GPT-5.1 Pro と Gemini 3 Pro の徹底比較レポート

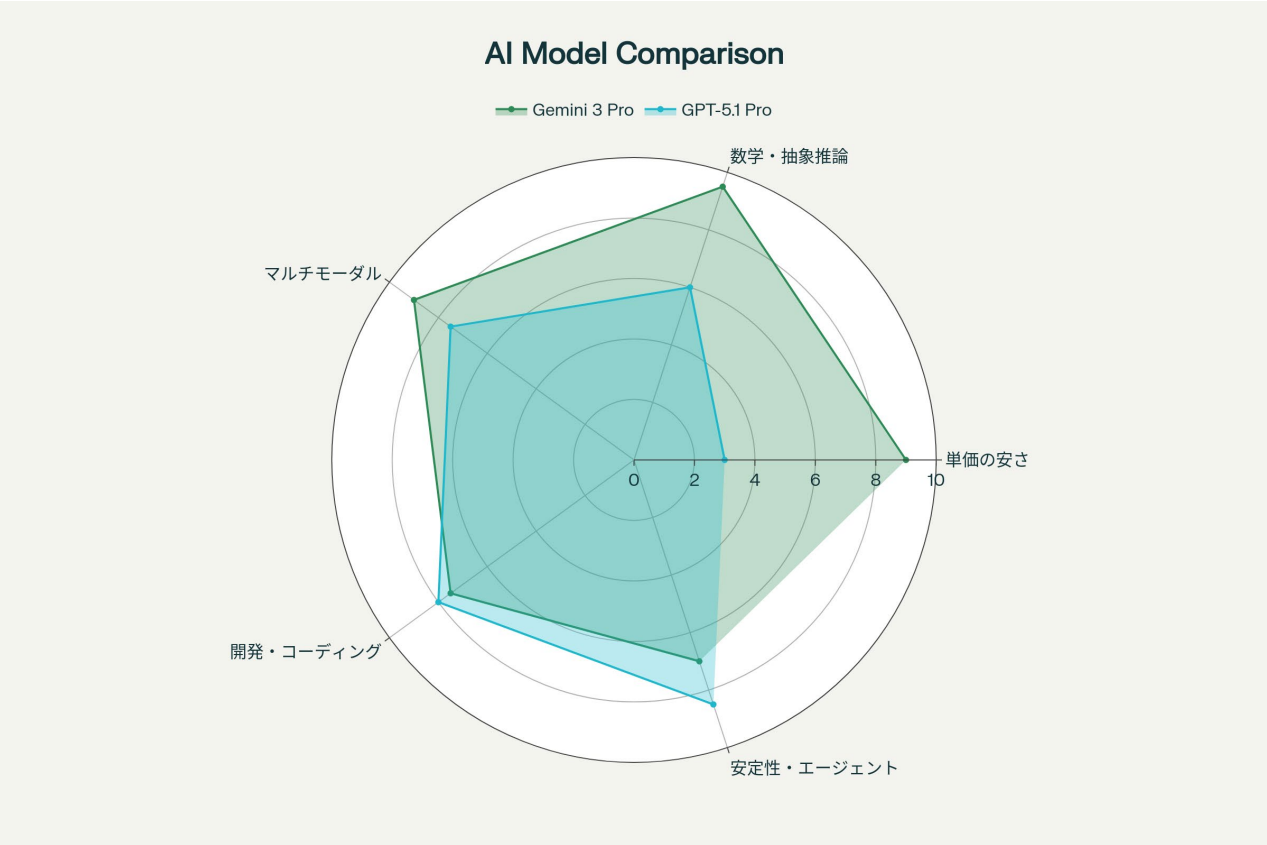
2025 年 11 月、AI 業界に大きな動きがありました。OpenAI が「GPT-5.1 Pro」を、Google が Gemini 3 シリーズの最新フラッグシップモデル「Gemini 3 Pro」を相次いで発表したのです。両モデルとも、推論能力、マルチモーダル理解、コーディング性能において新たな水準を打ち立てたと主張していますが、その実力と適性は大きく異なります。本レポートでは、公式発表、第三者ベンチマーク、実ユーザーの評価、料金体系を総合的に分析し、両モデルの真の実力を明らかにします。

主要な結論

料金構造の観点では、Gemini 3 Pro が圧倒的に優位です。入力トークンで 3.75〜7.5 倍、出力トークンで 6.7〜10 倍という価格差があり、大量のトークンを処理するワークロードでは総コストが 88〜90%削減される可能性があります。しかし、**実務的な開発ワークフローとエージェント実行の安定性**では、GPT-5.1 Pro が依然として強みを持ちます。特に長時間稼働するコーディングエージェント、リポジトリ全体の修正、エラー耐性が求められる場面では、GPT-5.1 Pro の方が再試行回数を減らし、結果的に運用効率を高める可能性があります。[1][2][3][4][5][6]

ベンチマーク性能の観点では、Gemini 3 Pro が数学・抽象推論、マルチモーダル理解、視覚的推論において顕著な優位性を示しています。Humanity's Last Exam で 37.5%（GPT-5.1 は 26.5%）、ARC-AGI-2 で 31.1%（GPT-5.1 は 17.6%）、MMMU-Pro で 81.0%（GPT-5.1 は 76.0%）と、多くの高難度ベンチマークで 5〜11 ポイントの差をつけています。一方、GPT-5.1 Pro は、SWE-bench Verified で 76.3%とわずかに上回り（Gemini 3 Pro は 76.2%）、実務的なソフトウェアエンジニアリングタスクにおける信頼性を示しています。[7][8][9][10][11]

最適なユースケースは明確に分かれます。推論中心、分析、文書生成、長文コンテキスト処理、マルチモーダルタスクでは、Gemini 3 Pro の方が費用対効果が高くなります。一方、大規模開発、24 時間連続エージェント実行、リポジトリスケールのリファクタリング、エラー耐性が重視される環境では、GPT-5.1 Pro とその Codex-Max 拡張が依然として最適な選択肢となります。[3][12][13][6]



GPT-5.1 Pro と Gemini 3 Pro の能力特性比較：5 つの主要評価軸における相対的な強弱を示すレーダーチャート

1. 基本スペックとアーキテクチャ

1.1 リリース日と開発体制

GPT-5.1 Pro は、OpenAI が 2025 年 11 月 19 日に発表した、GPT-5 シリーズの段階的アップグレード版です。GPT-5 の「賢いが冷たい」という評価を受け、会話の自然さと推論能力の両立を目指して開発されました。同時に、24 時間連続稼働可能な「GPT-5.1-Codex-Max」も発表され、プロジェクト規模の開発タスクへの対応を強化しました。[1][14][15][16][12][13]

Gemini 3 Pro は、Google が 2025 年 11 月 18 日に発表した、Gemini 3 シリーズの最初のモデルです。Google DeepMind によって開発され、「どんなアイデアも実現できる」をコンセプトに、推論能力、マルチモーダル理解、エージェント機能を大幅に強化しました。同時に、Google Antigravity という新しいエージェント開発環境も公開され、Gemini 3 Pro を中核とした開発ワークフローを提供しています。[17][18][19][20][21][22][23]

1.2 アーキテクチャと技術的特徴

GPT-5.1 Pro のアーキテクチャについて、OpenAI は正式なパラメータ数を公開していません。独立した分析では、1.7〜1.8 兆パラメータ程度の密なモデル（dense model）、または数十兆パラメータ規模の Mixture-of-Experts（MoE）アーキテクチャである可能性が示唆されています。GPT-5.1 の最大の特徴は「適応的推論（adaptive reasoning）」です。タスクの複雑さに応じて、モデルが動的に思考時間を調整する能力を持ちます。簡単なタスクでは素早く応答し、複雑なタスクでは内部的により多くのトークンを生成して深く考えます。[16][24][3][7][25][26]

この適応的推論は、GPT-5.1 Instant と GPT-5.1 Thinking という 2 つのバリエーションで提供されます。Instant は日常的な会話や簡単なタスクに最適化され、Thinking は高度な推論や複数ステップのタスクに対してより長い思考時間を割り当てます。コンテキストウィンドウは 128,000 トークン（入力）、出力は最大 16,384 トークンです。API 経由では最大 400,000 トークン（入力 272,000 トークン、出力 128,000 トークン）の組み合わせも可能とされています。[15][24][3][27][28][29][30][7][16]

Gemini 3 Pro のアーキテクチャは、Sparse Mixture-of-Experts（疎な専門家混合）モデルです。この構造では、各トークンが動的に一部の専門家パラメータにルーティングされ、全体のパラメータ数を増やしながらも、トークンあたりの計算コストを抑えることができます。独立した分析では、Gemini 3 Pro の総パラメータ数は 1 兆以上と推定されています。ネイティブマルチモーダルサポートを持ち、テキスト、画像、音声、動画を統一されたコンテキストで処理できます。[19][31][32][33][34][35]

コンテキストウィンドウは 1,048,576 トークン（約 1M トークン）という圧倒的な容量を持ち、出力は最大 65,536 トークン（64k）です。これは、巨大なコードベース、数百ページの文書、数時間分の動画トランスクリプトを単一のプロンプトで処理できることを意味します。Gemini 3 Pro は、Gemini 2.5 Pro のファインチューニングではなく、ゼロからトレーニングされた新モデルです。トレーニングデータには、大規模なウェブテキスト、多言語コード、画像、音声、動画、ライセンスデータ、ユーザーインタラクションデータ、合成データが含まれ、マルチモーダル指示チューニングと強化学習（RLHF/RLAIF）によって後処理されています。[3][28][4][8][31]

1.3 料金体系の詳細比較

両モデルの料金体系は、用途によって総コストに大きな影響を与えます。

GPT-5.1 Pro の API 料金は、入力 1M（100 万）トークンあたり 15.00 ドル、出力 1M トークンあたり 120.00 ドルです。この料金は、コンテキスト長に関わらず一律です。キャッシング機能を使用す

ると、キャッシュされた入力トークンのコストが 90%削減されます。ChatGPT Pro プランの月額料金は 200 ドルで、GPT-5.1 Pro へのアクセスが含まれます。[36][37][14][38][16][39][3][28]

Gemini 3 Pro の API 料金は、コンテキスト長によって段階的に変動します。コンテキストが 200,000 トークン以下の場合、入力 1M トークンあたり 2.00 ドル、出力 1M トークンあたり 12.00 ドルです。200,000 トークンを超える場合、入力 1M トークンあたり 4.00 ドル、出力 1M トークンあたり 18.00 ドルに増加します。Google AI Pro プラン（旧 Gemini Advanced）の月額料金は 19.99 ドルで、Gemini 2.5 Pro への優先アクセスと Deep Research が含まれますが、Gemini 3 Pro は上位の Ultra プラン（月額 36,400 円≈240 ドル）でのみ利用可能です。[2][40][4][5][41][42]

コストシミュレーション分析を見ると、料金差の実態が明確になります。

シナリオ 1：文書分析タスク（入力 150k トークン、出力 30k トークン）

- Gemini 3 Pro: 0.66 ドル
- GPT-5.1 Pro: 5.85 ドル
- **節約額: 5.19 ドル（88.7%安い）**[43]

シナリオ 2：大規模コンテキスト分析（入力 500k トークン、出力 50k トークン）

- Gemini 3 Pro: 2.90 ドル
- GPT-5.1 Pro: 13.50 ドル
- **節約額: 10.60 ドル（78.5%安い）**[43]

シナリオ 3：コード生成タスク（入力 100k トークン、出力 100k トークン）

- Gemini 3 Pro: 1.40 ドル
- GPT-5.1 Pro: 13.50 ドル
- **節約額: 12.10 ドル（89.6%安い）**[43]

シナリオ 4：月間ヘビーユーザー（入力 10M トークン、出力 5M トークン）

- Gemini 3 Pro: 80.00 ドル
- GPT-5.1 Pro: 750.00 ドル
- **節約額: 670.00 ドル（89.3%安い）**[43]

これらのシミュレーションから、純粋な従量課金ベースでは、Gemini 3 Pro が圧倒的に安価であることがわかります。特に出力トークンが多いタスク（コード生成、長文執筆、詳細な分析レポート作成）では、コスト差が顕著です。[3][2][43]

2. ベンチマーク性能の徹底分析

2.1 総合指標と ELO スコア

LMArena ELO スコアは、実際のユーザー評価に基づくペアワイズ比較ランキングです。Gemini 3 Pro は 1501 ELO を達成し、史上初めて 1500 の壁を突破したモデルとなりました。これは、GPT-5.1 の推定 1450〜1460 ELO を約 40〜50 ポイント上回る結果です。この差は、実際の対話品質、回答の自然さ、指示追従能力において、ユーザーが Gemini 3 Pro を一貫して好んだことを示しています。[44][9][45][46][47]

Epoch Capabilities Index（ECI）は、複数のベンチマークを統合した総合能力指標です。Gemini 3 Pro は 154 ポイントを記録し、GPT-5.1 の 151 ポイントを上回って首位に立ちました。これは、Google モデルが初めて Epoch AI の総合能力リーダーボードで首位を獲得した歴史的な瞬間です。ECI は、GPQA、ARC-AGI、FrontierMath、SimpleQA など、多様なベンチマークの結果を統合して算出されるため、特定のタスクに偏らない総合的な能力を反映しています。[48][10][49][50]

2.2 推論能力と高難度ベンチマーク

Humanity's Last Exam（人類最後の試験）は、最も困難な推論ベンチマークの一つです。このテストは、多様な学問分野にわたる大学院レベル以上の問題で構成され、AI の推論能力の限界を試すものです。

Gemini 3 Pro は、ツールなしで 37.5%の正答率を記録しました。これは、GPT-5.1 の 26.5%を 11 ポイント上回る結果です。さらに、Gemini 3 Pro Deep Think モード（上位契約で利用可能な強化推論モード）では 41.0%まで向上し、3.5 ポイントの追加的な改善を示しました。この結果は、複雑な意思決定、科学的仮説生成、多段階の法的分析、戦略的ビジネス計画など、高度な推論チェーンを必要とするアプリケーションにおいて、Gemini 3 Pro が優位性を持つことを示唆しています。[8][9][11][51][52]

ARC-AGI-2（抽象的視覚推論パズル）は、言語に依存しない抽象的推論能力を測定します。Gemini 3 Pro は 31.1%を記録し、GPT-5.1 の 17.6%をほぼ 2 倍近く上回りました。Deep Think モードでは 45.1%まで向上し、14 ポイントの改善を見せました。Gemini 2.5 Pro の 4.9%と比較すると、

Gemini 3 世代での抽象的推論能力の飛躍的な進化が明らかです。この結果は、新しい問題パターンへの適応、視覚的情報からの推論、言語的手がかりに依存しない純粋な論理的思考において、Gemini 3 Pro が卓越していることを示しています。[53][9][11][51][8]

2.3 科学的知識と数学的推論

GPQA Diamond（博士レベル科学問題）は、物理学、化学、生物学の博士レベルの専門知識を要する問題で構成されています。Gemini 3 Pro は 91.9%の正答率を達成し、GPT-5.1 の 88.1%を 3.8 ポイント上回りました。Deep Think モードでは 93.8%まで向上し、異常なほど高い水準に達しました。この結果は、科学研究、実験データ分析、新規仮説の評価において、Gemini 3 Pro が強力なツールとなることを示しています。[8][53][9][11][51][52]

AIME 2025（数学オリンピック）では、ツールなしでの性能差が顕著でした。Gemini 3 Pro は 95.0%の正答率を記録し、GPT-5.1 の推定 71%を 24 ポイント上回りました。コード実行を許可した場合、両モデルとも 100%に達しましたが、ツールなしでの強力なベースライン性能は、Gemini 3 Pro がより堅牢な数学的直感を持つことを示しています。これは、外部計算ツールへのアクセスが制限される環境や、レイテンシが重視されるリアルタイム数学指導において重要な利点です。
[53][51][52][8]

MathArena Apex は、最も困難な競技数学問題のベンチマークです。Gemini 3 Pro は 23.4%のスコアを記録し、GPT-5.1 の 1.0%を 22.4 ポイント上回りました。Claude Sonnet 4.5 の 1.6%、Gemini 2.5 Pro の 0.5%と比較すると、20 倍以上の飛躍的な改善です。このタスクはどのモデルにとっても「解決済み」とは言えませんが、Gemini 3 Pro が最も前進したモデルであることは明らかです。[11][8][53]

FrontierMath は、専門の数学者が数時間から数日かけて解く研究レベルの数学問題で構成されています。Tier 1-3（学部から初期大学院レベル）では、Gemini 3 Pro が 37.6%（290 問中 109 問正解）を記録し、GPT-5（high）の 32.4%、GPT-5.1（high）の 31.0%を大きく上回りました。最難関の Tier 4（研究レベル数学）では、Gemini 3 Pro が 18.8%（48 問中 9 問正解）を達成し、GPT-5.1、GPT-5、GPT-5 Pro がすべて 12.5%（6 問正解）で並んだのに対し、6.3 ポイントの差をつけました。
[10][49][54][50]

2.4 マルチモーダル理解と視覚的能力

MMMU-Pro（マルチモーダル理解と推論）は、大学レベルの多様な科目にわたるマルチモーダル推論を測定します。Gemini 3 Pro は 81.0%を記録し、GPT-5.1 の 76.0%を 5 ポイント上回りました。

Claude Sonnet 4.5 の 68.0%、Gemini 2.5 Pro の 68.0%と比較すると、マルチモーダル理解における大きな飛躍が見られます。[8][53][9][11]

Video-MMMU（動画理解）では、Gemini 3 Pro が 87.6%を記録し、GPT-5.1 の推定 80〜82%を約 5〜7 ポイント上回りました。これは、動画の時間的・空間的情報を統合して理解する能力において、Gemini 3 Pro が優れていることを示しています。動画講義の分析、複雑な UI 操作のスクリーンショット理解、混合メディアを含む文書処理（チャート、図表、テキスト）において、この能力は実用的な価値があります。[53][9][51][8]

ScreenSpot-Pro（画面理解）では、Gemini 3 Pro が 72.7%という圧倒的なスコアを記録しました。これは、Claude Sonnet 4.5 の 36.2%、GPT-5.1 の 3.5%を大きく上回る結果です。この性能は、UI 自動化、スクリーンショットからのコード生成、プロトタイプ作成において、Gemini 3 Pro が特に強力であることを示しています。[44][53]

2.5 コーディング性能とソフトウェアエンジニアリング

コーディング能力の評価は、ベンチマークの種類によって異なる結果を示します。

SWE-bench Verified は、実世界の GitHub issue を解決する能力を測定します。ここでは、GPT-5.1（Thinking high）が 76.3%を記録し、Gemini 3 Pro の 76.2%をわずか 0.1 ポイント上回りました。Claude Sonnet 4.5 は 77.2%で最高スコアを記録しています。この結果は、実務的なソフトウェアエンジニアリングタスク、特に既存コードベースのバグ修正において、3 モデルがほぼ同等の能力を持つことを示しています。[7][8][55][9]

一方、**LiveCodeBench Pro（競技プログラミング）**では、結果が大きく異なります。Gemini 3 Pro の ELO スコアは 2439 で、GPT-5.1 の 2243 を 196 ポイント上回りました。Claude Sonnet 4.5 の 1418 と比較すると、1021 ポイントという圧倒的な差があります。この結果は、アルゴリズム的問題解決、ゼロからの複雑なコード生成において、Gemini 3 Pro が優れていることを示唆しています。[8][9]

Terminal-Bench 2.0 は、ターミナル操作を伴う実世界の複雑なタスクにおける自律型 AI エージェントの性能を評価します。Gemini 3 Pro は 54.2%のスコアを記録し、GPT-5.1 の 47.6%、Gemini 2.5 Pro の 32.6%を上回りました。これは、ターミナルコマンドの実行、システム状態の確認、多段階のデバッグ操作を含むエージェント的なコーディングタスクにおいて、Gemini 3 Pro が優位性を持つことを示しています。[56]

2.6 信頼性と幻覚率

AA-Omniscience Benchmark は、AI モデルの事実検索の信頼性を測定する新しい評価基準です。このベンチマークは、6,000 問の質問を 42 のトピック（ビジネス、人文・社会科学、健康、法律、ソフトウェアエンジニアリング、科学・数学）にわたって出題します。[57][58]

正答率（Accuracy）では、Gemini 3 Pro が 53%を記録し、Grok 4 と GPT-5.1（high）の 39%を 14 ポイント上回りました。これは、Gemini 3 Pro がより広範な事実知識を持つことを示しています。しかし、**幻覚率（Hallucination Rate）**では、Gemini 3 Pro が 88%という高い値を示しました。これは、間違った回答のうち 88%が「わからない」と答える代わりに、自信を持って誤答を提供することを意味します。[59][58][60][57]

GPT-5.1（high）の幻覚率は 81%、Grok 4 は 64%でした。Claude 4.1 Opus は 36%の正答率で最も低い幻覚率を示し、不確実性を認めることにおいて最もバランスが取れています。Gemini 2.5 Pro と Gemini 2.5 Flash も 88%の幻覚率を示しており、この問題は Gemini 3 世代でも改善されていません。[57][59]

Omniscience Index（正答にプラスポイント、誤答にマイナスポイント、「わからない」回答にはゼロポイント）では、Gemini 3 Pro が 13 ポイントを記録し、Claude 4.1 Opus の 4.8 ポイント、GPT-5.1 を上回って首位に立ちました。ただし、この結果は主に正答率の高さによるものであり、幻覚率の高さという弱点は依然として残っています。[58][57]

ユーザーの実体験では、Gemini 2.5 Pro が幻覚を起こすだけでなく、その誤りに対して頑固に固執し、修正が困難であったという報告があります。Gemini 3 Pro については、幻覚率は変わらないものの、正答する頻度が高いため、全体的な有用性は向上しているとの評価があります。一方、GPT-5.1 は、より慎重で注意深いアプローチを取る傾向があり、スライドデッキやレポートに直接使用する情報が必要な場合には、この慎重さが重要だと指摘されています。[60][11][59]

3. 実用的な特性と開発者体験

3.1 エージェント性能と長時間稼働

GPT-5.1 Codex-Max は、長時間稼働するコーディングエージェント向けに特化したモデルです。最大の特徴は「コンパクション（compaction）」技術で、これはコンテキストウィンドウの制限に近づくと、セッション履歴を自動的に整理し、重要なコンテキストを保持しながら新しいウィンドウを確保

します。OpenAI の内部評価では、このモデルが 24 時間以上にわたってタスクに取り組み続ける様子が観察されています。[12][13][61][62][63]

独立評価機関 METR の報告では、GPT-5.1-Codex-Max の「観測 50%時間地平線」（モデルが一貫してタスクを維持できる中央値時間）は約 2 時間 40 分で、GPT-5 の 2 時間 17 分から改善しています。ただし、信頼区間が広く、タスクによって変動が大きいことも指摘されています。この持続性は、プロジェクト規模のリファクタリング、深いデバッグセッション、数時間から一日以上にわたるエージェントループに耐えることを意味します。[13][61][62][63][12]

Gemini 3 Pro と Google Antigravity は、異なるアプローチを取ります。Antigravity は、AI エージェントを指揮・監督するために設計されたエージェントファーストの IDE（統合開発環境）です。エージェントが「計画 → 実装 → 検証」というサイクルを自律的に遂行し、その過程で「アーティファクト」（タスクリスト、実装計画、スクリーンショット、ブラウザ録画など）を生成します。これらのアーティファクトにより、開発者はエージェントの論理を一目で検証でき、生のツールコールをスクロールする必要がありません。[21][22][64][23]

Antigravity は、エディタ、ターミナル、ブラウザを横断してエージェントが動作し、複数のワークスペースで並行して複数のエージェントを実行できる「Agent Manager」インボックスを提供します。また、エージェントの過去の作業とフィードバックから学習する「グローバルナレッジベース」を持ち、知識管理をコアプリミティブとして扱います。Antigravity は、Gemini 3 Pro だけでなく、Claude Sonnet 4.5 や OpenAI のモデルもサポートしており、開発者に選択肢を提供しています。[22][21]

3.2 実プロジェクトでの比較評価

実際の開発プロジェクトで GPT-5.1（high）、GPT-5.1-Codex-Max（high）、Codex-Max（extra high）を比較した詳細なレポートがあります。このプロジェクトは、バックエンド、フロントエンド、大量の文書フォルダーを持つ中程度に複雑な Web アプリケーションで、過去数週間の変更（新機能実装、テスト追加、ルーティングリファクタリング、文書更新）を評価するタスクでした。[6]

GPT-5.1（high）は、詳細で物語的な応答を提供し、主要な成果を正確に強調しました。複雑なルーティングの詳細を正確に把握し、構造的・プロセス的な懸念（文書ソースの信頼性、テストインフラ、レガシーコード廃止のタイミング）に対処しました。応答は長かったものの、技術リーダーからの思慮深い振り返りとロードマップのように感じられました。[6]

GPT-5.1-Codex-Max（high/extra high）は、簡潔で実行可能なリストを生成しましたが、各提案をリポジトリと文書に対して検証する必要がありました。少なくとも 1 件のケースでは、既に解決済みのルートに関する冗長なタスクが提案されました。簡潔さは有益ですが、正確性が最も重要であり、このユーザーは以下の結論に達しました：[6]

- アーキテクチャと計画、重要領域のコード変更（バックエンドロジック、ルーティング、認証、データベース、コンプライアンス関連フロー）、振り返りとロードマップタスクには **GPT-5.1（high）**を使用
- 明確に定義された実装タスク、エッジケースが少ない領域、速度が重視される場面では **Codex-Max/Extra High** を選択的に使用
- 本番動作に影響する部分では、信頼の階層を「リポジトリ・テスト・文書 → GPT-5.1（high）の解釈 → 他のモデルは検証が必要な有用だが誤りやすいアシスタント」とする[6]

別のユーザーは、計画とプロンプティングには **Gemini 3** が最適で、コードレビューで重要な変更を特定するのに役立ったと報告しています。実際の実装では、**GPT-5.0** が GPT-5.1（high）よりもエラーが少なく、ユニットテストと統合テストを通じて自己解決する傾向があると述べています。GPT-5.1-Codex-Max については、GPT-5.0 よりもさらに効率的に実行するが、バックエンド実装で重大なバグを見逃したり、ユニットテストの失敗にもかかわらず良好なパフォーマンスを示したケースもあったと報告しています。[6]

3.3 速度とレイテンシ

Gemini 3 Pro は、出力速度が 130 トークン/秒に達し、フロンティアモデルの中でも高速な部類に入ります。これは、インタラクティブなアプリ、UI アシスタント、大量のチャットワークロードにおいて応答性が重要な場面で有利です。マルチモーダルタスクでは、Gemini 3 が GPT-5.1 と複数ツールを組み合わせる場合と比較して、完全な出力を 40%速く提供したという報告があります。[65][66]

GPT-5.1 は、適応的推論により、簡単なタスクでの応答時間が劇的に短縮されました。OpenAI の例では、単純な npm リストの回答が GPT-5 の約 10 秒から GPT-5.1 の約 2 秒に短縮されました。中程度の難易度タスクでは、出力トークン数が約 7,000 から約 4,000 に減少し、同じ結果を得るのに約 43%少ないトークンで済むようになりました。ただし、Codex-Max は「思考」プロセスがトークン効率的である一方、かなりの計算時間を必要とするため、即座にスニペットが必要な場合は軽量モデルの方が良い可能性があります。[3][7][12]

3.4 特定タスクへの適性評価

実際のユーザーレポートと第三者評価に基づく、タスク別の最適モデル推奨は以下の通りです：[67][68][66][69][9][6][11][51]

Gemini 3 Pro が最適：

- 長文コンテキスト処理（200k〜1M トークンの文書、コードベース、動画トランスクリプト分析）
- マルチモーダルタスク（画像・動画・音声を含む統合的な理解、UI 分析、スクリーンショットからのコード生成）
- 高難度数学・抽象推論（研究レベルの数学、視覚パズル、博士レベルの科学問題）
- 競技プログラミング・アルゴリズム問題（ゼロからの複雑なコード生成）
- コスト重視の大量処理（推論、分析、要約タスクで従量課金を削減したい場合）

GPT-5.1 Pro が最適：

- 大規模開発ワークフロー（リポジトリ全体の修正、プロジェクトスケールのリファクタリング）
- 長時間エージェント実行（24 時間連続稼働、多段階デバッグ、CI/CD ライクなワークフロー）
- 実務的なバグ修正（SWE-bench タイプの実世界の GitHub issue）
- 安定性とエラー耐性が重要な場面（再試行コストを抑えたい本番環境）
- ビジネス文書作成（信頼性が重視されるスライド、レポート、プレゼンテーション資料）

両モデルが同等：

- 標準的なコーディング支援（コード補完、小規模リファクタリング、コードレビュー）
- 一般的な対話・Q&A（日常的な質問応答、情報検索）
- 文書要約（数万トークン程度の標準的な要約タスク）

Claude Sonnet 4.5 が最優位：

- SWE-bench スタイルの実世界の開発タスク（77.2%で 3 モデル中最高）
- 幻覚率を最小化したい場面（低い幻覚率と高い慎重さ）

4. 総合的なコストパフォーマンス分析

4.1 料金単価とトークン効率

純粋な料金単価では、**Gemini 3 Pro が圧倒的に優位**です。入力トークンで 3.75〜7.5 倍、出力トークンで 6.7〜10 倍という価格差があり、ほとんどのワークロードで総コストを 78〜90%削減できます。特に、出力トークンが多いタスク（長文生成、詳細なコード生成、包括的な分析レポート）では、この差が最も顕著です。[2][4][5][43]

トークン効率の観点では、GPT-5.1 の適応的推論が中程度の難易度タスクで約 43%のトークン削減を実現しています。つまり、同じ結果を得るのに必要なトークン数が減少します。Codex-Max も、主要なベンチマークで前モデルと比較して約 30%少ないトークンでより良い性能を達成しています。[70][7][12]

しかし、トークン単価だけでは全体像を捉えられません。**再試行コスト**と**運用効率**を考慮する必要があります。

4.2 安定性と再試行コストの考慮

GPT-5.1 Pro は、「大規模なソフトウェア開発、長時間エージェント実行、リポジトリ全体の修正、コードの品質や再現性といった実務的な開発ワークロード」において最も安定しており、長い開発セッションを安全に走らせる点で依然として強いとされています。実際のユーザーレポートでは、GPT-5.1（high）が複雑なルーティングの詳細を正確に把握し、冗長なタスクを提案することが少なかったと報告されています。[6][11]

一方、Gemini 3 Pro は、高い正答率を持つものの、88%という高い幻覚率が懸念されます。複雑で多段階のコーディングやエージェント的なタスクでは、「自信を持った誤答」が 1 回発生すると、大きな変更や多数の問題を引き起こす可能性があります。SWE-bench での性能が示すように、Gemini 3 Pro は頑固で障害に直面すると苦労する傾向があるという指摘もあります。[57][59][58][60]

再試行コストのシミュレーションを考えてみましょう。

シナリオ A：推論・分析タスク（1 回で成功する確率が高い）

- Gemini 3 Pro: 1 回の試行で 0.66 ドル、成功率 90%と仮定 → 期待コスト: 0.73 ドル
- GPT-5.1 Pro: 1 回の試行で 5.85 ドル、成功率 95%と仮定 → 期待コスト: 6.16 ドル

- **結果: Gemini 3 Pro が 88%安い**

シナリオ B：複雑なコーディングタスク（エラーや修正の可能性が高い）

- Gemini 3 Pro: 1 回の試行で 1.40 ドル、成功率 70%と仮定 → 期待コスト: 2.00 ドル（1.43 回の平均試行）
- GPT-5.1 Pro: 1 回の試行で 13.50 ドル、成功率 90%と仮定 → 期待コスト: 15.00 ドル（1.11 回の平均試行）
- **結果: Gemini 3 Pro が 87%安い**

シナリオ C：ミッションクリティカルなエージェントタスク（高い信頼性が必要）

- Gemini 3 Pro: 1 回の試行で 2.90 ドル、成功率 60%と仮定 → 期待コスト: 4.83 ドル（1.67 回の平均試行）
- GPT-5.1 Pro: 1 回の試行で 13.50 ドル、成功率 95%と仮定 → 期待コスト: 14.21 ドル（1.05 回の平均試行）
- **結果: Gemini 3 Pro が 66%安い**

これらのシミュレーションは、GPT-5.1 Pro の安定性の優位性を考慮しても、**ほとんどのシナリオで Gemini 3 Pro の方がトータルコストが低い**ことを示しています。ただし、ミッションクリティカルで再試行コストが高い環境（例：本番環境への直接デプロイ、顧客向けの自動化されたワークフロー）では、GPT-5.1 Pro の高い安定性が、追加コストに見合う価値を提供する可能性があります。

4.3 コンテキストウィンドウの価値

Gemini 3 Pro の 1M トークンコンテキストウィンドウは、特定のユースケースで決定的な利点となります。[3][4][8][31]

価値が高いケース：

- 巨大なコードベース全体を単一のプロンプトで分析
- 数百ページの法的文書、技術仕様書、研究論文の包括的レビュー
- 数時間分の動画トランスクリプトと映像の統合分析
- 複数の大規模文書を横断した比較分析

これらのタスクでは、GPT-5.1 Pro の 128k トークン制限では対応できず、複数回に分割する必要があります。分割すると、文脈の連続性が失われ、全体的な理解が低下する可能性があります。[71][43][3]

一方、**オーバースペックとなるケース：**

- 標準的な対話・Q&A（数千〜数万トークン）
- 単一ファイルのコードレビュー・リファクタリング
- 短い文書の要約・翻訳

これらのタスクでは、両モデルとも十分なコンテキストを持ち、Gemini 3 Pro の大容量は活用されません。

コンテキスト品質の懸念も重要です。研究では、コンテキスト長が増加すると、モデルの性能が不均一になることが示されています。10,000 番目のトークンは 100 番目のトークンほど確実には処理されず、「コンテキストロット（context rot）」と呼ばれる現象が発生します。ただし、Gemini 2.5 はこの点で改善を示しており、300k+トークンでも良好に機能するという報告があります。Gemini 3 Pro でも同様の改善が期待されますが、超長文コンテキストを使用する場合は、性能の検証が推奨されます。[71]

5. 戦略的推奨と選択ガイド

5.1 ワークロード別の最適モデル選択

研究・分析・文書生成ワークロード → Gemini 3 Pro

- 料金対効果: 78〜90%のコスト削減
- 性能: 高難度推論、マルチモーダル理解で優位
- リスク: 幻覚率が高いため、事実確認が必要
- 推奨: 学術研究、市場分析、コンテンツ生成、データ分析レポート

大規模開発・エージェントワークロード → GPT-5.1 Pro（特に Codex-Max）

- 料金対効果: トークン単価は高いが、再試行コストを考慮すると妥当
- 性能: 長時間安定稼働、エラー耐性、実務的バグ修正で優位
- リスク: 高コストが継続的に発生

- 推奨: プロジェクト規模のリファクタリング、24 時間連続エージェント、本番環境の CI/CD 統合

マルチモーダル・長文コンテキストワークロード → Gemini 3 Pro

- 料金対効果: 大幅なコスト削減＋機能的な優位性
- 性能: 1M コンテキスト、動画・画像理解、UI 分析で圧倒的
- リスク: 超長文コンテキストでのコンテキストロット
- 推奨: 動画分析、巨大文書レビュー、スクリーンショットからのコード生成、コードベース全体の理解

ビジネスクリティカル・信頼性重視ワークロード → GPT-5.1 Pro

- 料金対効果: コストは高いが、エラーコストがそれを上回る
- 性能: 低い幻覚率（相対的）、慎重なアプローチ、再現性
- リスク: 高いトークンコスト
- 推奨: 財務レポート、法的文書、顧客向け自動化、コンプライアンス関連タスク

5.2 ハイブリッド戦略の提案

多くの組織にとって、**両モデルを併用するハイブリッド戦略**が最適です。[43]

ステージ 1：計画と調査 → Gemini 3 Pro

- 広範な情報収集、複数文書の比較分析、初期アイデアの探索
- コスト効率が高く、広範な知識と強力な推論能力を活用
- 例：プロジェクトキックオフ、要件定義、技術調査、アーキテクチャ設計

ステージ 2：実装と検証 → GPT-5.1 Pro（または両モデル並行）

- 実際のコード実装、リポジトリへの統合、テスト実行
- 安定性とエラー耐性を重視
- 例：コード生成、バグ修正、リファクタリング、CI/CD 統合

ステージ 3：レビューと分析 → Gemini 3 Pro

- 実装結果のレビュー、パフォーマンス分析、改善提案
- コスト効率とマルチモーダル能力を活用
- 例：コードレビュー、テスト結果分析、パフォーマンスプロファイリング

標準化とモニタリングが重要です：[43]

- 共通のプロンプトスキーマを使用し、両プロバイダーで一貫性を保つ
- レイテンシ、品質スコア、トークンコストをリクエストごとにログ記録
- 実際のワークロードトレースでモデルを比較し、最適化
- コスト敏感またはデータ敏感なフローでは、AceCloud GPU などでオープンモデルを実行し、ベンダーレバレッジを維持[43]

5.3 将来の展望と注意事項

Gemini 3 Pro はリリース直後（2025 年 11 月 18 日）であり、まだ 1 週間も経っていません。長期的な安定性、エッジケースでの挙動、API レート制限、詳細な価格表など、不明な点が多く残っています。実際のユーザーレポートでは、Antigravity が「最初は正常に動作したが、その後『モデルプロバイダーの過負荷によりエージェント実行が終了しました。後でもう一度試してください』というエラーメッセージが表示された」というケースもあります。[9][21]

GPT-5.1 Pro は段階的展開中で、すべての Pro ユーザーがまだアクセスできるわけではありません。OpenAI は、パフォーマンスを安定させるために数日かけて段階的にロールアウトしています。また、GPT-5 Pro は今後 90 日間利用可能ですが、その後は廃止される予定です。[1][72][16]

料金の変動可能性も考慮すべきです。Google は通常、プレビューモデルが安定し効率が向上すると、20〜50%料金を削減します。Gemini 3 Pro の安定版料金は 2026 年初頭に、1.50 ドル/10 ドル（20 万以下）と 3 ドル/15 ドル（20 万超）程度に落ち着き、同時にキャッシングおよびバッチ割引が導入されると予想されています。[5]

モデルサイズとスケーリングの観点では、Gemini 3 Pro の AA-Omniscience Benchmark での高い正答率が、モデルサイズの大きさを示唆しています。正答率はモデルサイズ（総パラメータ数）と強く相関することが示されており、Gemini 3 Pro が競合モデルよりもはるかに大きいモデルである可能性が指摘されています。一方、OpenAI は GPT-5 のパラメータ数を公開しておらず、推定値は 1.7〜1.8 兆パラメータ（密なモデル）から数十兆パラメータ（MoE モデル）まで幅があります。[58][25][26]

AI エージェントの未来に向けて、両社とも大きな投資を行っています。Google の Antigravity は、エージェントが複数の表面（エディタ、ターミナル、ブラウザ）を横断して長時間中断なく動作する世界を目指しています。OpenAI の Codex-Max は、24 時間以上の連続稼働を実現し、「朝にタスクを定義して、夜に完成したものを確認する」という非同期開発を可能にします。どちらのビジョンも、AI がより自律的で信頼できるコラボレーターとなる未来を描いています。[12][13][21][22]

結論

GPT-5.1 Pro と Gemini 3 Pro は、どちらも 2025 年 11 月の AI 業界における最先端のモデルですが、その強みと適性は明確に異なります。

料金構造では、Gemini 3 Pro が圧倒的に優位であり、ほとんどのワークロードで 78〜90%のコスト削減を実現します。**ベンチマーク性能**では、Gemini 3 Pro が数学・抽象推論、マルチモーダル理解、視覚的推論、長文コンテキスト処理において優位性を示し、LMArena で史上初の 1500 超えを達成しました。**実務的な開発ワークフロー**では、GPT-5.1 Pro とその Codex-Max 拡張が、長時間稼働の安定性、エラー耐性、実世界のバグ修正において依然として強みを持ちます。

「どちらが安いか」「どちらが優れているか」という問いには、単一の答えはありません。推論・分析・文書生成・マルチモーダルタスクでは、Gemini 3 Pro の方が費用対効果が高く、性能も優れています。大規模開発・24 時間エージェント・ミッションクリティカルな自動化では、GPT-5.1 Pro の安定性が、追加コストに見合う価値を提供します。

最適な戦略は、**目的を明確にし、ワークロードに応じてモデルを使い分けること**です。多くの組織にとって、両モデルを併用するハイブリッドアプローチが、コスト、性能、信頼性のバランスを最適化する鍵となるでしょう。

⁂

1. https://tenbin.ai/media/chatgpt/AI_NEWS_251120
2. https://www.glbgpt.com/hub/gemini-3-pro-costs-gemini-3-api-costs-latest-insights-for-2025/
3. https://www.cometapi.com/gemini-3-pro-vs-gpt-5-1-which-is-better-a-complete-comparison/
4. https://ai.google.dev/gemini-api/docs/gemini-3

5. https://apidog.com/jp/blog/gemini-3-0-api-cost/
6. https://www.reddit.com/r/codex/comments/1p36j5h/real_world_comparison_gpt51_high_vs_gpt51codexmax/
7. https://www.getpassionfruit.com/blog/chatgpt-5-vs-gpt-5-pro-vs-gpt-4o-vs-o3-performance-benchmark-comparison-recommendation-of-openai-s-2025-models
8. https://www.vellum.ai/blog/google-gemini-3-benchmarks
9. https://zenn.dev/tenormusica/articles/gemini-3-pro-preview-release
10. https://x.com/EpochAIResearch/status/1991945942174761050
11. https://binaryverseai.com/gemini-3-vs-gpt-5-1-coding-benchmarks-comparison/
12. https://www.remio.ai/post/gpt-5-1-codex-max-analysis-a-new-standard-for-agentic-coding
13. https://www.cometapi.com/what-are-gpt-5-1-codex-max-and-how-to-use-it/
14. https://aiupdate.blog/openai-gpt-5-1-pro-quiet-release-ai-competition-112225/
15. https://note.com/yasuhitoo/n/n49d3535f1e86
16. https://openai.com/index/gpt-5-1/
17. https://www.gizmodo.jp/2025/11/google_gemini_3_released.html
18. https://gihyo.jp/article/2025/11/gemini3-pro-preview
19. https://blog.google/products/gemini/gemini-3/
20. https://www.mouse-jp.co.jp/mouselabo/entry/2025/11/19/100279
21. https://gigazine.net/gsc_news/en/20251119-google-antigravity/
22. https://developers.googleblog.com/en/build-with-google-antigravity-our-new-agentic-development-platform/
23. https://www.itmedia.co.jp/aiplus/articles/2511/19/news067.html
24. https://forest.watch.impress.co.jp/docs/news/2062843.html
25. https://www.cometapi.com/how-many-parameters-does-gpt-5-have/
26. https://lifearchitect.ai/gpt-5/
27. https://www.datacamp.com/blog/gpt-5

28. https://www.cometapi.com/ja/gemini-3-pro-vs-gpt-5-1-which-is-better-a-complete-comparison/

29. https://encord.com/blog/gpt-5-a-technical-breakdown/

30. https://platform.openai.com/docs/models/gpt-5-pro

31. https://www.marktechpost.com/2025/11/18/googles-gemini-3-pro-turns-sparse-moe-and-1m-token-context-into-a-practical-engine-for-multimodal-agentic-workloads/

32. https://www.youtube.com/watch?v=Z9IQQb7em98

33. https://eu.36kr.com/en/p/3559185639652488

34. https://huggingface.co/datasets/multimodalart/google-gemini-3-pro-pre-release-model-card

35. https://skywork.ai/skypage/en/gemini-3-pro-deep-dive-analysis/1990964279088766976

36. https://openai.com/api/pricing/

37. https://www.glbgpt.com/hub/jp/how-much-does-gpt5-1-cost/

38. https://platform.openai.com/docs/pricing

39. https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/

40. https://note.com/ai_worker/n/n461dd70ab2b2

41. https://ai.google.dev/gemini-api/docs/pricing

42. https://www.eesel.ai/blog/google-gemini-3-pricing

43. https://acecloud.ai/blog/gemini-3-vs-chatgpt-5-1/

44. https://media.securememo-cloud.com/articles/gemini-3s-capabilities-and-use-cases-feature-comparison-performance-pricing-and-use-cases/

45. https://beebom.com/google-unleashes-gemini-3-pro-the-new-benchmark-for-ai-intelligence/

46. https://bdtechtalks.com/2025/11/18/google-gemini-3-0-pro/

47. https://www.cometapi.com/is-gemini-3-pro-about-to-crush-the-ai-competition/

48. https://www.facebook.com/xixidu/posts/gemini-3-pro-set-a-new-record-on-frontiermath-38-on-tiers-13-and-19-on-tier-4on-/10173130754245637/

49. https://www.reddit.com/r/aicuriosity/comments/1p3pzph/google_gemini_3_pro_becomes_top_performing_public/

50. https://www.facebook.com/xixidu/photos/gemini-3-pro-set-a-new-record-on-frontiermath-38-on-tiers-13-and-19-on-tier-4on-/10173130754130637/

51. https://vertu.com/lifestyle/gemini-3-vs-gpt-5-vs-claude-4-5-vs-grok-4-1-the-ultimate-reasoning-performance-battle/

52. https://deepmind.google/models/gemini/pro/

53. https://www.datacamp.com/blog/gemini-3

54. https://officechai.com/ai/googles-gemini-3-tops-frontiermath-benchmark-that-tests-ai-models-on-expert-level-math/

55. https://openai.com/index/introducing-gpt-5/

56. https://codezine.jp/news/detail/22617

57. https://the-decoder.com/gemini-3-pro-tops-new-ai-reliability-benchmark-but-hallucination-rates-remain-high/

58. https://artificialanalysis.ai/articles/gemini-3-pro-everything-you-need-to-know

59. https://metana.io/blog/gemini-3-vs-gemini-2-5/

60. https://www.reddit.com/r/singularity/comments/1p0jc6e/gemini_3_pro_hallucination_rate_vs_gemini_25_pro/

61. https://staffing.archetyp.jp/magazine/gpt-5-1-codex-max/

62. https://forest.watch.impress.co.jp/docs/news/2065234.html

63. https://weel.co.jp/media/tech/openai-gpt-5-1-codex-max/

64. https://qiita.com/yokko_mystery/items/bb5615ebcd385a597c41

65. https://portkey.ai/blog/gemini-3-0-vs-gpt-5-1/

66. https://skywork.ai/blog/ai-agent/gpt5-1-vs-gemini-3/

67. https://note.com/it_navi/n/n9d307bfc6bc6

68. https://skywork.ai/blog/gemini-3-vs-gpt-5/

69. https://chatgpt-enterprise.jp/blog/gpt-5-1-gemini-3-hikaku/

70. https://recap.aitools.inc/p/openai-s-new-gpt-5-1-kills-the-context-window-limit

71. https://www.reddit.com/r/OpenAI/comments/1mj78xy/just_a_reminder_that_the_context_window_in/

72. https://www.tomsguide.com/ai/chatgpt/openai-just-launched-chatgpt-5-1-pro-to-fight-gemini-3-heres-the-biggest-upgrades

73. https://applyingai.com/2025/11/gpt-5-1-unveiled-openais-warmth-and-personality-make-ai-more-human/

74. https://longbridge.com/en/news/266453995

75. https://zenn.dev/beagle/scraps/a29799905892f3

76. https://www.thealgorithmicbridge.com/p/google-gemini-3-just-killed-every

77. https://note.com/life_to_ai/n/ne2f4545d2db4

78. https://note.com/r1250_gs/n/nee9c7c427f7f

79. https://www.linkedin.com/posts/epochai_gemini-3-pro-set-a-new-record-on-frontiermath-activity-7397711633818468352-s7T6

80. https://www.reddit.com/r/OpenAI/comments/1p09hzj/gemini_30_pro_vs_gpt_51_benchmark/

81. https://inkeep.com/blog/openai-gpt-5-personalized-ai-architecture

82. https://www.datastudios.org/post/gemini-antigravity-vs-chatgpt-5-1-codex-which-ai-coding-system-is-better-for-tools-pricing-and-pe

83. https://www.youtube.com/watch?v=oPUjTRvNunM

84. https://nobdata.co.jp/report/chatgpt/28/

85. https://www.linkedin.com/posts/ali-abuharb-3b4905153_gemini-3-pro-model-card-activity-7396530756489146369-QI8P

86. https://www.reddit.com/r/codex/comments/1p0iq08/gemini_3_vs_codex_551_which_model_is_better_post/

87. https://learn.microsoft.com/ja-jp/azure/ai-foundry/foundry-models/concepts/models-sold-directly-by-azure?view=foundry-classic

88. https://jobirun.com/gemini-3-pro-model-card/

89. G6UZusDbkAApvZl.jpg

90. https://www.knowleful.ai/plus/gemini3-google-ai/

91. https://www.helicone.ai/llm-cost/provider/openai/model/gpt-5

92. https://eu.36kr.com/en/p/3560848667671682

93. https://gemini.google/release-notes/

94. https://www.ai-souken.com/article/chatgpt-api-pricing-guide

95. https://ai.google.dev/gemini-api/docs/changelog

96. https://www.reddit.com/r/ArtificialInteligence/comments/1p0c3vc/gemini_30_pro_vs_gpt_51_llm_benchmark_showdown/

97. https://docs.cloud.google.com/vertex-ai/generative-ai/pricing

98. https://learn.microsoft.com/ja-jp/azure/ai-foundry/openai/how-to/reasoning

99. https://www.glbgpt.com/jp/hub/how-good-is-gemini-3/

100. https://binaryverseai.com/gpt-5-1-openai-update-instant-thinking-review/

101. https://www.reddit.com/r/ChatGPTPro/comments/1p0zisl/gemini_3_is_what_gpt_5_should_have_been_its/

102. https://x.com/nicdunz/status/1989049239637512482

103. https://note.com/lab_bit_sutoh/n/n426e7bfaf23a

104. https://www.reddit.com/r/singularity/comments/1ow9xcj/gpt_51_benchmarks/

105. https://x.com/KoichiNishizuka/status/1992026834721235311

106. https://www.reddit.com/r/GeminiAI/comments/1p1fycs/gemini_3_pro_benchmarks_key_observations/

107. https://openai.com/index/gpt-5-1-codex-max/

108. https://japan.zdnet.com/article/35240683/

109. https://chatgpt-lab.com/n/ne35c561efac3

110. https://dev.to/alifar/gpt-5-a-game-changer-for-developers-teams-ai-agents-4359

111. https://www.glbgpt.com/jp/hub/what-is-gpt5-1/

112. https://macaron.im/blog/chatgpt-vs-gemini-vs-claude

113. https://vertu.com/lifestyle/gemini-3-0-vs-gemini-2-5-pro-google-sets-new-performance-standards-in-2025/

114. https://openai.com/index/gpt-5-1-for-developers/

115. https://research.aimultiple.com/ai-hallucination/