

Gemini 2.5 Pro 「Deep Think」モードの性能徹底調査

1. 技術スペック（処理速度・モデル規模・学習データ・推論性能）

モデル概要: Gemini 2.5 ProはGoogle DeepMindが開発した次世代AIモデル「Gemini」ファミリーの最上位モデルです ¹。このモデルは**スパースMixture-of-Experts (MoE)**アーキテクチャを採用しており、入力ごとに一部のパラメータ（エキスパート）のみを活性化させる仕組みにより、**膨大な総パラメータ数を持ちながらも効率的な推論を実現**しています ²。具体的なパラメータ数は公開されていませんが、設計上は**実質的に数兆規模のパラメータ**を持つとも推測され、従来モデルを凌駕する容量を持ちながらトークンあたりの計算コストを抑えています ²。

コンテキスト長: Gemini 2.5 Proは**最大100万トークン**の文脈を扱える超長文対応モデルで、**1Mトークンのコンテキストウィンドウ**を備えています ³ ⁴。出力トークン数も最大64,000トークンに達し、非常に長い文章や巨大データセットを一度に処理・要約することが可能です ⁵。この長大なコンテキスト能力により、大規模文書全体を読み込んだ上での分析や、長時間の動画（最大3時間分）の内容理解など、**従来不可能だった規模の情報を一括処理**できる点が大きな特徴です ⁶ ⁷。

マルチモーダル対応: Gemini 2.5 Proは**ネイティブにマルチモーダル対応**しており、テキストだけでなく**画像・音声（オーディオ）・動画・PDF**といった多様なデータを入力として理解できます ⁸ ⁹。出力は基本的にテキストですが、音声生成にも対応しており、**24言語で自然な音声読み上げ**を行うことができます ¹⁰。また、**長大なコードベース**（リポジトリ丸ごと等）を入力しコード全体の構造を理解することも可能で、これにより大規模なコード解析や整合性チェックも行えます ¹¹ ¹²。

学習データ: モデルの事前学習には**超大規模かつ多様なデータ**が使用されました。公開情報によれば、**公開Web文書、各種プログラミング言語のコード、画像、音声（音声データや環境音）、動画**など、多岐にわたる領域のデータを含む巨大データセットで学習しています ¹³。また、学習後期には**指示文と応答のペアデータ**や**人間の好み（RLHF）、ツール使用データ**などを用いた指示追従・能力向上のための**微調整**が行われています ¹⁴。安全性確保のため、データからの**重複削除や不適切な内容のフィルタリング**も徹底されています ¹⁵。

処理速度: Gemini 2.5 Proは非常に高度な推論を行うため、推論には相応の計算資源と時間を要します。特に**Deep Thinkモード**（後述）を有効化すると、**通常モードより応答生成に時間がかかり、回答まで数十秒～数分程度を要**することがあります ¹⁶。これはDeep Thinkが裏側で**複数の仮説を並列検討し、自己修正を繰り返して最適解を導く**ためです ¹⁷ ¹⁶。一方でGeminiファミリーには**軽量・高速な変種**も用意されており、例えば**Gemini 2.5 Flash**や**Flash-Lite**は品質と速度のバランスを調整したモデルです ¹⁸ ¹⁹。Flash系は**遅延を抑え日常タスクに適した応答速度**を実現しており、プロンプトに対する応答をより短時間で返すことが可能です ¹⁸。

推論性能: Gemini 2.5 Pro（Deep Thinkモード有効時含む）は現時点で世界トップクラスの性能を示しています。例えば**2025年米国数学オリンピック（USAMO）**問題に対し、Gemini 2.5 Pro Deep Thinkは**49.4%**という驚異的なスコアを記録し、OpenAIの強力モデル（OpenAI o3やGPT-4相当）の約20%前後というスコアを倍以上引き離しました ²⁰。また**競技プログラミング性能指標**の一つである**LiveCodeBench v6**でも、Gemini 2.5 Pro Deep Thinkは**80.4%**というトップ級の正解率を達成しています（通常の2.5 Proでも71.4%と高水準） ²¹。さらに、研究分野で注目される「**Humanity's Last Exam (HLE)**」といった総合知識テストで

も他の最先端モデルを大きく上回るなど、**高度な数学・科学知識の推論や複雑なコード生成で卓越した成果**を示しています²²²³。Google公式発表によれば、Gemini 2.5 Proは学術ベンチマークで軒並み最高性能を記録し、**Web開発用コード生成コンペ「WebDev Arena」でも世界トップ（ELOスコア1420）**、AIモデルの多次元評価「LMArena」でも全指標で首位に立っています²⁴²⁵。加えて、100万トークンの長大な文脈や動画入力を活かし、**長文読解や動画理解においても最先端の性能**を示しています²⁶。

2. 他の主要AIモデルとの比較（GPT-4／Claude／Mistral等）

Gemini 2.5 Pro (Deep Think) は、競合する主要AIモデル（OpenAI GPT-4系、Anthropic Claude系、Mistralなど）と比べても**多くの点で卓越した性能と特徴**を持ちます。以下に各モデルとの比較をまとめます。

- ・ **モデルの位置付けと強み:**
 - ・ **Gemini 2.5 Pro:** 超長文・マルチモーダル処理を得意とし、**高度な推論力とコーディング能力で突出**しています²⁷²²。特にDeep Thinkモードにより数学証明や高度なプログラミング問題で人間専門家に迫る成果を上げています。
 - ・ **OpenAI GPT-4:** 汎用会話AIの代表格で、**対話の流暢さや創造的応答**に優れ、市場シェアも大きいモデルです²⁸。GPT-4(特に2025年時点のGPT-4.1)はマルチモーダル入力（画像等）もサポートし、**会話の自然さと多分野で安定した高性能**が強みです²⁸。
 - ・ **Anthropic Claude:** 安全性と長文対応に注力したモデルで、**最大20万～50万トークンもの広大な文脈**（Claude 3.5 Sonnetでは基本20万、企業版で最大50万）を扱え²⁹、**コーディング支援や高度な倫理基準への準拠**に強みを持ちます²⁸³⁰。Constitutional AIによる**毒性の低減**やプライバシー配慮にも定評があります。
 - ・ **Mistral:** フランス発のオープン系モデルで、軽量かつコスト効率を重視したアーキテクチャが特徴です³¹。**モデルの軽量最適化により高速で、大規模処理においてコスト当たりのパフォーマンスが高いことから、大量トラフィックやデータ主権（欧州内運用）を重視する用途に適しています**³¹。ただし総合的な知能や創造力では、大規模モデルのGPT-4やGeminiに一步譲ります。
- ・ **性能比較:** 推論精度のベンチマークではGemini 2.5 Pro (Deep Think)が**数学・科学・プログラミングの難問領域で首位**に立っています²²²³。GPT-4も総合力が高く創造的タスクで強みがありますが、**数学のようなステップ推論問題ではGemini (Deep Think) に及ばない**とされています²²。Claudeは**長文要約やコード補完で優秀な結果**を示しつつ、安全性の高さから**企業利用で評価**されています³⁰。Mistralは公開モデルとしては高性能なものの、**上位モデル（GPT-4/Gemini）ほどの推論力は持たず**、主に効率とオープン性で差別化されています。
- ・ **コンテキスト長:** Gemini 2.5 Proと最新のGPT-4.1は**いずれも最大約100万トークン**のコンテキストを扱える超長文モデルです³²⁷。Claude 3.5も標準で20万トークン（エンタープライズ版50万）と長文に強く²⁹、一方で**Mistralの主力モデルは約12.8万トークン程度**の文脈長となっています³³。したがって**超長文ドキュメントの一括処理**においては、Gemini 2.5 ProおよびGPT-4系が突出しており、Claudeも実用十分な長大コンテキストを提供しています。
- ・ **マルチモーダル対応:** Gemini 2.5 ProとGPT-4は**いずれもテキストだけでなく画像や音声などマルチモーダル入力に対応**しています⁹²⁸。GPT-4（Vision機能）は画像解析が可能で、Geminiはさらに動画・音声まで統合的に処理できます⁹。Claude（Opus系）は2025年時点では主にテキスト特化（画像入力等は限定的）であり³⁴、Mistralも基本はテキストモデルとして提供されています（画像対応の派生は報告なし）。**多言語対応**では、Geminiは音声出力含め24言語を流暢に操り¹⁰、GPT-4も主要言語で高性能、Claudeも100以上の言語で堅実に動作するなど**いずれも多言語サポートは強力**です。Mistralはオープンモデルとして多言語データで学習されており、規模の割に多言語能力は高いものの、**トップモデルほどの緻密な多言語最適化は未成熟**と言えます。

・**利用コスト**: API利用時の料金や提供形態も重要な比較ポイントです。以下に各モデルの参考コストと提供状況を示します ³⁵ ³⁶ :

モデル (提供元)	コンテキスト長	API利用料金(入力/出力)	提供形態・プラン
Gemini 2.5 Pro (Google)	1,048,576トークン (約100万) ⁷	\\$1.25 / \\$10.00 (100万 トークンあたり) ³⁷ ※音声・画像入力も 同一レート	Google AI Studio/Vertex経 由API、 Geminiアプリ (要AI Ultra契約 \\$249.99/ 月) ³⁸
GPT-4 (OpenAI)	1,000,000トークン (GPT-4.1) ³² 従来GPT-4は32k~ 128k	\\$2.50 / \\$10.00 (100万 トークン) ³⁹	OpenAI API (要利用料従量 課金)、 ChatGPT Plus \\$20/月 (32k版)
Claude 3.5/4 (Anthropic)	200,000~500,000 トークン ²⁹	\\$3.00 / \\$15.00 (100万 トークン) ³⁹	Anthropic API (従量課 金)、 Claude Instant/ Proプラン (商用契約)
Mistral (Mistral AI)	128,000トークン ³³	\\$0.40 / \\$2.00 (100万 トークン) ³⁵ ※小 型モデル例。大規模版は 非公開	オープンモデル (7B他) 無 償公開、 商用向けAPI 提供あり

表注: 上記料金は一例で、モデルやプランにより変動します。またGeminiの「AI Ultra」は最上位サブスクリプションでDeep Think利用に必須であり、月額約37,000円と高価ですが、現在は期間限定で半額キャンペーンも実施されています ³⁶ ⁴⁰。

この比較から、**Mistral**は価格面で極めて廉価であり大量トークン処理でもコストを抑えられますが、性能面では**Gemini/GPT-4/Claudeのプレミアムモデルが圧倒的**です ²⁸。**GPT-4**は市場シェアが大きく幅広い用途に使われますが、OpenAIのAPI利用料は決して安価ではありません (高度なGPT-4モデルでは入出力合わせて\\$100+/Mトークン程度のプランも存在)。**Gemini 2.5 Pro**はGoogleクラウド経由で使えるものの、高性能ゆえ利用コストも高めで (単価自体はGPT-4並かやや割安ですが、Deep Thinkはサブスク契約が必要)、その価値は高度な課題に挑む場合に発揮されます ⁴⁰。**Claude**は価格設定が柔軟で、低スペック版 (Claude Instantなど) は低コスト、高スペック版 (Claude Opus 4など) はGPT-4級の料金設定となっています ⁴¹。総じて、**Gemini 2.5 Pro (Deep Think)**は他モデルより高価な部類ですが、**得られる推論力・長文処理能力は群を抜いている**ため、費用対効果は用途次第と言えるでしょう ⁴⁰。

3. マルチモーダル対応力 (画像・動画・コード・多言語・長文推論)

画像・動画理解: Gemini 2.5 Proは**画像や動画を直接入力**として解析できる数少ない大規模モデルです ⁹。例えば画像を与えて内容を説明させたり、動画ファイルを入力してその中の出来事を要約・質問応答するといったことが可能です。実際、社内評価では**視覚的な推論ベンチマーク**において他のテキスト特化モデルを上回る高スコアを記録しています ³⁴。一例として、視覚情報を含む総合テストMMMUでは、Gemini 2.5 Proは**単一試行で82.0%の正答率**を示し、OpenAIのモデル (約81.6%) やAnthropicのモデル (76.5%) よりも高い成績でした ³⁴。さらにGeminiは**動画コンテンツも3時間分まで解析可能**であり ⁶、映像内の物体・シーン理解や、映像ストーリーの要約などを行えます。例として「職場でヘルメットを着けていない人を映像から検出してハイライト」するといった**業務用ビデオ解析**のデモも報告されています ⁴²。

コード理解・生成: Gemini 2.5 Proはコード処理能力が非常に高く、特にPro版は「コーディングに最適」なモデルと位置づけられています⁴³。単純なコード生成だけでなく、複雑なプログラムのデバッグや大規模コードベースの理解も得意です¹¹。実際にGeminiは対話中にプログラムを記述・実行するツール統合機能を持ち⁴⁴、ユーザーの指示に応じて関数呼び出しでコードを実行し結果を反映することができます。Googleのデモでは、Gemini 2.5 Proが数行の指示から「エンドレスランナー」のブラウザゲームを自動コーディングしたり、マンデルブロ集合のフラクタル可視化プログラムを生成した例が紹介されています⁴⁵⁴⁶。ベンチマークでも、コード生成コンテストLiveCodeBenchでトップクラスのスコア（上記の通りDeep Thinkで80%超）を記録し、コード編集タスク（Aiderベンチマーク）でも他モデルを上回る修正成功率を示しています²¹⁴⁷。多言語のコード（PythonだけでなくJSやC++など）にも対応しており、競技プログラミングからWeb開発まで幅広く活用可能です⁴⁸。

多言語対応: Gemini 2.5 Proは多言語での高度なやり取りが可能で、テキスト出力はもちろん、前述の通り24言語の音声読み上げも同一の合成音声で行えます¹⁰。これはユーザーが日本語で質問し、それに対する回答を英語音声で読み上げる、といった場面でも一貫した声質で各国語を発話できることを意味します。また多言語テキストの理解・翻訳も得意で、社内評価のGlobal MMLU（多言語学力テスト）では89.2%という非常に高い平均正答率を達成しています⁴⁹。この数字はGPT-4やClaudeを含む他モデルと比べても最高水準の多言語対応能力を示しています（参考: ClaudeやGPT-4も主要言語で高得点だが、Geminiはより幅広い言語で均質な性能とされる）²⁸。日本語での回答品質も極めて高く、事前学習に大規模日本語コーパスが含まれていることから、専門的な日本語文章の生成・要約なども自然に行えます。

長文推論: 前述のようにGemini 2.5 Proは桁違いの長文コンテキストを保持できるため、長大な文章の一括分析や複数文書横断の推論に強みがあります。例えば数百ページに及ぶ技術論文集や書籍データをすべて与え、その内容を理解・要約させたり、共通するテーマを抽出させるといったことが可能です。Googleの発表でも、「広範な社内ナレッジベースの分析」などにGeminiの長大コンテキスト活用が適していると述べられています⁷⁵⁰。実験的ベンチマークのMRCR (Massive Recall Reading) v2では、128kトークン規模の入力でGeminiは約58%のスコアを記録し、OpenAIや他社モデルの36~57%を上回りました⁵¹。さらに桁違いの100万トークン入力が必要な極限状況でも、Geminiは一応回答を返すことができ（ポイントワイズ評価で16.4%の正答率）⁵¹、他モデルはそもそも対応できませんでした⁵²。このことから、Gemini 2.5 Proは長大な履歴や大量データを踏まえた推論力で群を抜いていることがわかります。例えば企業内の何年分もの文書やチャットログをすべてモデルに与え、横断的に質問に答えさせる、といった高度な知識統合タスクにも適したプラットフォームと言えます⁷⁵⁰。

4. 利用可能なAPIや開発者向け機能（拡張性・ツール統合など）

API提供とプラットフォーム: Gemini 2.5 ProはGoogleの提供する複数のプラットフォーム経由で利用可能です。⁵³にあるように、一般ユーザー向けには「Geminiアプリ」（チャットUI、現在は招待/有料）、開発者向けには「Google AI Studio」（ブラウザ上でモデルを試せる環境）、エンタープライズ向けには「Vertex AI」（Google Cloud上のAIサービス）で提供されています⁵³。またREST APIとしても「Gemini API」が提供されており、アプリケーションから直接モデルを呼び出すこともできます⁵⁴。これらを通じてチャットボットや自社システムへの組み込みが可能であり、OpenAI APIと同様にHTTP経由でプロンプトを送信し応答を受け取るスタイルです。開発者向けSDKも用意され、PythonやNode.js等から簡単にGeminiモデルを利用できます⁵⁵。

拡張性とツール統合: Gemini 2.5は外部ツールの統合機能が充実しています。他のLLM同様、関数呼び出し（Function Calling）のインターフェースを持ち、モデルが回答中に特定の関数（API）を呼んで追加情報を取得したり、計算を実行することが可能です⁵⁶。例えばモデルが「現在の天気」を尋ねられた際、自身でウェブ検索APIを呼び出し回答に反映するといったツール使用が組み込まれています⁵⁶。またコード実行環境との連携もサポートされており、ユーザーの要求に応じてPythonスクリプトを生成・実行し、その結果を返すといった高度な対話もできます⁴⁴。これらはGoogle DeepMindのProject Marinerなどの研究成果を取り入れた機能で、Gemini 2.5 Proは対話中に自律的に外部システムを操作するエージェント機能を備えている

といえます²⁷。実際、Geminiは対話エージェントがPokémonゲームをプレイする実験（Gemini Plays Pokémon）など、**エージェント的な応用例**も報告されています⁵⁷。

思考の可視化・制御: 開発者向け機能としてユニークなのは、モデルの「思考過程」を制御・観察できる仕組みが提供されている点です。Gemini 2.5 Proでは推論中のチェーン・オブ・ソート（内部思考）の**ステップ数（思考バジェット）を調整可能**であり、予算を増やせばより綿密な検討を行う代わりに時間とコストが増え、減らせば簡略思考で応答を高速化する、といったトレードオフ設定ができます⁵⁵。また、API経由では**モデルが裏で考えた内容のサマリーを取得**するオプションも追加されました⁵⁵。これはモデルがどのような推論経路を辿ったか、回答根拠の要約を返す機能で、**ブラックボックスなAIの透明性向上**につながります。例えばユーザからの質問に対し、Geminiが内部でどんな事実や知識を参照して結論に至ったかを要約させ、説明責任を果たす用途が考えられます⁵⁵。これらの機能は**フロンティアAIの信頼性を高める取り組み**の一環であり、開発者がモデルの出力を検証・制御しやすくするためのものです。

Model Context Protocol (MCP): GoogleはGemini APIおよびSDKにおいて、**Model Context Protocol (MCP)**という標準仕様のサポートも表明しています⁵⁵。MCPはモデルに対して追加コンテキストやツールをシームレスに組み込むためのプロトコルで、例えば社内のデータベースやCMSとLLMを連携させ、**最新社内情報をモデルのコンテキストに含めて応答**させるような統合を容易にします。これにより、**オープンソースの外部ツールやプラグイン**をGeminiと接続し、モデルの知識や機能を拡張することができます（OpenAIのPluginsやLangChain的な機構の標準化版と捉えられます）。MCP対応により、開発者はGeminiを**自社独自ツールやデータと統合したカスタムAIシステム**として活用しやすくなると期待されています⁵⁵。

提供エコシステム: Gemini 2.5 ProはGoogleの他サービスとも統合が進んでいます。例えばGoogleドキュメントやスプレッドシート上でGeminiを呼び出して高度な要約や分析を行う「**Duet AI**」機能や、AndroidデバイスでのAI支援機能など、Google製品群への組み込みが段階的に図られています（2025年I/Oでの発表より）。現時点でGemini 2.5 Pro自体をオンデバイス動作させることは困難ですが、将来的に**モバイル向け軽量モデルやデバイス組み込み専用モデル**（例：上記のGemmaファミリー）も想定されています⁵⁸⁵⁹。開発者はVertex AI上でGeminiを用いる際、データ所在地のリージョン選択やアクセス制御など**クラウドインフラとしての管理機能**も利用でき、企業システムと安全に統合することができます⁵³。

5. セキュリティ・プライバシーへの配慮（データ利用方針・情報保護・オンデバイス処理の可否等）

安全性とガードレール: Google DeepMindはGeminiの開発において「安全性」と「責任あるAI」を最優先事項に掲げており⁶⁰、Gemini 2.5ファミリーは**これまでで最も安全性の高いモデル**となるよう設計・テストされています⁶¹。特に最上位のDeep Thinkモードについては「フロンティア安全性評価」の対象として、通常以上に時間をかけた厳格なテストと専門家レビューを実施しており、**信頼できるテスターにのみ限定公開してフィードバック収集**する段階を経ています⁶²。このような慎重なプロセスにより、**重大な危険能力（Critical Capability Level）に達しないことを確認**しつつ一般提供を進めています⁶¹。実際、内部の安全評価ではサイバー悪用・バイオリスク等に関するシナリオでGemini 2.5 Proがいずれも警戒ラインを超えないことが確認されており、既存モデルより安全性指標が向上しました⁶³⁶⁴。

プロンプト悪用対策: 特筆すべきセキュリティ強化として、「**間接的なプロンプト注入**」攻撃への防御があります。これはAIに読み込ませたテキスト内に悪意ある指示を隠し、モデルを不正動作させる攻撃手法ですが、Google DeepMindはこれに対抗するため**自動レッドチームing（ART）**を活用した継続的な検証を行いました⁶⁵。社内のセキュリティチームがGeminiを相手に様々な攻撃プロンプトを自動生成・適用し、防御の抜け穴を探す仕組みです⁶⁶。その結果得られた知見をもとに、**Spotlighting**や**自己検証（Self-reflection）**等の多層防御策を組み合わせ、加えてモデル自体を**有害指示を無視するよう追加学習（model hardening）**させました⁶⁷⁶⁸。このモデル強化により、Geminiは巧妙な隠し命令を**高精度で識別して無視**できるよう

になり、ツール使用時の間接プロンプト注入に対する耐性が大幅に向上しました⁶⁹⁷⁰。結果として**Gemini 2.5は歴代モデル中もっともセキュアなモデルファミリーになった**と報告されています⁶⁹。

出力のフィルタリング: コンテンツ安全性についても、Googleは多角的な対策を実装しています。訓練段階で不適切コンテンツをフィルタしたのに加え、生成時にも**トラウマ誘発やヘイト・差別、違法助長などの出力を防ぐフィルタ**が統合されています。例えば、児童に有害な出力に関しては専門チームが設定した高基準のスコア閾値を採用し、Gemini 2.5 Proではリリース時よりさらに誤出力率を下げています⁷¹⁷²。また**ユーザーからのフィードバック**（低評価等）を逐次モデル改良に反映し、有害発言や事実誤りの低減を図っています。Google DeepMindは安全性に関するホワイトペーパーも公開し、モデルの防御アプローチを透明化するとともに外部監査も受けています⁷³⁷⁴。

プライバシーとデータ利用: プライバシー面では、Googleは**ユーザーデータの取り扱いに関する設定**を提供しています。GeminiアプリやAI Studioでは「Gemini Appsアクティビティ」というオプトイン項目があり、これを**有効にしている場合のみユーザーとの対話データをモデル改善に利用**すると明記されています⁷⁵。逆に言えば、デフォルトでデータを学習利用しない設定も可能であり、ユーザーは履歴を削除したり学習への利用を拒否できます⁷⁶。企業向けのVertex AI経由でGeminiを利用する場合、**送信したプロンプトや出力が第三者のモデル訓練に使われることはない**ポリシーとなっており（Google Cloudのデータ利用ポリシーに準拠）、機密情報を扱うシナリオでも安心です。⁷⁷で述べられているAnthropic Claudeの例では、90日以内にプロンプト・応答ログを自動削除し学習に使わないとされていますが、Googleも同様に**一定期間後の自動削除や暗号化保管**など業界水準のプライバシー対策を講じています（詳細はGoogle Privacy PolicyおよびGemini Privacy Hubに記載）。さらに、政府機関向けにはオフライン環境で動作する特別モデルの提供も検討中とされ、例えば**機密情報専用のGeminiモデル**をクラウド隔離環境で提供することで、高度な秘密保護要求に応える計画もあります。

オンデバイス処理: Gemini 2.5 Proほどの巨大モデルになると、その**推論には数十～数百GB級のメモリと専用アクセラレータ**が必要となるため、現状スマートフォンやPCといったオンデバイスで直接実行することは現実的ではありません。基本的には**クラウド上のTPU/GPUでホスティングされたAPIサービス**として利用する形になります⁷⁸。一方で、Googleはモバイル向け小型モデルとして**Gemini Flash-Lite**や「**Gemma**」シリーズの提供も発表しています⁵⁹。Gemma等は軽量・オープンなモデルで、必要リソースを大幅に削減しており、将来的にエッジデバイスでの推論も視野に入れた設計です（例えば数億～数十億パラメータ級であればスマホでも実行可能）。2025年現在、Gemini 2.5 Proそのもののオンプレミス提供やローカル実行版はありませんが、GoogleはAIをデバイス上で動かす取り組み（例：Pixel端末でのオンデバイス音声モデルなど）も進めており、**シナリオに応じてクラウドと端末内AIを使い分けるハイブリッド戦略**を採っています。機密性が特に重要な場合は、クラウド接続を断った**オフラインモード**で小型モデルを走らせ、必要時にクラウドのGemini Proを使うといった運用も可能です。総じて、現時点では**最高性能を発揮するにはクラウド上でGemini 2.5 Proを利用するのが前提**ですが、用途次第でデータをクラウドに出さずローカル処理する選択肢（小型モデル利用など）も用意されています。

まとめ: Gemini 2.5 Pro Deep Thinkは、その卓越した性能と多機能ぶりから「AI技術の転換点」と評されています²²。技術スペックでは桁違いの規模と能力を示し、他モデルとの比較でも多くの面でリードしています。画像・動画・コード・多言語といった様々な分野に対応でき、開発者にとっても統合・拡張がしやすいプラットフォームが整備されています。さらに、安全性・プライバシーにも最新の配慮がなされており、業務利用にも耐えうる信頼性を備えています。**世界最高峰のAI**として、研究開発やビジネスに革新をもたらすポテンシャルを持つのがGemini 2.5 Pro Deep Thinkです⁷⁹。

1 2 5 11 12 13 14 15 78 Gemini 2.5 Pro - Model Card

<https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf>

3 8 9 10 34 37 43 44 45 46 47 48 49 51 52 53 56 58 59 63 64 71 72 Gemini Pro - Google DeepMind

<https://deepmind.google/models/gemini/pro/>

4 17 20 21 22 【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/>

6 18 19 57 storage.googleapis.com

https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf

7 28 29 30 31 32 33 35 39 41 50 77 Claude, GPT, Gemini, or Mistral? Here's What Powers the Best AI Chatbots in 2025

<https://quidget.ai/blog/ai-automation/claude-gpt-gemini-or-mistral-heres-what-powers-the-best-ai-chatbots-in-2025/>

16 23 36 38 40 79 世界最高性能AI 「Gemini 2.5 Deep Think」 を徹底解説！実際に使った結果も紹介 | SHIFT AI TIMES

<https://shift-ai.co.jp/blog/33115/>

24 25 26 27 54 55 61 62 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/>

42 Gemini 2.5 を使用した会話型画像セグメンテーション

<https://developers.googleblog.com/ja/conversational-image-segmentation-gemini-2-5/>

60 普遍的な AI アシスタントの実現に向けた Google のビジョン

<https://blog.google/intl/ja-jp/company-news/technology/google-deepmindgemini-universal-ai-assistant/>

65 66 67 68 69 70 73 74 Advancing Gemini's security safeguards - Google DeepMind

<https://deepmind.google/discover/blog/advancing-geminis-security-safeguards/>

75 Google Gemini Chats Just Got More Personal - Yahoo! Tech

<https://tech.yahoo.com/ai/articles/google-gemini-chats-just-got-160100045.html>

76 Revoke permission to train on data. - Gemini Apps Community

<https://support.google.com/gemini/thread/342177585/revok-permission-to-train-on-data?hl=en>