ChatGPT

# Gemini 2.5 Deep Think：リリース後の評判と総合評価

**2025年8月1日に正式リリース**されたGoogle DeepMindの最新AIモデル「**Gemini 2.5 Deep Think**」は、超高性能な大規模言語モデル（LLM）として大きな注目を集めています。本記事では、その性能や新機能、ユーザーや専門家の評判、競合モデルとの比較、さらにコーディングやビジネス支援など特定のユースケースでの有用性・課題について、信頼性の高い情報源に基づき徹底的にまとめます。

## 性能評価とベンチマーク結果

**図：Gemini 2.5 Deep Thinkと他モデルのベンチマーク比較**（数学（USAMO）・コーディング・マルチモーダル推論） [1] [2]

**Gemini 2.5 Deep Think**は、客観的な指標で現行トップクラスの性能を示しています。特に**数学分野の推論力**が飛躍的で、**2025年の米国数学五輪（USAMO）ベンチマークで約49.4%**という驚異的スコアを記録し、従来最強と目されたOpenAIモデル（約19–22%）を大きく上回りました [1] [3]。また**競技プログラミング**指標の「LiveCodeBench」では**80.4%**の正解率を達成し、同じGemini 2.5 Pro通常モード（約71.4%）やOpenAIモデルの成績（70%前後）を凌駕しています [4]。さらに**マルチモーダル推論**ベンチマークMMMUでも**84.0%**という高スコアを記録し [2]、総じて**幅広い高難度タスクで最先端（state-of-the-art）の性能**を示しました。

他のベンチマークでもGemini 2.5 Proは突出しています。学術知識テストのMMLUでは**94.8%**と、人間専門家に迫る正答率を達成（GPT-4は約86%、Claudeは約88%） [5]。大学院レベル物理質問集（GPQA）でも**85.2%**とGPT-4（約78%）を上回り [5]、コード生成のHumanEvalも**92.1%**と他モデルを引き離しています [5]。100万トークンという**超長文コンテキスト**への対応力も特筆すべきで、従来のGPT-4（最大128kトークン）やClaude 2系（最大100k〜200kトークン）をはるかに凌ぐ容量です [6] [7]。実際、数十万トークン規模の複数コードリポジトリを一括投入し、Geminiがそれらを踏まえた高度な設計提案を返した例も報告されています [8]。応答生成**速度**も高速化されており、コード生成はGPT-4比で平均2.3倍のスピードというデータもあります [9]。これら客観指標から、Gemini 2.5 Deep Thinkは**精度・推論力・速度・長文処理**のいずれにおいても現行トップ水準であることが示されています。

## 新機能と前バージョンからの改善点

**Gemini 2.5 Deep Think**では、前身モデル（Gemini 1.5や2.0世代）から多数の改良が加えられています。Gemini 1.5世代（2024年発表）は**「マルチモーダル × 超長コンテキスト」**を特徴とし、テキスト・画像・音声・動画・コードを統合処理でき、最大100万トークンの文脈保持が可能という画期的な土台を築きました [10]。Gemini 2.0（2024年末）では**「思考プロセス生成（Flash Thinking）」**が初導入され、応答前に内部で推論ステップを踏む仕組みが試験的に導入されています [11]。そして今回の2.5では、これらをさらに発展させ**モデルにネイティブに“思考するAI”能力を統合**しました [10]。具体的には、人間の問題解決に倣い**複数の仮説を並行検討してから回答を導く「Deep Think」モード**がGemini 2.5 Proに実験機能として追加されています [12] [13]。この並列思考アプローチにより、複雑な問題に対する推論精度や一貫性が大幅に向上しています [10]。

新機能も多数盛り込まれました。**ネイティブ音声出力**機能により、テキストだけでなく自然な多言語の音声応答が可能となり、会話体験がより豊かになっています [14]。また**Live API**の強化により、開発者が対話中に

モデルの思考過程やツール使用を追跡できる「思考サマリー」機能が追加され、モデルの透明性・デバッグ性が向上しました [15]。加えて、**Project Mariner**という機能を通じて外部ブラウザ操作などの自動タスク実行（エージェント機能）も試験提供され、ユーザーの指示に基づきウェブ検索やアプリ操作を行う**「Agent Mode」**が導入されています [16]。このAgent Modeにより**ユーザーの意図を汲んだ自律的なタスク実行**が可能になり、AIの可能性が大きく拡張されました [16]。

**安全性**の面でも強化が図られています。Gemini 2.5では**高水準の安全評価（フロンティア安全性評価）**が実施されており、特にDeep Thinkモードは提供前に入念なテストと専門家レビューが行われています [17]。**間接的なプロンプトInjection攻撃**への耐性も向上し、ツール使用時に悪意ある指示が紛れ込むケースへの対策が強化されました [18]。実際、Gemini 2.5 Deep Thinkは従来モデルより**有害コンテンツ生成の抑制や客観性**が改善した一方、無害な質問にも慎重になり**回答を拒否しがちになる傾向**も見られたと報告されています [19]。これは安全重視の調整の結果であり、今後さらなる最適化が期待されています。

そのほか、**SDKのMCP（Model Context Protocol）対応**による他社AIツールとの互換性向上や、Gemini API経由での制御性向上（思考バジェット機能拡張）など、開発者向けの改良も多数行われています [15]。総じて**Gemini 2.5**は、**前バージョン比で性能面も機能面も飛躍的に進化**しており、**「考えるAI」**へのシフトとユーザビリティ向上が同時に図られたアップデートと言えるでしょう [10]。

## ユーザー・専門家による評判とSNS上の反応

**Gemini 2.5 Deep Think**に対するリリース直後の評判は、**絶賛**と**懐疑・批判**が入り混じっています。開発者や研究者からはその能力を高く評価する声が多く聞かれ、ある開発者は**「複雑なアルゴリズム設計で、自分では思いつかない創造的アプローチを示してくれる。まさに優秀なチームメンバーを雇ったようだ」**とReddit上で賞賛しています [20]。大手企業CTOからは**「月額36,800円で大容量ストレージ(30TB)やYouTube Premiumも付くのだから、ストレージ代だけで元が取れる。Deep Think機能は言わばボーナスだ」**という声もあり、高額ながら付帯価値も含め妥当との意見もあります [21]。こうしたユーザーからは**「まさにAI分野の革命だ！」**との感嘆も上がっており、専門家の中には**“人間の専門家レベルであらゆる知的作業をこなす”**とその汎用知能ぶりを評価する向きもあります [22] [23]。

一方で**批判的な意見や不満**も少なくありません。特に多いのが**価格と利用制限**に関する指摘です。あるIT企業経営者はX（旧Twitter）上で**「月3.6万円も払うのに1日数回しかDeep Thinkを使えないなんて冗談だ。ヘビーユースできなければ価値がない。他のAIサービスの方がマシ」**と怒りを露わにしました [24]。実際、UltraプランではDeep Thinkモードの**利用回数に1日あたりの上限**が設けられており [25]（※具体的な回数非公表ながら数件程度との報告あり）、この点がヘビーユーザーには不満の種となっています。また一般ユーザーからは**「期待してUltraプランに入ったけど、普通の質問なら無料版のGeminiで十分。Deep Thinkは完全にオーバースペック」**との失望の声も上がっています [26]。つまり「高いお金を払った割に日常用途では宝の持ち腐れ」という評価です。

総じて、**「技術的革新性は認めるがコストに見合うかは用途次第」**というのが現時点での評判と言えるでしょう。研究者や開発者コミュニティでは革新的機能を歓迎する声が多い一方、一般ユーザー層からは「様子見」「価格が下がるまで待ちたい」といった慎重な反応も見られます [27] [28]。Google自身も「Gemini 2.5 Deep Thinkは困難な問題に取り組む研究者・科学者・学者向けのプレミアム体験」と位置付けており [29]、現状ではプロフェッショナル用途を想定したサービスであることが明言されています。

## 他の大規模言語モデルとの比較（強み・弱み）

Gemini 2.5 Deep Thinkの優位性を把握するため、主要な競合LLMとの比較にも触れておきます。**OpenAI GPT-4**（2023年公開の代表的LLM）や、その後継モデル群、**Anthropic社のClaude**シリーズ、**Meta社のLLaMA**などとの**強み・弱みの対比**は次の通りです。

- **GPT-4（OpenAI）との比較:** GPT-4は長らく汎用LLMの代表格でしたが、Gemini 2.5は多くの客観評価でGPT-4を凌駕しています。例えば前述のMMLUや物理問答、コード生成精度で**GeminiがGPT-4を7〜10ポイント以上上回る**結果が報告されています [5]。特に**数学的推論力**での差は顕著で、GPT-4が解けなかった超難問をGemini Deep Thinkは複数解決できるといったケースも出ています [1]。**コンテキスト長**もGeminiは最大100万トークンと、GPT-4（拡張版でも128k程度）の約8倍に及び、大量の文章やコードを一度に分析・生成できる強みがあります [6]。一方**GPT-4の強み**としては、2023年時点から公開APIや幅広い統合が進んでおり、エコシステムの豊富さや安定性で先行していた点が挙げられます。またGPT-4は若干Geminiより**応答が洗練されている**との指摘も一部にあり（会話の自然さや創造的文章のスタイルなど）、細かなチューニング面では依然評価が高いとの声もあります。ただ総合的には、**Gemini 2.5 Proは事実上GPT-4の性能を上回る次世代モデル**との見方が強く、Google自身も「主要な次元で人間の好み評価（Human preference）において全モデル中トップ」と自信を示しています [30] [31]。

- **Claude 2/3（Anthropic）との比較:** Claudeシリーズは**長文読解**や**安全性**に定評があり、Claude 2では最大100kトークンの入出力や安定した応答が特徴でした。しかしGemini 2.5は文脈長と推論力でそれを凌駕し、**100万トークンの文脈＋高度推論**という点でClaudeの強みだった長文処理能力を大きくリードしています [7]。Anthropicの次世代モデル「Claude 3」（開発コード: Sonnet 等）でも一部**拡張思考モード**が搭載され、論理的推論やコーディング能力を強化していると報じられています [32]。実際、Claude 3系列は高度なコーディングでGeminiに肉薄する性能を示すベンチマークもありますが [33]、総合評価ではGemini 2.5 Proが依然リードしています [34]。例えば先述のHumanEvalコードテストではGeminiが92%に対しClaudeは89%程度 [5]、学術知識でもClaude 3が80%台前半に達する一方でGeminiは90%超と、**依然数ポイント〜十数ポイントGemini優位**の結果が見られます [5]。Claudeの強みとしては**応答の一貫した丁寧さ**や**有害発言の抑制**で評価が高く、Geminiも安全性は重視しているもののDeep Thinkでは応答拒否が増える副作用があったため、絶妙な安全度合いではClaudeも競争力があります [19]。またAnthropicは「憲法AI」アプローチによる倫理調整で知られ、クリエイティブな文章生成や要約などで根強いファンを持っています。総じて**Claudeシリーズは安全・安定志向、Geminiは攻めの高性能志向**という住み分けとの指摘もあります。

- **LLaMA系（Meta）との比較:** Meta社のLLaMAおよびLlama 2は**オープンソース**で提供されている点が最大の特徴です。コミュニティ主導でモデルをカスタマイズ・軽量化でき、オンプレミスで利用可能なため、**プライバシーやコスト管理**の面で優れます。しかし**性能面ではトップモデルに一歩及ばない**のが現状です。例えばLlama2 70BモデルのMMLUスコアは約68.9%に留まり、GPT-4（86.4%）やGemini 2.5（94.8%）には遠く及びません [35] [5]。Metaは今後「Llama 3」や「Llama 4」で大幅な改良を計画中とされ、80%台に迫るとの情報もありますが [36]、少なくとも2025年8月時点では**Gemini 2.5の独走状態**に対抗できるオープンモデルは登場していません。LLaMA系の強みは**無料で使えること**と**モデルを自由に手元で動かせる拡張性**であり、特定業務に絞った微調整モデル（例：医療特化やコード特化モデルなど）は商用LLMにない柔軟性があります。ただし汎用能力や総合的な知識・推論力では、**Geminiのような超大規模モデルが依然大きなリード**を保っているのが実情です。

## ユースケース別の有用性と課題

**Gemini 2.5 Deep Think**の真価は、その高度な性能を各種ユースケースで発揮できるかにかかっています。以下、**コーディング支援**、**ビジネス支援**、**学術研究**、**創作支援**の観点で有用性と課題を整理します。

- **コーディング支援:** Gemini 2.5 Proは**「開発者に最も高く評価されているコード生成向けモデル」**と公式に謳われ[12]、実際に**Web開発やアルゴリズム設計で驚異的な成果**を見せています。Deep Thinkモードでは**競技プログラミングレベルの難問**にも取り組め、時間計算量やトレードオフを慎重に考慮した最適解を提示する能力があります[37]。前述のようにLiveCodeBenchやHumanEvalといったベンチマークで最高性能を発揮し[4][38]、**フルスタックのアプリ生成**や**複雑なバグ修正**もこなせると報告されています。実例として、通常モードでは粗い出力だった**VoxelアートのHTML生成**が、Deep Thinkを使うと芸術的センスを感じる洗練されたコードに進化した例もあり、筆者は**「同じAIが作ったとは思えないプロ級の差」**と驚きを述べています[39][40]。課題としては、Deep Thinkは高品質なコードを返す反面**実行に時間が掛かる**点です（場合によっては数分〜数十分）[41]。また大規模コードベースの理解には強いものの、**開発者が意図しない過剰最適化**や**スタイルのクセ**など細部調整が必要になるケースも考えられます。ただ総じて、**高度なプログラミング支援AIとしてGemini 2.5 Deep Thinkは現行最強クラス**であり、専門家レベルのコードレビューアやペアプログラマを得たような効果が期待できます[20]。

- **ビジネス支援:** 複雑なビジネス上の問題解決や意思決定支援においても、Deep Thinkは有用だとされています。例えば**「競合他社の戦略を分析し、自社の差別化戦略をリスクとリターンを含め具体提案せよ」**のような高度な戦略立案の問いに対し、Gemini Deep Thinkは**複数のアプローチを並行検討し最適解を構築する**能力を発揮します[42]。公式にも「反復的な開発・設計」「戦略的計画」がDeep Thinkの得意分野として挙げられており[43]、現実のビジネス課題（市場分析、プロジェクト計画、意思決定支援など）でも**人間のコンサルタントに匹敵する洞察**を与える可能性があります。実際、あるCTOはGemini Ultraプランを**「優秀なビジネスアナリストをチームに加えた感覚」**と評しています[20]。一方課題として、ビジネス利用では**常に最新データへのアクセス**が必要ですが、現状Geminiの情報アップデート頻度や外部データ接続（例えばリアルタイムの市場データ参照）は限定的です。ただしAgent Modeにより限定的ながらウェブ検索で情報収集することも可能になりつつあり[16]、将来的にはより動的なビジネスデータ分析も期待されています。

- **学術研究:** Googleは本機能を**「研究者・科学者・学者向け」**と位置付けているだけあり[29]、学術用途でのメリットが大きいと考えられます。Deep Thinkは**高度に複雑な問題を論理立てて推論**できるため、例えば**数学の未解決問題の仮説検証**や**科学論文の解釈・新規発見**に威力を発揮します[44]。実際、今年の国際数学オリンピックで金メダルを獲得した内部モデルをベースにしており、社内テストでは日常使用可能な速度に最適化しつつIMO銅メダル相当の性能を達成したとされています[45][46]。**論文の内容を踏まえた実験計画の提案**や、複数論文を跨いだ体系的なレビューの作成など、人間研究者の頭脳を拡張するツールとなり得ます。実用上の注意点は、やはり**推論に時間がかかる**ことで、複数時間スケールの考察が必要な場合もあります[45]。また専門領域の深い知識については組み込まれたデータに依存するため、最新の学術知見を取り込みきれていない可能性もあります。しかし今後API経由でツールとの連携（例えばコード実行や外部データ参照）も拡充予定で[47]、例えば実験データを取り込んで考察させる、といった**インタラクティブな研究支援**も視野に入っています。総じて、**難解な研究課題に対する洞察を飛躍的に得られるブレーンとして期待**されており、既に一部の数学者や科学者から「問題解決のアプローチが劇的に広がった」との声も報告されています[48][49]。

- **クリエイティブ創作支援:** Deep Thinkは**創造性を要するタスク**にも強みを発揮します。例えば**物語のプロット作成**や**ゲームデザイン**、**アート作品のアイデア出し**など、従来のLLMでは単調になりがちな領域で「時間をかけた熟考」により質の高いアウトプットを生成できます。実験例として、あるユーザーがWeb上で**「美しい日本庭園の塔をボクセルアートでHTML生成してほしい」**と依頼したところ、通常モードでは陳腐な構成だったのが、Deep Thinkモードでは**桜の花びらの細部や色彩のグラ**

**デーションまで配慮した芸術的シーン**を描き出したといいます [50] [39]。このように、**芸術的センスを持った専門家レベルの創作**が可能との評価もありました [51]。他にも作曲支援や映像脚本のプロット生成などでも有用との声があります。課題としては、創作分野ではしばしば**明確な評価軸が無い**ため、AIの提案が本当に「優れているか」を判断しにくい点です。またDeep Thinkは安全策で**既存作品の過剰な模倣を避ける**傾向が強く（著作権や倫理への配慮）、斬新さを求める一部クリエイターには物足りない場合もあるかもしれません。しかし総合的には、**創造的プロセスを加速・補完するパートナー**として非常に心強く、アイデア発想や高度なリファレンス生成において人間クリエイターの能力を高めるツールとなっています [52] [53]。

---

以上、Gemini 2.5 Deep Thinkのリリース時点での総合的な評判と評価をまとめました。圧倒的な性能向上と革新的機能によって**AIの新たな地平を切り開いたモデル**である一方、その恩恵を最大限活かすには用途を選ぶ面もあります。Googleはまずは専門家コミュニティでのフィードバックを経て安全性・有用性を高めつつ、今後より広範なユーザーに展開するとしています [54] [55]。まさに**「AI推論の歴史的転換点」**とも称されるGemini 2.5 Deep Think [56]が、今後どのように普及し私たちの仕事や生活を変えていくのか、引き続き注目が集まっています。

**参考文献・情報源:** [1] [2] [4] [5] [6] [7] [9] [12] [37] [51] [20] [24] [57] [16] [44] 他、Google公式ブログ、専門家のNote投稿、技術ニュースサイト記事など多数。

---


[1] [3] [4] [17] 〖o3超え〗Gemini 2.5 Pro「Deep Think」モードが数学・コーディングベンチマークで革命的な成果 – OpenAI o3を圧倒的に上回る精度を実現 - チャエンのAI研究所
https://digirise.ai/chaen-ai-lab/gemini-2-5-pro-deepthink/

[2] [12] [13] [14] [15] [18] [30] [31] Google I/O 2025: Google DeepMind から Gemini 2.5 のアップデート
https://blog.google/intl/ja-jp/company-news/technology/google-gemini-updates-io-2025/

[5] [7] [9] [38] 〖⚡速報〗GoogleがChatGPT、Claudeを超えたNo.1のAIモデル『Gemini 2.5 Pro 06-05』を公開 - チャエンのAI研究所
https://digirise.ai/chaen-ai-lab/gemini-2-5-pro-06-05/

[6] [8] [10] [11] 無料で使えるハイエンドのAIモデル「Gemini 2.5 Pro Experimental」の性能が凄い | 天秤AIメディア byGMO
https://tenbin.ai/media/generative_ai/gemini-25-pro-experimental-performance

[16] 〖2025年最新版〗Google Gemini AIエンジニアが徹底解説！‣はてなベース株式会社
https://hatenabase.jp/blog/gemini-models-comparison-guide/

[19] [37] [43] [44] [45] [46] [49] [52] [53] Gemini 2.5 Deep Think の概要｜npaka
https://note.com/npaka/n/ncf6ddb58a2cf

[20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [39] [40] [41] [42] [47] [50] [51] [55] [56] [57] 〖衝撃検証〗Google Gemini 2.5 Proの新機能「Deep Think」が別次元〜月3.6万円の価値を徹底分析｜AidX 研究所（アイデックス ラボ）
https://note.com/aidxlab/n/nc9ad550b5f92

[32] [33] [34] ジェミニ2.5プロ：AIのライバルとの比較分析（2025年の風景）
https://dirox.com/ja/post/gemini-2-5-pro-a-comparative-analysis-against-its-ai-rivals-2025-landscape

[35] Meta Llama 2 vs. OpenAI GPT-4 - by Diana Cheung - Medium
https://medium.com/@meetdianacheung/meta-llama-2-vs-openai-gpt-4-785589efe15e

[36] Llama 4の性能と評価を徹底比較！GPT-4、Gemini、Claudeとの ...
https://no1s.biz/blog/8007/

48 54 Gemini 2.5 Pro、人間的な並列思考で問題解決する「Deep Think」 - Impress Watch
https://www.watch.impress.co.jp/docs/news/2036289.html