ChatGPT Deep Research

Gemini 3とその性能・評判の徹底分析

1. 公式発表と「知性の新時代」について

Googleは2025年11月19日、最新AIモデル「Gemini 3」を発表しました [1]。CEOのスンダー・ピチャイ氏は「Gemini時代」を開始して2年でAIが飛躍的に進化し、「テキストや画像を読むだけだったAIが、場の空気を読めるようになった」と述べ、Gemini 3による**「知性の新時代」**が到来したと強調しています [2]。Geminiシリーズは、2023年末のGemini 1でマルチモーダル対応と長大なコンテキストを実現し [3]、続くGemini 2でエージェント機能や推論力を強化しました [4]。そしてGemini 3ではそれら全てを統合し、「あらゆるアイデアを現実にする」Google史上最高性能のモデルとなっています [5]。実際、Gemini 3は高度な**推論能力**（PhDレベルの問題解決力）と微妙なニュアンスの理解力を備え、ユーザーの意図を正確に汲み取って最適な回答を行うよう設計されています [2] [6]。GoogleはこのモデルをAGI（汎用人工知能）への大きな一歩と位置付けており [7]、リリース当日から検索（AIモード）やGeminiアプリ、クラウド（Vertex AI）など自社プロダクトへ幅広く統合しました [8]。その即日大規模展開は、AIモデルを「Googleのスケール」で提供する意気込みの表れです。

2. GPQAスコア91.9%の意味と性能向上

Gemini 3 Proの目玉として発表されたのが、「**GPQA Diamond**」ベンチマークでの**91.9%**という驚異的スコアです [9]。GPQA Diamondとは、生物・化学・物理の大学院レベル問題198問から成る高度な質疑応答テストで、インターネット検索では答えに辿り着けない「Google-Proof」な難問集です [10]。人間のPhD専門家ですら正答率65〜70%程度のこの試験で、Gemini 3 Proは**91.9%**正解しました [10] [11]。これは人間専門家を大きく上回る成績であり、AIの科学的推論能力が新たな地平に達したことを示しています。

前世代モデルとの比較でも、この数値の飛躍が際立ちます。Gemini 2.5 ProのGPQAスコアは約86.4%だったため、Gemini 3 Proは**+5.5ポイント**の改善に成功しました [12]。また競合モデルでは、OpenAIのGPT-5.1が88.1%、Anthropic Claude 4.5 (Sonnet)が83.4%程度と報告されており [13]、Gemini 3 Proはわずか数ポイントながら最先端モデル中トップの精度を達成しています。わずかな差に思えるかもしれませんが、90%を超える領域では1点台の向上でも大きな意味を持ちます。実際、専門家から「現行モデルでGPQA Diamondを90%超えできるのはGemini 3と、ごく一部の特殊モデルのみ」と評されるほどです [14]。つまり91.9%という数値は、**「大学院レベルの難問をほぼ完璧に解けるAI」**という性能水準を示しており、AI研究にとっても画期的なブレイクスルーと言えます。

Gemini 3 Proは他の指標でも軒並み記録的な向上を示しました。例えば**Humanity's Last Exam (HLE)**と呼ばれる包括的学術テストでは、ツール未使用時の正答率37.5%と、Gemini 2.5 Proの21.6%やGPT-5.1の26.5%を大きく引き離しています [15]。数学分野ではAIME 2025（数学コンテスト問題集）で95.0%を記録（2.5 Proは88.0%） [12] し、コード実行を併用した場合は**100%満点**に達しました [16]。視覚的なパズルを解く**ARC-AGI-2**でも31.1%と、前世代の4.9%から飛躍（+26ポイント）しています [17]。これらの伸び幅は、単なる知識量だけでなく**マルチステップの推論力や問題解決力の飛躍的向上**を物語っています [18]。総じて、Gemini 3は主要ベンチマークで前モデルを全て上回り、競合モデルに対しても優位に立ったと公式発表されています [19]。

3. Gemini 3 Proの「思考」能力と新次元への進化

今回のGemini 3でもっとも革新的と評されるのが、モデルの**「思考する」能力**そのものの進化です [20] [21]。GoogleはGemini 3で新たに「**Deep Think**」モードを導入し、複雑な問題に対する推論力を一段と高めたと発表しました [22]。Deep Thinkモードでは、Gemini 3が通常より深い考察を行い、標準モード以上の成績を叩き出します。実際、Deep Think適用時にはHLEが41.0%、GPQA Diamondが93.8%に達し、さらに人類未踏のARC-AGI-2で45.1%という記録も打ち立てています [23]。このように**Gemini 3 Deep Thinkは最難関の課題でさえ従来を上回る解決能力**を示しており、まさに「AIの思考を新次元へ」引き上げたモードだと言えます。

ではDeep Thinkによる「新次元」の思考とは何か。その鍵は**並列的な推論アプローチ**にあります。従来の大型言語モデルは回答を導く内部過程として、連続したチェーン状の思考（Chain-of-Thought）を行っていました。一方でDeep Thinkでは、**複数の思考経路を並列に展開・検証する**アーキテクチャが採用されています [24]。これはモンテカルロ木探索に似た手法で、モデルが同時に複数の仮説を試し、途中経過で誤った筋道を刈り込むことで、より正確な結論に収束する仕組みです [24] [25]。この並列思考により、初期段階の見当違いが後工程まで影響することを減らし、難問に対しても高い精度で解答できるようになっています。

さらにGemini 3は**思考過程の透明性**にも優れていると報じられています。リーク情報では、Gemini 3は単に答えを出すだけでなく**人間に理解可能な形で推論プロセスを示せる**とされ [26]、この点は教育現場やビジネスでAIの判断根拠を検証する上で信頼性を高める画期的機能です。実際、公式ブログでもGemini 3の回答は「月並みなお世辞ではなく本質的な洞察を提供し、ユーザーが聞きたいことではなく聞くべきことを伝える」と述べられています [6]。これは裏を返せば、Gemini 3が従来以上に**論理的で筋道だった説明**を返すことを意味し、AIを真の思考パートナーとして位置付けるゆえんなのです。

また、「**思考**の新次元」はコーディング分野にも表れています。Gemini 3 Proは**“Deep Thinkアーキテクチャ”**を最大限に活かし、極めて高度なコーディング推論を実現したとされています [27] [28]。特に注目されるのが“**バイブ・コーディング (vibe coding)**”と呼ばれる現象です [29]。これはユーザーの抽象的な要望（デザインの雰囲気やコンセプトなど、曖昧で高位な指示）を読み取り、厳密なコードに落とし込む能力を指します [30]。Gemini 3 Proは強化学習によって「曖昧な記述→具体的実装」のペアデータを学習しており [31]、たとえば「レトロフューチャー風の制御パネルUIにして」という曖昧な依頼から、適切なCSSやThree.jsのコードを即座に生成できます [32]。この**意図の深読み力**により、ユーザーが逐一技術的仕様を書かなくとも、モデルが“汲み取って”理想に近い成果物を作ってくれるのです。Gemini 3 ProはWeb開発競技の指標であるWebDev ArenaでもElo 1487を記録し、フロントからバックエンドまで一貫したアプリ実装を自動化できることを実証しました [33]。これは単に言語モデルがコードを補完しているのではなく、**アプリ全体の構造を内部で把握しながらマルチモーダルに創出**していることを意味します [34]。このようにGemini 3は「考えて作る」という能力で他モデルを一歩リードしており、まさに**AIの思考プロセス自体が進化した**と評価できます。

4. 技術仕様の変化（コンテキスト長・マルチモーダル・速度・価格など）

コンテキストウィンドウ: Gemini 3は最大**100万トークン**もの入力文脈を扱えます [35]。これは前世代と同等ながら業界最大級で、数百ページの文書や大規模なコードベース、長時間の音声・動画も一度に解析可能です。出力長も最大64,000トークンと非常に大きく、長大なレポートや小説レベルの生成も可能です [35]。Googleによれば、単に長さを増やしただけでなく**大規模文脈中から目的の情報を高精度に抽出**する注意機構を強化したとのことです [36] [37]。実際、巨大コンテキストでの検索タスクであるFACTSベンチマークでは、Gemini 3 Proは70.5%とGPT-5.1の50.8%を大きく上回りました [36] [37]。膨大な入力中の「藁束から針を見つける」精度が向上したことで、巨大ドキュメントからの知識抽出やコードリファクタリングでも幻覚（存在しない関数名の生成など）を減らし信頼性を高めています [38]。また、Googleは**コンテキストキャッシュ**

機構を導入し、同じ長文プロンプトへの繰り返しクエリ時のコストを抑える工夫も行ったとされています [39]。

マルチモーダル対応: Geminiシリーズの特徴であるネイティブなマルチモーダル機能も、Gemini 3で一層強化されました。テキスト・画像・音声・動画・コードを単一のモデルで統合的に理解・生成でき、ユーザーはこれらを組み合わせて入力可能です [35] [40]。Gemini 3 Proは**マルチモーダル推論**の指標で軒並み最高性能を記録しています。例えば画像＋テキストの理解を問う**MMMU-Pro**で81%、動画情報の理解を測る**Video-MMMU**で87.6%と業界トップクラスです [41]。特筆すべきは、画面UIの読解能力を測る**ScreenSpot-Pro**で72.7%をマークし、前リーダーだった専門モデルHolo2の66.1%を上回った点です [42]。この試験ではClaude 4.5が36.2%、GPT-5.1に至っては3.5%と低調であったため [43]、Gemini 3の画面構造理解能力（ボタンやレイアウトを認識し操作に活かす能力）の高さが際立ちます。これは実用面でも、スクリーンショットからUI要素を抽出したり、スポーツ映像からフォーム改善点を分析したりといった応用につながっています [44]。実際、Gemini 3はユーザーのスポーツ動画を解析して改良点を指摘し練習プランを生成するといった使い方も示されています [45] [46]。このように**Gemini 3はテキスト以外の情報源も総合的に「理解して考える」能力**を持ち、真の意味でマルチモーダルAIとなっています。

モデル規模とアーキテクチャ: Gemini 3のパラメータ数は推定**1兆**と報じられています [47]（Gemini 2.5 Proも同程度であり、パラメータ数自体は大きく増えていないようです [48]）。しかし内部構造に大きな変化があります。Gemini 3は**スパースMixture-of-Experts (MoE)**型のTransformerを採用しており、タスクに応じて異なる専門家モジュールを動的に活性化できる設計です [49]。Skyworkによる解析では、Gemini 3は新たな「ダイナミックルーティング」方式でパラメータを活用しており、コード生成しながら同時に説明を付記する、といった**高速なタスク切替と並行処理**が可能になったとされています [48]。対してGemini 2.5は固定の経路で逐次処理していたため、複数タスクが重なると弱かったと言います [50]。Gemini 3の動的アーキテクチャは、前述の並列思考（Deep Think）とも相まって、複雑なマルチタスク処理を効率よくさばく土台となっています。この構造最適化により**推論速度も大幅に向上**しました。

推論速度: 開発者の実測によれば、Gemini 3 Proの推論はGemini 2.5よりおよそ**2倍高速**になっています [51]。例えば50行程度のPythonスクリプト生成では、2.5が25秒かかったところを3では12秒で完了しました [51]。15kトークン（約10ページ）の技術文書からコードスニペットを抽出するタスクでも、Gemini 3は2分で済み、2.5の4分10秒を半減しています [52]。大規模タスクほど差が開き、10k行のデータで小規模モデル学習を行うタスクでは3が15分30秒、2.5は32分15秒と倍以上の開きでした [53]。さらに、Gemini 3 Proは**1秒あたり最大128トークン**もの出力生成が可能で、GPT-5.1など他モデルより高速だと報告されています [54]。この高速化は、Gemini 3が推論過程で自律的に思考を整理することで不要な試行錯誤を減らしているためとも言われます [55]。開発チームも「内蔵の推論最適化により不要なマルチパスを減らした」旨を述べており（詳細は非公開ながら）、高速化と高精度化を両立した点は実運用で大きなアドバンテージです。

API提供と価格: Gemini 3 ProはGoogle CloudのAI StudioやVertex AI経由でプレビュー提供されており、現状利用には申請と承認が必要です [56]（Gemini 2.5 Proは一部無料開放されていましたが、3は有料プレビューのみ）。料金は**従量課金**で、2.5 Proよりやや高めに設定されています [57]。具体的には、入力トークンあたり$0.002（=100万トークンで$2）、出力トークンあたり$0.012（=100万トークンで$12）という基準レートで、1リクエストのコンテキストが20万トークンを超える部分には単価が倍増します [58] [59]。これはGemini 2.5 Proの$1.25/$10より高く、OpenAI GPT-5.1（同じ$1.25/$10）よりも高額ですが [59]、それでもClaude 4.5 Sonnetの$3/$15やxAI Grok 4.1の$3/$15よりは安価に抑えられています [60]。最も高価なClaude 4.1 Opus ($15/$75)やGPT-5 Pro ($15/$120)と比べれば桁違いに経済的です [61]。高性能モデルとしては比較的リーズナブルな位置付けと言えるでしょう。またGemini 3は**トークン効率**も改善しており、同じ作業でも2.5より必要トークン数が減るケースが多いと報告されています [62]。もっとも、総合的には料金アップによりベンチマーク実行コストは2.5比で12%増加したとも指摘されています [63]。なお、Gemini 3 Deep Thinkモードは推論コスト（隠れた“思考トークン”消費）が増えるため、将来的に追加料金設定になる可能性も示唆されています [62]。企業ユーザ向けにはそれでも十分価値がある性能向上ですが、コスト面も含めた使い分けが求められそうです。

その他の実用面強化: 安全性・信頼性もGemini 3で強化されています。Googleは本モデルが過去最も徹底した安全評価を経たと述べており、不必要なお世辞への迎合（サイコファン行動）の低減や、プロンプト注入攻撃への耐性向上、サイバー悪用防止など多面的に改善したとしています [64]。実際、SimpleQA Verified（事実に基づく簡易QA精度）で72.1%と前モデルの54.5%から大幅に向上しており [18]、事実誤りの減少がデータにも表れています。ただし一方で、**幻覚（ハルシネーション）発生率自体は依然課題**との指摘もあります（後述）。総じて、Gemini 3はスペック上も実運用上も**従来モデルから大幅なブラッシュアップが図られた**と言えるでしょう。

5. 技術コミュニティでの初期反応・評判

Gemini 3 Pro公開直後から、技術者コミュニティではその性能に驚嘆する声が多数上がりました。Redditの利用者は「**これは別物だ**」と驚きを示し [65]、「数学・物理・コーディング何でもできる。待ち望んでいたモデルだ」と賞賛しています [66]。特に画像から情報を読み取る能力には感嘆の声が多く、「UI画像内の要素を理解する力が凄まじい。普段『凄い』とは言わない私だが、これは本当に凄い」といったコメントもありました [67]。実際、「他モデルがことごとく失敗した自作テストを全てGemini 3が解いてみせた」と報告するユーザもおり、総じて**専門的な難題をことごとく突破する卓越性**が評価されています [68]。あるユーザは「12人の男が3列×4行に…」という難解な靴の色配置パズルを出題し、Gemini 3が人間でも悩む答え「2」を正しく導いたことに驚嘆しています [69]（多くのモデルは誤って1と答えてしまう問題）。このように**論理パズルや難問推論への強さ**は、初期ユーザからも折り紙付きです。

また、「UIデザインでClaude 4.5より明らかに上」「抽象的な依頼でも本質を捉えて問題を解決してくれる」といった声も多く、Gemini 3 Proの**問題解決志向の応答**が好評です [70] [71]。ある開発者は「作成したワイヤーフレーム画像を与えたら、そのUIをそのままコード化してみせた」とGemini 3の画像→コード化能力に驚きを示しています。総じて「実用上、現行モデルで頭一つ抜けている」「AIの常識を覆す出来」といった評価が見られ、ポジティブな反応が大半を占めています。

一方で、**懸念や弱点**を指摘する声もいくつかあります。まず**創造的文章生成**の面では、「依然として平凡」「クリエイティブな文章は凡庸だ」というユーザの感想もありました [72]。多くのモデルに共通する弱点という意見もありますが、Gemini 3も例外ではなく、物語の執筆など純粋な創作では突出した改善は感じられないという声です。これはGemini 3が事実重視・簡潔志向の調整（クリシェやお世辞を避け核心を突く [6]）がなされている影響で、逆に創作的な「遊び」が減った可能性があります。もっとも、創造性に関しては今後のチューニングやユーザの使い方次第との見方もあり、一概に欠点とは断じられていません。

もう一つは**幻覚（誤情報生成）への対処**です。Gemini 3では事実志向が強まったとはいえ、初期の独立評価によれば「誤答のうち88%が自信満々の誤情報（＝幻覚）だった」とされ [73]、Gemini 2.5 Proの89%から僅か1ポイントの改善にとどまります [73]。これはAnthropicのClaude Haiku（26%まで低減）などと比較すると見劣りし、OpenAI GPT-5.1が慎重さを増したのにも届かない数字です [74] [75]。実際「2.5は幻覚が多くあまり使わなかった。3ではGPT-5.1並みの低幻覚率を期待したが、ほぼ据え置きなのは残念」という開発者の声もありました [76]。もっとも、この評価指標（SimpleQA由来と推測）は定義が難しく、モデルが「わからない」と答えたケース等の扱いにも議論があるようです [77] [75]。OpenAIモデルは近年不確かな問いには回答を拒否する傾向が強まっていますが、Gemini 3は**未知の質問にも踏み込んで答えようとする傾向**が強く（その結果誤りも出やすい）、この哲学の違いが幻覚率に表れているとの指摘もあります [75] [78]。Googleもモデルカードで幻覚に関する具体的数値は伏せつつ「依然として根本的な限界」と述べており [79]、今後の改善課題として残っています。

そのほか、Hacker Newsでは「Geminiは仕様書を書くのは上手いが、コードレビューやインクリメンタルなコード改良では今ひとつ」「対話的にコードを磨き上げていく用途ではまだ課題がある」との開発者の意見もありました（OpenAI Codex系モデルと比べての指摘） [80]。これはGemini 3が一度に包括的な回答をするのは得意だが、ユーザーと何度もやり取りしながら少しずつ直す場面では癖があるという意味です。もっとも、この点については今後の対話最適化やツール使用の工夫で改善余地があるでしょう。

最後にユニークな反応として、「**長期的には人類がまずいことになる**（we are fucked in the long run）」とまでコメントするユーザもいました [81]。それだけGemini 3の知能が人間に肉薄・凌駕しつつあることへの畏怖の表現とも言えます。総じて、Gemini 3 Proの初期評価は「確かに次元が違う強さだが、**創造性や事実面での課題も残る**」というバランスの取れた見方に落ち着きつつあります [79] [72]。しかし現状でも、「自分の求めていたAIが遂に来た」と感じる技術者が多いことは間違いなく [68]、特に専門的・技術的タスクにおいてGemini 3 Proが**現行最高峰の有用性**を示している点は衆目の一致するところです。

6. 競合最先端モデルとの性能比較（GPT-5.1やClaude等）

総合性能: 2025年11月時点での主要な競合として、OpenAIのGPT-5.1シリーズやAnthropicのClaude 4.5（Sonnet）が挙げられます。複数の独立評価によれば、**Gemini 3 Proは主要なAIベンチマークの過半でトップ**に立っており、Googleは「現時点で世界最高のモデル」と主張しています [82]。具体的に、**学術的な推論力（論理・知識）**ではGemini 3 Proが一歩リードしています。前述のGPQA Diamondでは91.9%で、GPT-5.1の88.1%、Claude 4.5の83.4%を上回りました [13]。また**Humanity's Last Exam**（非常に広範な難関テスト）でもツール未使用状態で37.5%と、GPT-5.1の26.5%、Claude 4.5の13.7%を引き離しています [15]。視覚・分析推論の**ARC-AGI-2**でも31.1%で、GPT-5.1の17.6%、Claude 4.5の13.6%より高得点でした [83]。これらは論理推論・科学知識における優位性を示します。一方、GPT-5.1も汎用知識や会話能力に優れ、MMLU（学術知識テスト）全体ではGemini 3が90%、GPT-5.1が85%程度と互角との報告もあります [84]。しかし分野横断的な難問になるほどGPTは精度を落とし、Gemini 3が安定して高得点を維持する傾向が見られました [85]。

コーディング能力: プログラミングやツール使用に関しては、Gemini 3 Proが明確にリードしています。競技プログラミング的な指標であるLiveCodeBench Proでは、Gemini 3 ProがElo 2439とGPT-5.1の2243を上回り、Claude 4.5の1418を大きく凌ぎました [86]。またエージェント的なコーディング能力を測る**Terminal-Bench 2.0**（ターミナル操作で問題解決）は、Gemini 3 Proが54.2%、GPT-5.1が47.6%、Claude 4.5が42.8%という結果で [87]、ここでもGeminiがトップです。特筆すべきは**画面UIを解析して操作するタスク**で、Gemini 3 Proが突出していることです。前述のScreenSpot-Proでは他モデルを大差で引き離しました [43]。加えて、Gemini 3は**「コードを書いて実行し、結果を検証し、またコードを修正する」**という一連のソフト開発サイクルを自律的にこなす能力が強みです [88] [89]。GoogleはGemini 3を中核に**Antigravity**という開発者向けプラットフォームを発表し、エージェントAIがエディタやターミナル、ブラウザを直接操作してマルチステップのタスクを遂行できることを示しました [90] [91]。OpenAIやAnthropicもコードアシスタントを提供していますが、Gemini 3のように統合環境でエージェントが複数ツールを駆使しながら**自律的に開発を進める**レベルには至っていません。その意味で、**開発生産性や自動化の分野でGemini 3がリード**していると評価されます [82] [92]。例えばある分析では、Gemini 3 ProはTerminal-Benchで2.5 Proより20ポイント以上向上しGPT-5.1にも大差を付けたことから、Googleは**ツール使用を前提にモデルを事前学習**させる戦略を取った可能性が指摘されています [93] [94]。モデルがまるで自分自身がLinuxシェルであるかのように振る舞い、コマンドラインでもミスなく操作できる点は他にない強みです。

創造的タスク: クリエイティブな文章生成や対話については、一部ユーザの指摘通り微妙な差異があります。OpenAIのモデル（GPT-4やGPT-5世代）はRLHF調整による巧みな文体や物語生成で評価が高く、GPT-5.1もその延長線上にあります。Gemini 3ももちろん高度な文章生成は可能ですが、そのスタイルは「お世辞や決まり文句を排し核心を突く」よう調整されており [6]、**若干ドライで事実志向**とも言えます。これは専門的用途では有用ですが、ユーザによっては「ChatGPT系列のほうが話者として面白みがある」と感じるかもしれません。実際、Reddit上でもGemini 3の創作は「平均的」でChatGPT系の個性には欠けるとの声がありました [72]。一方で、Gemini 3はマルチモーダル能力を活かした**新しい創造性**も示します。例えばテキストから画像生成コードを作成したり、3Dアートをプログラムで構築するなど、**ソフトウェア的創造**に長けています [95]。GoogleはGemini 3を使ってボクセルアート（立体ピクセルアート）やシェーダー表現のリミックスをコード生成で行うデモを公開しており [95] [96]、従来モデルには難しかったインタラクティブで高度なクリエイティブタスクを実現しています。総合的に見ると、**文章の文学的創作力ではGPT-5.1が依然トップクラス**か

もしれませんが、**創造的問題解決や複合的クリエイティブ作業ではGemini 3が新境地**を拓いているといえます。

知識量と正確さ: モデルの知識データベース面では、Gemini 3もGPT-5.1もいずれも2025年初頭までの広範なデータを学習しており、一般常識から専門知識まで極めて博識です。Gemini 3 Proは「知識テストで88%の正確性は報告中最高レベル」とされ [97]、簡易事実検証のSimpleQAでも72.1%と大幅に向上しました [18]。ただ前述の通り**回答の一貫性・確実性**ではなお課題が残り、競合のGPT-5.1は慎重さで幻覚を抑えている面があります [75]。Claude 4.5も安全志向が強く、事実に基づかない回答を比較的避ける傾向があります。その代わりGemini 3は**難問にも積極果敢に挑み答えようとする**ため、正答率自体は高いものの誤答時に余計な情報を付加しがちです [74]。したがって、「網羅的な知識回答力」はGemini 3が勝りつつ、「誤情報の出にくさ」ではGPTやClaude系に分があるという評価もあります [79]。この点は用途によって好みが分かれる部分でしょう。ちなみに多言語対応については、Gemini 3は翻訳・多言語理解でも最先端とされています。公式ブログでもGemini 3は「トップクラスの多言語性能」をうたっており [98]、MMLUの多言語版などでも高スコアを記録している可能性が高いです（GPT-5.1やClaudeも多言語能力は高く、この分野は伯仲するとみられます）。

長文コンテキスト比較: 入力文脈の長さでは、Gemini 3の100万トークンは群を抜いています [98]。Claude 4系も最大約10万〜約20万トークン程度（Claude 2が100k提示）と長文対応を売りにしていますが、Gemini 3は桁違いの長文を保持できます [28]。もっとも、100万トークンすべてで高精度を出せるかは別問題ですが、前述の通りGemini 3は長文内からの情報検索性能を磨いており [37]、**大量コンテキスト下での処理信頼性**で先行している可能性があります。GPT-5.1も最大128kトークン版が存在しますが [99]、100万には遠く及びません。従って、**超長文や多数文書を跨ぐ分析タスク**では2025年現在、Gemini 3がもっとも適したモデルと言えるでしょう。

速度とコスト: 処理速度に関しては先述の通り、Gemini 3はレスポンス生成が速く開発効率に優れます [54]。対するGPT-5.1はモデルサイズや設計上、やや応答に時間がかかる傾向があります（もっともOpenAIもターボ版などで高速化を図っています）。コスト面では、OpenAIの価格設定は入力1M/$1.25・出力1M/$10とGemini 2.5並みで**割安**です [100]。Gemini 3 Proはそれより高価になるものの [59]、先述したように**性能あたりの費用対効果**を考える必要があります。Claude 4.5 Sonnetはさらに高価であるため [60]、高性能を求めるならGemini 3はまだ現実的な価格設定と言えます。企業が自前で大規模モデルを運用するよりは遥かに低コストで、Google自身も自社TPU v5eなど専用インフラで低コスト推論を実現しているはずです [101]。したがって、**速度・費用面でもGemini 3は総合的に有利**との見方があります [54]。

以上を総合すると、**2025年11月時点の最先端LLM競争において、Gemini 3 Proは論理推論・マルチモーダル理解・コード生成といった主要分野でライバルを僅かにリードする**状況です [82]。OpenAI GPT-5.1も依然強力で、特に対話の自然さや創造的応答で根強い支持がありますが、学術的難問や実務タスク解決ではGemini 3が頭一つ抜けた印象です [82]。Claude 4.5は安全志向と長文対応で一定のポジションを保つものの、性能指標ではGeminiやGPTに遅れを取っています [82]。業界アナリストも「GoogleがGemini 3でAI性能レースの先頭に立った」と評価しており [102]、今後OpenAIやAnthropicが次世代モデル（GPT-5.2やClaude 5など）で巻き返すかが注目されています。いずれにせよ、Gemini 3の登場はAIモデル間競争をさらに加速させ、私たちが享受できるAIの**知的レベルを大きく底上げした**と言えるでしょう [103] [104]。

Sources: [3] [6] [13] [12] [24] [26] [31] [42] [51] [59] [79] [66] [68] [86] [87] [74] [82]

---

[1] [2] [3] [4] [5] [6] [7] [8] [9] [19] [22] [23] [41] [45] [46] [64] [88] [89] [90] [91] [95] [96] [98] [103] Gemini 3: Introducing the latest Gemini AI model from Google
https://blog.google/products/gemini/gemini-3/

[10] [11] [14] GPQA-Diamond Explained: The AI Scientific Reasoning Benchmark | IntuitionLabs
https://intuitionlabs.ai/articles/gpqa-diamond-ai-benchmark

12 17 18 40 92 Gemini 3 Pro released: Is Gemini 3 Pro About to Crush the AI Competition? - CometAPI - All AI Models in One API
https://www.cometapi.com/is-gemini-3-pro-about-to-crush-the-ai-competition/

13 15 16 35 57 58 83 86 87 100 Trying out Gemini 3 Pro with audio transcription and a new pelican benchmark
https://simonwillison.net/2025/Nov/18/gemini-3/

20 21 26 27 28 Gemini 3.0完全ガイド｜Googleが放つ次世代AIの全貌！これを読めば全てがわかる！ - AI OTAKU / AIオタク
https://ai-otaku.jp/398/

24 25 29 30 31 32 33 34 36 37 38 39 93 94 Gemini 3 Pro, DeepThink, and Antigravity | by Yuri Trukhin | Nov, 2025 | Yuri Trukhin
https://blog.trukhin.com/gemini-3-pro-deepthink-and-antigravity-7cbdce2a9137?gi=99522200ef1c

42 43 44 49 54 59 60 61 62 63 79 82 97 102 104 Analysts say Google now leads the AI performance race with Gemini 3 Pro
https://the-decoder.com/analysts-say-google-now-leads-the-ai-performance-race-with-gemini-3-pro/

47 55 Gemini 3.0 Pro. This model is mindblowig in humanity's… | by Balthasar | Nov, 2025 | Medium
https://medium.com/@ibokkon/gemini-3-0-pro-a06f528c7720

48 50 51 52 53 56 84 85 Gemini 3 vs Gemini 2.5 2025 Head-to-Head Benchmark - Skywork ai
https://skywork.ai/blog/llm/gemini-3-vs-gemini-2-5-2025-head-to-head-benchmark/

65 66 67 68 69 70 71 72 80 81 Gemini 3 Pro first impressions : r/singularity
https://www.reddit.com/r/singularity/comments/1p0f6uw/gemini_3_pro_first_impressions/

73 74 75 76 77 78 Gemini 3 Pro Hallucination Rate Vs. Gemini 2.5 Pro : r/singularity
https://www.reddit.com/r/singularity/comments/1p0jc6e/gemini_3_pro_hallucination_rate_vs_gemini_25_pro/

99 Google unveils Gemini 3 claiming the lead in math, science ...
https://venturebeat.com/ai/google-unveils-gemini-3-claiming-the-lead-in-math-science-multimodal-and

101 Google Releases Gemini 3, Calling It Most Powerful AI Model Yet
https://techstrong.ai/articles/google-releases-gemini-3-calling-it-most-powerful-ai-model-yet/