgpt-5-1-and-one-ai-crushed-the-competition
3. Gemini 3 Pro vs GPT 5.1: which is better? A Complete Comparison- CometAPI - All AI Models in One API, 11月 23, 2025 にアクセス、 https://www.cometapi.com/gemini-3-pro-vs-gpt-5-1-which-is-better-a-complete-comparison/
4. Gemini 3 Pro set a new record on FrontierMath: 38% on Tiers 1–3 ..., 11月 23, 2025 にアクセス、 https://www.reddit.com/r/accelerate/comments/1p3qwet/gemini_3_pro_set_a_new_record_on_frontiermath_38/
5. Chubby ♨ on X: "Gemini 3.0 Pro tops new physics benchmark at 9.1% CritPt is a new graduate-level frontier physics benchmark built by over 60 researchers to test models on truly novel, research-grade problems across 11 physics subfields—and no current model scores above 9%. Even top systems / X : r - Reddit, 11 月 23, 2025 にアクセス、 https://www.reddit.com/r/accelerate/comments/1p3gnat/chubby_on_x_gemini_30_pro_tops_new_physics/
6. Gemini 3.0 vs GPT-5.1: a clear comparison for builders - Portkey, 11 月 23, 2025 にアクセス、 https://portkey.ai/blog/gemini-3-0-vs-gpt-5-1/
7. Simon Willison's Weblog, 11 月 23, 2025 にアクセス、 https://simonwillison.net/
8. Gemini 3 Developer Guide | Gemini API - Google AI for Developers, 11 月 23, 2025 にアクセス、 https://ai.google.dev/gemini-api/docs/gemini-3
9. Gemini 3.0 vs GPT-5.1 vs Claude Sonnet 4.5: Which one is better? - Bind AI IDE, 11

月 23, 2025 にアクセス、 https://blog.getbind.co/2025/11/19/gemini-3-0-vs-gpt-5-1-vs-claude-sonnet-4-5-which-one-is-better/
10. FrontierMath: LLM Benchmark for Advanced AI Math Reasoning, 11 月 23, 2025 にアクセス、 https://epoch.ai/frontiermath
11. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI - arXiv, 11 月 23, 2025 にアクセス、 https://arxiv.org/abs/2411.04872
12. [2509.26574] Probing the Critical Point (CritPt) of AI Reasoning: a Frontier Physics Research Benchmark - arXiv, 11 月 23, 2025 にアクセス、 https://arxiv.org/abs/2509.26574
13. LLM Performance Leaderboard - a Hugging Face Space by ArtificialAnalysis, 11 月 23, 2025 にアクセス、 https://huggingface.co/spaces/ArtificialAnalysis/LLM-Performance-Leaderboard
14. Gemini 3.0 Pro vs GPT 5.1-Codex-Max: Tried Python Coding : r/GeminiAI - Reddit, 11 月 23, 2025 にアクセス、 https://www.reddit.com/r/GeminiAI/comments/1p3td75/gemini_30_pro_vs_gpt_51codexmax_tried_python/
15. Gemini 3.0 Pro vs GPT 5.1: LLM Benchmark Showdown : r/ArtificialInteligence - Reddit, 11 月 23, 2025 にアクセス、 https://www.reddit.com/r/ArtificialInteligence/comments/1p0c3vc/gemini_30_pro_vs_gpt_51_llm_benchmark_showdown/
16. Google Gemini 3 Benchmarks - Vellum AI, 11 月 23, 2025 にアクセス、 https://www.vellum.ai/blog/google-gemini-3-benchmarks
17. Gemini 3 Vs ChatGPT 5.1: Benchmarks, Pricing And Real-World Use - AceCloud, 11 月 23, 2025 にアクセス、 https://acecloud.ai/blog/gemini-3-vs-chatgpt-5-1/
18. API Pricing - OpenAI, 11 月 23, 2025 にアクセス、 https://openai.com/api/pricing/
19. Azure OpenAI Service - Pricing, 11 月 23, 2025 にアクセス、 https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/
20. Gemini 3 Pro - Everything you need to know - Artificial Analysis, 11 月 23, 2025 にアクセス、 https://artificialanalysis.ai/articles/gemini-3-pro-everything-you-need-to-know
21. GPT-5 Thinking vs Gemini 2.5 pro review (for scientific applications) : r/OpenAI - Reddit, 11 月 23, 2025 にアクセス、 https://www.reddit.com/r/OpenAI/comments/1mwqqfg/gpt5_thinking_vs_gemini_25_pro_review_for/
22. How does GPT 5 Pro (the $200 version) stack up to Gemini 3?, 11 月 23, 2025 にアクセス、 https://www.reddit.com/r/ChatGPTPro/comments/1p1bnpi/how_does_gpt_5_pro_the_200_version_stack_up_to/