ChatGPT

日本の主要LLMと数学・物理推論能力の比較

国内開発の最新LLMとその特徴

NTT：「tsuzumi 2」（つづみ2）

- **モデル概要**：NTTが開発した純国産LLMの最新版で、約300億パラメータ規模（前版は70億）に拡張されています [1]。日本語の文章理解・要約・指示応答性能が大幅に向上し、日本語領域では巨大モデル（GPT相当）に匹敵する実力を軽量構成で実現した点が特徴です [1]。1台のGPUで動作可能な軽量設計でオンプレミス運用もでき、機密データを安全に扱える高セキュリティモデルです [2]。
- **数学的推論能力**：日常レベルの算術や文章題には対応できますが、高度な数学問題や定理証明の領域では**ChatGPT-5など海外先端モデルに及ばない**と報じられています [3]。特に物理・数学・コーディングの性能はChatGPT-5に劣るとの評価があり、専門的な数式処理や難解な問題解決は不得意とみられます [4]。
- **物理分野への対応**：一般的な物理知識の質問には回答可能です。例えば高校レベルの物理法則の説明や教科書的な問題には対応できます。しかし、**高度な物理の応用問題や専門的な推論**になると正確性が低下しやすく、ChatGPT-5クラスのモデルとの差が見られます [3]。日本語での物理会話や資料要約は得意な一方、数式を用いた複雑な物理計算の正確な解答は難しい場合があります。
- **ステップバイステップ推論**：基本的な逐次推論（段階的な説明や理由付け）は可能です。指示に対して推論過程を示す回答も生成できます。ただし**複雑な論証や長いチェーンオブソート**になると誤謬が生じやすく、推論過程の厳密さでは最先端モデルに及びません。とはいえ日本語対話での理由説明や手順解説など、ビジネスで重視される基本的な推論能力は備えています [1]。
- **数式や論理記述の出力精度**：簡単な方程式や数式の記述は可能です。例えば一次方程式の解や簡単な数式変形をLaTeX風に出力することもできます。しかし**複雑な数式変形や長い論理式の正確な出力**は苦手です。証明問題では論理の飛躍や誤りが混入しやすく、計算ミスも起こり得ます。NTTは金融・医療・公共分野の知識強化には注力していますが [5]、数学の専門的記述の正確性向上は今後の課題です。
- **API・商用利用**：**NTTは「tsuzumi 2」を2025年10月より法人向けに提供開始**しました [6] [7]。オンプレミスやプライベートクラウド環境で運用可能で、企業のニーズに応じて導入できます。現時点で一般開発者向けの公開APIはなく、NTTと契約した企業・自治体向けに提供される形です。純国産モデルのため海外クラウドを介さず利用でき、**高セキュリティ・低コスト**なDX支援基盤として位置付けられています [2]。

NEC：「cotomi」（コトミ）

- **モデル概要**：NECが独自開発した日本語特化の生成AI基盤モデルです。約130億パラメータと報じられており [8]、**日本語の知識・読解タスクで世界トップクラスの性能**を示しています [9]。日本語ベンチマークJGLUEでは知識量81.1%、読解力84.3%という高スコアを記録し、海外製LLMを大きく上回る日本語処理能力を実現しています [9]。特徴として128Kトークンという超長文コンテキストに対応し、一度に20万字規模の日本語文書を扱える点が挙げられます [10] [11]。また**MCP（Model Context Protocol）仕様**に準拠したエージェント機能を備え、他AIや外部ツールと連携してタスクを自動化できるのも特色です [12] [13]。
- **数学的推論能力**：一般常識問題や基礎的な数学計算には対応可能ですが、NEC自身は**高度数学タスクへの特化はまだ限定的**としています。ウェブ上の学習データに数学向けの情報が少ないため、専門的な数学問題の正解率は突出して高いわけではありません。ただし**問題解決の過程に着目した学習を強化**する取り組みを行っており、推論過程を踏んだ回答の精度向上が図られています [14]。例えば文章

で与えられた算数ストーリー問題をタスク分解し、一つ一つ解いていくような推論には一定の強みを発揮します。

- **物理知識・応用問題への対応**：物理分野について明確なベンチマーク結果は公表されていませんが、**一般的な物理学知識や教科書レベルのQ&Aには対応**できます。専門的な物理応用問題については、cotomi単独で解を出すよりも、**長文コンテキストを活かして関連資料を読み込ませたりツール連携することで解決を図る**想定です [15]。実際、128Kもの長大な文脈を保持できるため、例えば研究論文や技術資料を与えて内容を要約・質疑応答させるといった応用が可能であり、物理分野でも資料理解型のタスクには強みがあります。
- **ステップバイステップ推論の正確性**：cotomiは**エージェント機能強化によりタスクプランニングやツール選択を適切に行う能力**を向上させています [14]。ユーザーの依頼を自律的にタスク分解し、最適な手段を選んで順次実行する能力があり、複雑なマルチステップ推論でも回答の誤りを減らしています [14] [16]。例えば「まず数式を解き、次にその結果を元に文章を生成する」といった複数工程の指示に対しても、適切に段取りを踏んで回答できる傾向があります。**高度な論理推論力や計画実行力が強化**されており、処理速度と回答品質の両面で向上が見られます [14]。
- **数式や論理記述の出力精度**：長文コンテキストを扱えるため、**証明問題などで途中の論理ステップを保持した説明**を出力することも可能です。数学記号やLaTeX風表記で数式を回答に含めることもできます。ただし、cotomi自体が数学専用に最適化されたモデルではないため、**出力された数式や証明の厳密性には注意が必要**です。特に長い証明過程では一部飛躍があったり、計算ミスのチェックが甘いケースも考えられます。とはいえ、マルチステップ推論能力の向上により**以前より論理的に筋道立った回答**を示す傾向があります。
- **APIや商用利用**：NECは**自社ソリューションの一環としてcotomiを提供**しています。外部にはエンタープライズ向けに組み込む形で展開しており、2025年7月には強化版を発表しました [12]。MCP標準への準拠により社内外サービスとの連携が容易になっており、例えばドキュメント管理クラウド（Box社）との連携実証も進めています [17]。**現時点で一般開発者向けの公開APIはなく**、NECと協業する形で特定用途に導入するケースが多いようです。ただ、企業内AIエージェントとして業務自動化を支援する位置付けのため、必要に応じて個別案件でカスタマイズ導入される形になります。性能強化版では高速性も維持しており、128k長文処理など高度機能も含め社内検証を経て安定提供されています [12] [15]。

ソフトバンク：「Sarashina」（更科）シリーズ

- **モデル概要**：ソフトバンクグループのSB Intuitions社が開発した国産LLMです。**2024年度に推定4,600億パラメータの超大規模モデル「Sarashina」を完成**させたとされ、現在は約700億パラメータの実用モデル「Sarashina mini」の社内トライアルを進めています [18]。桁違いの巨大モデルを「教師AI」と位置づけ、その知見を小型モデルに継承・蒸留する戦略を採っています [19]。日本独自のAIエコシステム構築を目指した国家レベルの取り組みの一環であり、**日本語特化かつ高性能な基盤モデル**として開発されています [20]。
- **数学的推論能力**：初期のSarashinaモデルでは、日本語の一般質問応答で高い性能を示す一方、**数学の文章題やプログラミング問題の性能が十分でない**課題がありました [21]。これはWeb上の学習データに数学・コードのデータが相対的に少ないためと分析され、2025年には**数学・コーディングタスクのデータを抽出・増強して学習することで性能向上を図った新版「Sarashina2.2」**が開発されています [22]。その結果、小規模版(3B)でも旧版の70Bモデルを上回る**日本語数学タスク性能（MGSM-jaで62.4%）**を達成しており [23]、大規模モデル側でも数学推論力が強化されたと推測されます。現時点で具体的なベンチマークスコアは未公表ですが、**超巨大モデルの容量を活かして高度な数学推論にも対応し得るポテンシャル**を備えていると考えられます。
- **物理分野への対応**：4.6兆という桁違いのパラメータ数により、物理学を含む幅広い領域の知識を大量に保持していると期待されます。一般的な物理Q&Aや科学常識はもちろん、専門的な物理用語や概念についても学習している可能性が高いです。ただし、実際に**大学レベル・研究レベルの物理問題を解決できるか**は未知数です。社内ではモデルの知識を小型版へ蒸留する試みも行われており、将来的には物理計算やシミュレーション的な応用も視野に入れていると考えられます。現在のところ外部評価

はありませんが、**国内最大規模のモデルとして物理を含む科学技術分野の網羅性は極めて高い**でしょう。一方、精密な数値計算は外部ツール連携に頼る可能性があります。

- **ステップバイステップ推論の正確性**：Sarashinaでは**巨大モデルを「教師」、小型モデルを「生徒」**とする知識蒸留で効率向上を図っています [19]。このアプローチにより、大モデルが持つ高度な推論パターンを小型モデルにも継承させ、ステップバイステップの思考を再現させる狙いがあります。事実、社内検証では複雑な逐次推論課題にも対応できる手応えがあるようです [18]。段階的な理由説明や問題分割など、**チェイン・オブ・ソート（思考の連鎖）を活かした推論**をモデルに習得させる研究開発が進められています。もっとも、公開情報が限られるため細部は不明ですが、ChatGPT等に匹敵する推論正確性を目指していると言えます。
- **数式や論理記述の出力精度**：初期モデルではこの点が弱みとされてきました（Sarashina2-70Bでの数学文章題正解率56.4%、コード生成22.0%に留まる [24]）。しかし新版ではデータ強化によって大幅改善しており、**数学式を含む応答やコード出力の精度が向上**しています [23]。例えば、小規模版3Bモデルで日本語数学（MGSM-ja）62.4%・日本語コード(JHumanEval)39.6%に達しています [25]。大モデルではさらに高い精度で数式展開や論理的文章生成が可能とみられ、オープンなGPT-4系に迫る表現力も期待されています。ただし外部による検証データが不足しているため、実運用では慎重な検証が必要でしょう。
- **API・商用利用**：**2025年現在、Sarashinaモデルは社内トライアル段階**であり、一般には公開されていません [26]。ソフトバンクは今後このモデルを商用提供する計画で、国内の企業や官公庁向けに**「ソブリンクラウド/ソブリンAI」**構想のもとで提供していく意向を示しています [18] [27]。現時点でAPIは非公開ですが、将来的にはソフトバンク系のクラウドサービス上で提供されたり、自社サービス（通信ネットワーク最適化など）に組み込まれる可能性があります。また、同社は**マルチLLM戦略**を掲げており、用途に応じて他社LLMも含め使い分ける中でSarashinaを位置付けるとしています [28]。商用版提供時には大規模モデル版と軽量版を組み合わせ、高性能とコスト効率を両立する形でユーザー企業に提供される見込みです。

ELYZA：「ELYZA-LLM」シリーズ

- **モデル概要**：東京大学 松尾研発のスタートアップELYZAが開発する日本語特化LLM群です。2025年7月時点で**最高性能のモデル「ELYZA-Shortcut-1.0-Qwen-32B」**を公開しており、これは中国・アリババ社の公開モデルQwen2.5-32B-Instructを基に独自改良したものです [29]。同社はさらに、推論能力強化に特化した「ELYZA-Thinking-1.0-Qwen-32B」という**Reasoningモデル**も開発しており、そこで得られたデータを生かしてShortcutモデルの高速化を実現しています [29]。**日本語対話・タスク実行性能はグローバルLLMに匹敵するレベル**で、国内勢の中では最高水準とされています [30] [31]。金融業など大企業での導入実績もあり、安全性・カスタマイズ性に優れた企業向けLLMソリューションを提供中です [32] [33]。
- **数学的推論能力**：ELYZA-LLMの性能評価では、**数学問題集「MATH-500」およびその日本語版でOpenAI GPT-4（2024年版）に匹敵するスコア**を達成したと公表されています [34]。具体的な正解率は明示されていませんが、GPT-4相当であることから**大学入試レベルの数学問題でも高い正答率**が見込まれます。また、日本語版GSM8K相当の算数文章題にも対応可能です。推論過程を要する問題ではThinkingモデルを用いることで一段と精度が向上し、段階的に考えながら回答する能力に優れます。総じて、**国内モデル中では数学分野に強いモデルの一つ**と言えます。
- **物理知識・応用問題への対応**：ELYZA-LLMは広範な日本語データから学習しており、物理分野の知識もカバーしています。日本語版MMLUサブセットや自社タスク100種評価などでもグローバルモデル並みの性能を示したとされ [34]、一般的な物理QAや科学常識問題にも的確に答えられます。例えば「音の速度は？」「ニュートンの第一法則とは？」といった質問には正答できるでしょう。応用物理の問題については、Thinkingモデルでステップ解析させることで対応可能です。とはいえ、量子力学のような高度理論の新規問題を解くのは困難で、**既知知識の範囲内で回答を組み立てるレベル**です。数式処理も含め、ChatGPT-4相当の実力があるため物理定数の計算や簡単な実験数値の推算はこなせますが、未知の問題に対する創造的推論はトップモデルほどではありません。
- **ステップバイステップ推論の正確性**：ELYZAは**Thinkingモデルによって逐次的な思考能力を強化**しています。Thinking版は回答までの思考プロセス（いわゆるChain-of-Thought）を内部で展開しなが

ら、高度な推論を行う設定になっています。Shortcut版は高速応答重視ですが、その開発過程でThinking版の知見が活かされているため、通常のモデルでもある程度のステップ推論は可能です[29]。ユーザが「途中経過を詳しく説明して」と指示すれば、数ステップの論理展開を示すこともできます。特に**推論が必要なタスクでGPT-4相当のスコア**を出せていることから[34]、段階的推論の正確性は非常に高い水準です。複雑な数学パズルの解説や、長文文章の論理誤り検出なども得意としています。

- **数式や論理記述の出力精度**：ELYZA-LLMは**コード生成評価（JHumanEval）でも高スコア**を達成しており[34]、構文の厳密さが要求される出力に強みがあります。LaTeX形式の数式や論理式も整った形で出力可能で、フォーマット面の精度は良好です。たとえば「二次方程式の解の公式」を尋ねれば適切な数式で回答し、証明問題では箇条書きや(1)(2)...のように論理を整理して示すこともできます。もっとも、モデル自身に高度数学を解く創発的能力が備わっているわけではないため、**出力された数式が本当に正しいかの検証は必要**です。しかし、GPT-4に匹敵すると称するだけあって形式的な誤りは少なく、計算問題でも途中計算を書き出して正答に至るケースが多く見られます。
- **API・商用利用**：ELYZAは自社モデルを用いた**安全なAPIサービス提供を準備中**で、大企業との共同開発プロジェクトなども進めています[35]。実際に2023年頃から一部企業においてELYZAモデルの運用実績があり、金融機関での社内利用例も報告されています[32]。現在、ウェブ上でデモ版チャットが公開されており（登録者向け）[30]、順次API経由で企業利用できる計画です[35]。商用利用に際しては契約により提供され、ELYZA側でクラウド環境を用意した上で顧客に専用インスタンスを割り当てる形が想定されています。モデルそのものも一部は公開されており、たとえば**「Llama-3.1-ELYZA-JP-70B」や最新版「ELYZA-Shortcut-32B」**の情報が公開されています[36]。なお、ELYZAは医療特化モデル（ELYZA-LLM-Med）も開発しており[37]、特定分野向けにも積極的に展開しています。

Preferred Networks：「PLaMo」（プラモ）シリーズ

- **モデル概要**：Preferred Networks (PFN) が**フルスクラッチ開発**した国産LLMシリーズです。商用版フラッグシップとしてPLaMo 2.0 Prime（モデル規模非公開、1000億パラメータ級とも言われる）が提供されており、**日本経済新聞社の2025年優秀製品賞・最優秀賞**を受賞しています[38]。日本語に対する高い生成性能、長文32kコンテキスト対応、検索強調型生成（RAG）機能など、日本語実務で求められる精度と機能を安定して発揮できる点が評価されています[39]。費用対効果にも優れ、同等サイズの海外モデルより利用しやすいコストで必要十分な性能を提供することも特徴です[40]。さらに最新版PLaMo 2.2では指示追従性能が日英両言語で大きく向上し、複数ターンのロールプレイや医療QAなど**高度専門性や文脈理解が必要な場面で回答の一貫性・信頼性が向上**しています[41]。
- **数学的推論能力**：PLaMoシリーズは公式には数学専用のベンチマーク結果を公開していませんが、コードデータや多言語データも大規模に学習しているため、**基礎的な数学推論力は高い水準**にあると推察されます。特にPLaMo 3（開発中）では英語・日本語・コードをバランス良く学習しており、コード内のアルゴリズム理解などを通じて計算論的な推論力も強化されています[42][43]。実務シーンでは、文章中の数値を扱う財務計算や統計処理の質問などに安定して答えられるよう調整されているようです。もっとも、最先端の数学難問（数学五輪レベルの問題など）の解答は難しく、ここはGeminiやGPT-5のような超巨大モデルが依然優位でしょう。しかし日本語で出題される数学文章題に関しては、PLaMoもほぼ問題なく解答できるだけの能力を備えていると考えられます。
- **物理知識・応用問題への対応**：PLaMoは日本語・英語の大量のテキストを学習しており、日本の文化・制度から科学技術知識まで幅広くカバーしています[44]。そのため、物理法則や化学法則に関する知識問題には正答できる場合が多いです。例えば「光の速度はいくらか」「熱力学第二法則の内容は？」といった質問には正確に答えられるでしょう。応用的な物理問題については、モデル単体での推論には限界もありますが、**PFNは視覚と言語を統合したマルチモーダルモデル（PLaMo 2.1-8B-VLなど）も開発**しており[45][46]、画像データと組み合わせた異常検知や物体認識といった**フィジカルAI領域**への応用も見据えています[47]。これにより、将来的には物理実験データの解析やシミュレーション結果の説明など、人間の専門家を支援する用途にも使われていく可能性があります。現状のPLaMo Prime自体はテキスト専門ですが、物理の**公式や単位系の扱いなども含め日本語での専門QAに対応できる**よう独自データでチューニングされているとのことです[48]。

- **ステップバイステップ推論の正確性**：PFNは計算資源提供や国研(NICT)との協力で次世代モデル開発を進めており、PLaMo 3では**Sliding Window型の注意機構**を採用するなど長文推論時の効率と精度向上を図っています [49] [50]。最新版2.2では追加学習データによって**複数ターンにまたがる高度な質問応答でも一貫した回答**を示すようになっています [48]。例えば医療分野で患者の症状を複数聞き出しながら診断推論するようなケースでも、前後の文脈を踏まえた整合的な応答が可能です [48]。推論過程の明示という点では、ユーザが望む場合に理由を付けて説明することも可能ですが、Chain-of-Thoughtをそのまま出力する機能はデフォルトではありません。しかし**内部的な論理整合性は高められており**、JetBrains社のテストでも「Gemini2.5比で解決できるコーディング課題が50%以上増加した」ことが報告されています [51]。これはモデルの問題解決力（推論力）が着実に向上していることを示します。
- **数式や論理記述の出力精度**：PLaMoは社内プロダクトとして**翻訳特化モデルや金融特化モデル**も展開しており [52]、Markdownやコードを含む文書の出力にも対応しています [53]。そのため、数式を含むレポートや対話文中に論理記号が出てくる場合でも適切に処理できます。実際、政府の**ガバメントAI「源内」**プロジェクトではPLaMo翻訳モデルが採用されており [53]、PDFに含まれる数式や特殊記号を崩さず翻訳・要約するといったタスクも想定されています。コード生成についても、PFNの発表によれば100B規模モデルの運用経験から課題を克服してきており [54]、オープンなコードベンチ（HumanEval等）の成績も良好と推測されます。総じて、**フォーマット整形や表記の正確さに定評**があり、論理記述の体裁は整った出力が期待できます。ただし計算間違いそのものをモデルが検出・修正することは難しいため、最終的な結果検証は必要です。
- **API・商用利用**：**PLaMo 2.0 Primeは商用提供中**であり、PFNは2026年1月には新規利用者に100万トークン分のクレジットを付与するキャンペーンも実施しました [55] [56]。利用希望者はPFNのプラットフォームに登録することでAPI経由でモデルを利用できます。オンプレ版提供についての言及はありませんが、セキュリティ重視の顧客には日本国内データセンター経由でサービス提供するなどの対応も考えられます。実際に自治体や企業での利用が拡大しており [57]、**国産AIの選択肢として安心して使える**ことが売りです [57]。また、PFNはNICTとの共同で**次世代モデルPLaMo 3.0**を2026年春に公開予定としており [58]、これは研究コミュニティにも提供される見込みです。オープンソースへの姿勢もあり、一部の小規模モデル（2B, 8B, 31B）はHugging Face上で公開されています [59]。総じて、商用サービス（API利用やクラウド提供）と研究用途公開を両立する形で展開が進んでいます。

サイバーエージェント：「CyberAgentLM」シリーズ

- **モデル概要**：サイバーエージェントが自社の日本語データを活かして開発したLLMです。2023年5月に130億パラメータ規模の独自日本語LLMを公開し、その後2023年11月に68億パラメータ版、2024年7月には**225億パラメータの最新モデル「CyberAgentLM3-22B-Chat」**を一般公開しました [60]。**既存の外部モデルに頼らずスクラッチで開発**されたモデルで、LLM日本語評価指標でMeta社Llama-3 70Bモデルと同等性能を達成しています [61]。このモデルは商用利用も可能なライセンスで公開されており、国内企業による生成AI活用促進を目指しています [60]。また、画像や音声も扱える対話AI（VLM版）など視覚入力対応の派生モデルも研究されています [62]。
- **数学的推論能力**：CyberAgentLM3（22B版）は、日本語総合能力で大規模モデル(70B)に匹敵することから、基本的な数学文章題や算術にも対応できると考えられます [61]。実際、同社公開のリーダーボード結果では、中学生程度の数学問題で高い正解率を示しているとのことです（詳細数値は非公開）。しかし、モデル規模が海外トップモデル（数百億〜数兆パラメータ）より小さいため、**高度な数学推論（例：複雑な証明問題や難関大学入試問題）では性能限界**があります。精度向上にはさらなるモデル大型化や専門データでの再学習が必要でしょう。一方で日本語の細かいニュアンスを理解するため、**問題文の意味取り違えが少なく安定した回答**を出す傾向にあります。簡単な数列問題や単位換算などは正確にこなせます。
- **物理知識・応用問題への対応**：CyberAgentLMは日本語特化とはいえ、Webデータ中の科学記事やWikipediaも学習しているため、**物理学の基礎知識は十分備えています**。力学・電磁気・化学など高校レベルの質問に対して概ね正しい説明が可能です。ただ、応用的な物理問題（例えば初見の複雑な力学計算や電気回路設計問題）となると、モデル単独での解答は難しくなります。出力された回答も概念説明が中心で、定量的な計算結果を要する場合には誤差が出ることがあります。総じて**物理に関**

する一般的なQ&Aには強いが、難易度の高い問題解決は不得意という評価です。もっとも、ユーザ自身が誘導する形でヒントを与えれば、それを反映して答えを更新する対話能力はあります。これは日本語運用能力の高さによるもので、専門用語を正しく理解・回答に組み込める点は評価できます。

- **ステップバイステップ推論の正確性**：CyberAgentLM3はチャット指向に調整されており、ユーザから「手順を詳しく教えて」と頼まれれば、**考えられる手順を順序立てて回答**します。例えば「ある料理のレシピ手順を箇条書きで説明して」といった依頼には適切に応じられます。論理パズルなどでも、自身で問題文を分解して考える程度の推論は可能です。ただしモデルの規模上、GeminiやGPT-5のような**高度な思考チェーン**は実装されておらず、複雑な推論になると推論飛躍や見当違いな回答が増えます。言い換えると、**日常的な範囲の段階的推論は正確だが、長大な推論チェーンは苦手**です。それでもver.3の22Bモデルではver.1（13B）やver.2（6.8B）に比べ改善が見られ、最新の微調整によって回答の一貫性・安定性が向上しています。例えば、ある質問に対し途中で矛盾した説明に陥るケースが減り、ユーザーの追加質問にも整合的に答え続ける傾向があります。
- **数式や論理記述の出力精度**：サイバーエージェントのモデルは**オープンソースコミュニティでも試用が進んでおり**、実際にローカル環境で動かして検証したユーザ報告があります [63]。それによれば、**基本的な数式やプログラムコードの出力も可能**で、たとえばPythonの簡単な関数を書く程度なら正確に行えるとのことです。数式についても、「E=mc^2」を含む文や簡単な数式展開をそのままテキスト生成することができます。ただ、高度な論理記述（証明全体の記述など）になると、論理の飛躍や書式の乱れが散見されます。コード出力に関しても、長めのコードでは抜け漏れが起こりがちです [64]。それでも**手軽に使えるローカルLLMとしては極めて優秀な部類**であり、68億パラメータ版でも商用利用可能なライセンスで公開されていることから、ユーザ自身で微調整して専門フォーマットの出力精度を高める余地があります。
- **API・商用利用**：CyberAgentLMは**モデルデータそのものが公開**されており、GitHubやHugging Face経由でダウンロードして利用可能です [60]。ライセンスも商用利用を許諾しているため、企業が自社サービスに組み込むこともできます [60]。実際、同社は広告事業などで自社LLMを活用し始めており、社内システムでの対話AIやクリエイティブ生成支援に応用しています。また外部向けにはオウンドメディア上でAPIエンドポイントや利用方法を紹介しており、開発者コミュニティへの普及も図っています。「誰でも使える日本語LLM」を標榜しており、**国内では数少ない、商用利用可能なオープンモデル**として注目されています [60]。今後もパラメータ数の大きいv4モデルの開発や、視覚/音声を扱うマルチモーダル対応のLLM研究も進める意向が示されています [62]。

海外先端LLM（Gemini 3 Pro・ChatGPT-5.2・Claude Opus 4.5）との比較

Google Gemini 3 Pro

- **モデル概要**：GeminiはGoogle DeepMindが開発する次世代マルチモーダルLLMです。2025年11月にリリースされた**Gemini 3.0 Pro**は、テキストだけでなく画像・音声・コードなど複数のモーダルを統合した最上位モデルとなっています [65] [66]。Gemini 2.5 Proから飛躍的に性能が向上しており、**推論深度・信頼性が大幅に強化**されています [51]。長大なコンテキストでの動作やツール使用（コード生成やAPI呼び出し）にも長け、開発者向けツールへの組み込みが進んでいます [67]。マルチモーダル能力もトップクラスで、画像や動画を解析してテキスト出力したり、逆にテキストからインタラクティブなグラフやUIを生成することも可能です [68] [69]。Gemini 3 ProはGoogle CloudのDuet AIや検索エンジン・生産性ツール群に統合されつつあり、広範なサービスで利用可能となっています。
- **数学的推論能力**：Gemini 3 Proは**高度な数理推論で突出した性能**を示しています。専門家が作成した最難関の数学ベンチマーク「FrontierMath」ではTier 1-3問題で**正解率38%**を記録し、従来モデルを大きく引き離しました [70]。ChatGPT-5.1世代では到達できなかったレベルの**思考の連鎖（チェインオブソート）**を獲得しており、複雑な数学パズルや未解決問題にも一定の筋道だった解答を出せるようになったと評されています [71]。例えば難解な確率問題や幾何問題でも、一歩一歩推論を進めていき正答に近い答えを導くケースが増えています。Gemini 3の数学性能向上は従来比「23倍の圧倒的進歩」とも言われ [72]、**ChatGPT-5シリーズすら追いつけない推論力**を獲得したとの指摘もあります。

競技数学的な問題解決や自動定理証明に強みを発揮しており、ツールを併用せずとも驚異的な正答率を達成しています [73]。

- **物理的知識・応用問題への対応**：Gemini 3 Proは科学分野にも圧倒的です。AI研究者らによる**最先端物理学テスト「CritPt」**では、Gemini 3 Proが**スコア9%でトップ**となり、Claude 4.5 Opusの5%やGPT-5.1と比肩する性能を示しました [74]。CritPtは研究者のリサーチアシスタント能力を測る難関ベンチマークであり、この高スコアは**未知の物理問題にも一定の洞察を示せる**ことを意味します [74]。また、自身で方程式を立て数値計算する能力も飛躍的に伸びており、例えば**MathArena**という物理数学混合チャレンジでは他モデルが1%程度の正解率しか出せない問題群でGemini 3 Proは23%に達したとの報告があります [75] [71]。これは通常人間でも解けないレベルの問題に部分的とはいえ食らいついていることを示します。総合的に、Gemini 3 Proは**物理法則の深い理解とそれを応用した問題解決において現行トップクラス**です。理論物理の難問へのチャレンジや、未知の現象を説明する仮説立案補助など、人間研究者の創造的作業を助けることが期待されています。
- **ステップバイステップ推論の正確性**：Gemini 3 Proの大きな強みは**計画立案や段階的推論の飛躍的向上**です。従来モデルでは難しかった長い思考チェーンをGeminiは安定して維持でき、途中で方針がぶれることが少なくなりました [71]。例えば複数の制約条件を含むパズルを解く際、条件をひとつずつ整理して検討し、最終解を導くプロセスを適切に踏むことができます。Google内部テストでも、Gemini 3 Proは2.5 Proに比べ**50%以上多くのベンチマーク課題を解決**したとの結果があり [51]、これは難問でのステップ推論力が格段に増したことを裏付けます。さらにGeminiは**ツール使用を組み合わせたエージェント的推論**にも優れています [76]。例えば外部の電卓やデータベースを自律的に呼び出しながら問題を解決するような場面でも、適切にタスクを分割し順序立てて実行できます。総じて、**推論過程の一貫性・正確性では現状最も信頼できるモデル**と評価されています。
- **数式や論理記述の出力精度**：Gemini 3 Proは**数式やプログラムコードの生成も極めて高品質**です。コード分野では、Gemini 2.5 Pro比でフロントエンド開発の品質が向上したという開発者の証言があり、VSCode上でのテストでは35%高い精度でソフトウェア課題を解決したとのことです [77]。数式についても、美しいテキストフォーマットで複雑な式を出力できます。例えばユーザーが「与えられたデータに対する回帰分析の結果をプロットして」と要求すれば、GeminiはPythonコードを生成しつつ対話型のグラフをレンダリングすることも可能です [68]。**推論結果を図表やダッシュボードとして提示**できる点は他モデルにない特徴です [68]。また、RNA転写のような複雑トピックも視覚的に解説するなど [78]、テキストと図を組み合わせた高度なアウトプットが可能です。論理記述では、証明問題に対して適切な形式で解答をまとめあげます。Gemini 3 Proの出力する証明は厳密かつ簡潔で、従来モデルにありがちだった論理の飛躍がほとんど見られません。**数式処理・論理表現の面でもGemini 3 Proは群を抜く精度**を誇ります。
- **APIや利用形態**：Gemini 3 ProはGoogleの各種サービスを通じて利用可能です。一般開発者向けにはGoogle CloudのAI API（PaLM APIの後継）で提供されており、コードエディタ「Studio Bot」や対話AI「Bard」の高機能モードにも統合されています。企業向けにはDuet AI（Google Workspace向けAI機能）や、提携各社のアプリケーション（例：Box社のコンテンツAI、ReplitのAIコーディング支援等）にも**Geminiの能力がバックエンド実装**されています [76] [79]。つまりユーザは直接モデルにアクセスしなくとも、Googleの製品群を介してGemini 3の先進機能を利用できる形です。現時点でモデルそのもののオープン提供はなく、利用はGoogleのクラウド経由となります。また、一部機能（例えば画像入力を含むマルチモーダル対話）はPro版専用で、Googleの有料サービスに加入する必要があります。Gemini 3は**世界的に見ても最高峰の性能**であり、その利用は基本的に商業プラットフォーム上でコントロールされています。

OpenAI ChatGPT-5.2 (GPT-5.2 系列)

- **モデル概要**：ChatGPT-5.2はOpenAIのGPT-5世代に属する大規模言語モデルで、2025年末時点の最新版です。GPT-4からの中間アップデート（GPT-4.5等）を経て、大幅な性能向上と新機能追加が行われています。特に**「GPT-5.2 Thinking」モード**では内部でチェインオブソートを用いて推論する設定になっており、複雑な課題解決に卓越した能力を発揮します [80]。また**「GPT-5.2 Pro」モード**では応答速度と創造性が強化され、ビジネス文書作成やビジュアルコンテンツ生成（表・グラフやMarkdownでのPPT生成など）も可能となりました [81]。コンテキスト長は標準で256kと非常に長く、OpenAI社

内の評価では256kの長文質問応答で**ほぼ100%に近い正確性**を示したとのことです [82]。これは数百ページに及ぶドキュメント内から正確に情報を見つけ出す能力を意味します。さらに画像理解や音声入力にも対応し、ツール呼び出しも高度に統合された**汎用AIエージェント**として完成度が高まっています。

- **数学的推論能力**：ChatGPT-5.2は、前世代まで苦手だった数学分野で**飛躍的進歩**を遂げました。競技数学レベルの難問にも取り組めるようになり、**専門家級数学ベンチマーク「FrontierMath」(Tier1-3)で40.3%という過去最高の解答率**を達成しています [83]。これは直前にトップだったGemini 3 Proの約38%を上回り、事実上**数学問題解決で世界最先端**に立ったことを示します [83] [70]。また大学院レベルの数学試験でも高得点を記録し、研究者から「人間の優秀な大学生に匹敵する」と評されました [73]。GPT-5.2では**推論途中に計算が必要な場合、自律的にPythonツールを呼び出して計算**する機能があり、複雑な数値計算もミス無くこなします [84] [85]。例えば多重積分や大型行列の固有値計算といった手計算困難な処理もツールを併用して正確に行い、最終回答に反映できます。**自動定理証明**の分野でも部分的な成功例が報告されており、GPT-5.2 Proが統計的学習理論の未解決問題に対し提案した証明が査読を通過したという驚くべき事例も公開されました [86]。以上のように、ChatGPT-5.2は**数学に関しては人間専門家に肉薄、一部領域では凌駕する性能**を獲得しています [83]。
- **物理的知識・応用問題への対応**：OpenAIはGPT-5シリーズで科学研究支援を重視しており、「世界で最も科学者の助けになるモデル」と評価しています [87]。大学院生レベルの科学質問集「GPQA Diamond」では**GPT-5.2 Proが93.2%**という極めて高い正答率を記録し [87]、高度な物理・化学・生物の知識を正確に使いこなせることを示しました。例えば量子力学の概念や最先端の宇宙論について尋ねても、正確かつ詳細な解説を返します。応用問題への対応も優れており、物理の文章題で必要なら中間式を書き出して解を導きますし、工学的な最適化問題にもPythonツールで計算しながら答えることができます [88]。ただし、Geminiが得意とする創発的な仮説提案や未解明領域での推測では、GPT-5.2はやや保守的な傾向があります。CritPt物理ベンチマークでもGPT-5.1世代でGeminiに劣ったとの報告があり [74]、GPT-5.2になってほぼ肩を並べましたが依然トップではありません。それでも**既知の科学知識の網羅性と組み合わせの巧みさ**は群を抜いており、研究論文の要約や実験データの分析など実務的タスクでは最有力と言えます。実例として、GPT-5.2 Proは3時間に及ぶ多言語の会議音声から参加者ごとの発言を区別して完全書き起こしするデモを成功させています [89]。こうした高度な音声・画像統合もできることから、科学技術分野のあらゆる資料を横断して知見を引き出す能力に長けています。
- **ステップバイステップ推論の正確性**：ChatGPT-5.2は**Thinkingモードで顕著なステップ推論力**を発揮します。内部で数千トークンにわたる思考展開を行いながら回答するため、ユーザへの最終応答は洗練されていますが、その裏では綿密な推論が行われています。OpenAIの社内テストでは、GPT-5.2 Thinkingは**256kの長文コンテキスト上でも高精度な「針探し問題 (MRCRv2)」**を解くことができ、これは膨大な情報から特定の答えを見つけ出すタスクで4ステッププロンプト（4-needle）版においてほぼ100%の正確さだったと報告されています [82]。このように、必要に応じて多段の推論を行ってもエラーを蓄積しない強力な推論安定性があります。また、**ツール呼び出しを絡めた複雑なワークフロー**でも、Tau 2ベンチ（電話応対シナリオ）で98.7%、小売シナリオで82%という極めて高い成功率を示しています [90]。これは各ステップで適切なAPIやDBに問い合わせを行い、結果を解釈して次の行動を決めるという一連のマルチステップ推論を高精度でこなせることを意味します。総じて、**ChatGPT-5.2は長大な思考プロセスや複雑なタスクフローでも破綻しない信頼性**を備えており、ステップバイステップ推論の分野でGeminiに匹敵するトップランナーです。
- **数式や論理記述の出力精度**：GPT-4の時点でも定評があった数式出力やコード整形の精度が、5.2ではさらに向上しています。**競技プログラミング評価（SWE-Bench Verified）では80%、難度の高いSWE-Bench Proでも55.6%という史上最高スコア**を叩き出しました [91]。Pythonのみならず JavaScriptやGo言語にも対応しており、工業的なコーディング課題に強くなっています [92]。前世代に比べ、特にフロントエンド開発やUI生成のコード品質が向上しており、3D要素を含むウェブページのコーディングなどもこなします [93]。数式表現については、LaTeXで複雑な積分記号や行列を書いてもまずミスがありません。科学論文中の図表読み取りもでき、グラフの数値を解析してテキストで報告するといったことも可能です [94]。さらに、**5.2 Thinkingモードでは画像中の数式や表を読み取って解釈**することすらできます [84]。論理記述では、例えば法律文書を論理構造ごと要約したり、文章から論理矛盾を指摘したりといった高度なタスクも高精度です。幻覚（事実誤り）発生率もGPT-5.1の

8.8%から6.2%へ低減しており [95]、出力内容の信頼性がさらに増しています。要約すれば、**ChatGPT-5.2は数式・コード・論理の全てにおいて極めて正確かつ洗練された出力**を行えるモデルとなっています。

- **API・商用利用**：ChatGPT-5.2はOpenAIの有料プランで利用可能です。ChatGPT Plus/Pro/Enterpriseユーザーは、UI上でGPT-5.2（ThinkingまたはPro）を選択して使うことができます [81]。生成に時間を要する高度な機能（表やPPT生成など）もバックグラウンドで数分かけて処理され提供されます [81]。開発者向けにはGPT-5.2 APIが提供されており、トークン課金制でモデルの機能を自社アプリに組み込めます。コンテキスト長256k対応のAPIはビジネス契約が必要な場合もありますが、100k程度までは通常APIで扱えるようになっています。OpenAIはモデルの詳細を非公開としつつも、多数の企業と提携しており、MicrosoftのCopilotサービス群（OfficeやGitHub等）にもGPT-5.2が組み込まれています。モデルの重みは非公開・独占利用ですが、その分**徹底した安全対策とサポート**が付随します。企業ユーザには専用インスタンスでの提供も行われ、機密データを扱う用途にも対応しています。総じて、ChatGPT-5.2は**広範な商用エコシステムに組み込まれ、最先端機能をサービス経由で提供する形**となっています。

Anthropic Claude Opus 4.5

- **モデル概要**：Claude Opus 4.5はAnthropic社のClaudeシリーズにおける最新強化版モデルです（2025年時点）。Claude 2をベースにさらなる性能向上を図ったもので、特に**コーディング能力とエージェント機能で世界最高レベル**と謳われています [96]。コンテキスト長は約200kと非常に長く、対話形式のインストラクションにも忠実に応答します。Claude Opusは「憲法AI」として倫理指向の調整が特徴でしたが、4.5では安全性を維持しつつ**視覚・推論・数学すべてで前世代より大きな強化**がなされています [97]。特に**ソフトウェア開発支援（コーディング）では依然トップ**であり、他モデルに対する優位性を保っています [98]。価格改定も行われ、従来モデルと同等価格で利用できるようになったため、コスト面の強みも出てきました [99]。
- **数学的推論能力**：Claude Opus 4.5は総合的に見るとGPT-5.2やGemini 3 Proに次ぐ性能と評価されます。数学ベンチマーク「MATH (500題)」では**85%**もの高得点をマークし（Claude 2比大幅向上）、代数・幾何・解析など幅広い数学領域で強力です [100]。これはGPT-4の約2倍近い正解率であり、Opus 4.5が**人間の大学生を上回る数学力**を身につけたことを示唆します。また、数学コンテストレベルの難問集であるAIME問題でも37%を正答しています [100]。GeminiやGPTがChain-of-Thoughtで一歩リードしていますが、Claudeも自己修復型の推論を行える特徴があります [64]。一度間違えても、それを指摘すると自身の誤りに気付き解答を修正できる柔軟性が報告されています [64]。総じて、**Claude 4.5はトップモデルに僅差で迫る数学性能**を持ち、とりわけユーザフィードバックを活かした見直しに強みがあります。
- **物理的知識・応用問題への対応**：Claude Opus 4.5は知識面では膨大なテキストから学習しており、物理・化学・生物の知識は極めて豊富です。科学QAであるMMLUベンチマークでも上位の成績を収めています。難解な物理問題への挑戦ではGeminiに一歩譲りますが、**CritPt物理評価ではGeminiに次ぐスコア（5%）**を記録し [74]、最先端モデルの一角として健闘しています。これは極めて難しい前沿研究の問題集での結果であり、Claudeも一定の推論力を発揮したことになります。また、Claudeは**文章の文脈保持や推論一貫性が高い**ため、例えば長い物理の設問（実験の状況説明→問に答える）にも粘り強く付き合い、正答に近づけます。OpenAIやDeepMindのような外部ツール統合機能は弱めですが、代わりに**ユーザの追加質問や訂正に対する応答が丁寧**で、対話を通じて物理問題を解決するのに向いています。例えばユーザがヒントを与えれば、それを取り入れて答えを改善していくインタラクティブなやり取りが可能です。したがって、**Claude 4.5は知識量で他に劣らず、対話的な問題解決で力を発揮するモデル**と位置付けられます。
- **ステップバイステップ推論の正確性**：Claudeシリーズは一貫して長文での理由説明が得意でしたが、Opus 4.5ではそれがさらに磨かれています。例えばある問題に対し、一度で正答に至らなくとも、ユーザが「本当にそれで良いか考え直して」と促すと、モデル自身が推論ログを再評価して誤りを検知し、修正を加えた解答を提示する、といった**内省的プロセス**が見られます [64]。この挙動はAnthropicの目指す憲法AIアプローチの成果とも言え、**自己訂正しながら推論精度を高める**点でユニークです。もっとも、最初から完璧な推論チェーンを組み立てる能力ではGeminiやGPT-5に及ばな

いという指摘があります [71] 。特に、極めて複雑なパズルで一発回答を求める場合、Claudeはやや試行錯誤が多くなる傾向です。しかし日常的タスク（例えば文章要約→要約に基づく分析→結果報告の3段階処理等）では高精度で、**実用上問題ないレベルの逐次処理性能**を持っています。また、会話調整も上手く、長いやり取りの中でも質問の意図を見失わず回答し続ける安定性があります。要約すると、**Claude Opus 4.5は若干のヒューマンループを許容すれば最終的に非常に正確な推論結果を導けるモデル**と言えます。

- **数式や論理記述の出力精度**：Claude 4.5は**ソフトウェアエンジニアリング領域で依然トップ**との評価があり、コード生成では他モデルを凌駕しています [98] 。実際、一部のベンチマークではGPT-5.2やGemini 3を上回るスコアも出ています [101] 。例えば40B規模のOSSモデルが81.4%を出したSWE-Bench Verifiedで、Opus 4.5やGPT-5.2は80%前後だったとされ、**ほぼ肩を並べる最高水準**です [102] 。数式出力に関しても、ClaudeはLaTeX形式での回答が可能であり、MATHベンチ85%という数字がそれを裏付けます [100] 。論理記述では、Claudeは長文解答を丁寧に段落構成する傾向があります。例えば法的な判断を求める質問では、箇条書きや章立てで論点を整理しつつ結論を述べるため、可読性の高い出力となります。ただ、一部で指摘されるのは「**複雑なタスクでの見落とし**」です [64] 。例えば長いコードを書かせる際、要件の一部を漏らすケースがあります。しかしこれは他モデルのレビューを受けた際に修正できるレベルであり（人または別AIが指摘すると直せる）、Claude自身も指摘を受けての修正に前向きです [64] 。幻覚出力についてはAnthropicのポリシーで抑制されていますが、完全ではなく、最新報告では依然GPT-4相当の幻覚率があるとのことです。ただし**対話を重ねることで最終的な回答の正確性を高める設計**のため、ファイナルアンサー時点ではかなり質の高い回答に到達します。まとめると、**Claude Opus 4.5は形式張った出力やコード生成に強く、必要に応じて修正も効く柔軟なモデル**と言えます。
- **API・商用利用**：AnthropicはClaude 4.5を**低価格でAPI提供**しており、以前のClaude 2に比べ大幅なコストダウンが発表されました [98] 。これにより、開発者や企業はより安価に最強クラスのモデルを利用できます。Claude APIは最大100k超のコンテキストを扱え、用途に応じてOpus（高性能版）とSonnet（廉価高速版）の2種類が選べます。Anthropic自体は安全性を重視した設計から企業ユーザの支持があり、SlackのAIアシスタントやQuoraの対話ボットなどにClaudeが組み込まれています。また、2025年末にはAnthropicがOpenAIと戦略提携を結び、OpenAIプラットフォーム上でClaudeモデルも利用可能になるという報道もありました（※実現すればChatGPTからClaudeを呼び出すことも可能になる見込みです）。現状、モデル重みは非公開ですが、一部学術機関との共同研究で提供されるケースがあります。総じて**Claude Opus 4.5はAPI経由で広く商用利用可能**であり、特にコード生成やカスタマーサポート分野での採用が進んでいます。

主要モデルの比較まとめ

以上より、日本の主要LLM6種（NTT・NEC・ソフトバンク・ELYZA・PFN・サイバーエージェント）と海外先端LLM3種（Gemini 3 Pro・ChatGPT-5.2・Claude Opus 4.5）の数学・物理推論能力等を比較すると以下のようになります。

モデル（提供企業）	数学推論能力	物理知識・応用問題	ステップバイステップ推論	数式/論理記述出力	API提供・商用利用
tsuzumi 2（NTT）	日本語での基本計算・文章題には対応。高度な数学・定理証明は苦手で、**ChatGPT-5に及ばない** [3]。	一般的な物理QAは可。応用的・専門的な物理問題は精度低め（ChatGPT並みの巨モデルとの差あり）[4]。	段階的推論は可能だが、**長い論証では誤り増**。基本的な理由説明・手順解説は良好。	シンプルな数式は出力可。**複雑な数式変形や証明は不正確さ**が残る。論理展開に飛躍が出る場合あり。	**法人向け提供中**（NTTと契約してオンプレ/プライベートクラウドで利用）[2]。一般APIは未公開。
cotomi（NEC）	日本語読解・知識問答で**世界トップ級** [9]。数学特化ではないが、長文問題も128K文脈で解ける可能性。	科学知識は豊富。**長文コンテキスト活用で技術文書理解**◎。物理計算自体は不得意だが外部ツール連携で補完。	**タスク分解・ツール選択による高度推論**が可能 [14]。Agenticな自律実行で複雑手順も処理。	長文中の数式も保持し出力可。フォーマットも整うが、**厳密性は要検証**。論理的には一貫性高め。	**企業向けソリューション内提供**。128K対応モデルを社内適用。公開APIなし（MCP準拠で他サービスと連携）[15]。
Sarashina（SB）	4.6兆パラメータの超巨大モデルで潜在力大。**データ強化により数学性能改善** [103]。一部ベンチでGPT級との噂。	膨大な知識を保持し物理も網羅。未知問題は要検証だが、**科学知識網羅性は非常に高い**。数値計算はツール併用か。	**教師AI→生徒AIの知識蒸留**で推論力向上狙う。複雑推論も可能性高いが詳細不明（商用化準備中）[18]。	数式出力精度は当初低めだったが改善中 [24]。コードや論理の一貫性も強化されつつある。	**社内トライアル中**。商用提供予定 [26]。一般公開なし。将来はソブリンクラウド経由提供見込み。

モデル（提供企業）	数学推論能力	物理知識・応用問題	ステップバイステップ推論	数式/論理記述出力	API提供・商用利用
ELYZA-LLM（ELYZA）	**MATHベンチでGPT-4相当** [34] と高水準。数学問題も高正解率で日本トップ級。	一般常識・科学QAはGPT-4並み [34]。物理も基本網羅。応用問題もThinkingモデルでかなり対応可。	**Thinkingモデル**でチェインオブソート駆動し高精度推論。通常モードでもGPT級の推論力。	**数式・コード出力精度は極めて高い**（GPT-4並） [34]。論理的に整理された回答を出力。	**安全なAPI提供を準備中** [35]。一部機能はデモ公開。企業と共同実証多数、商用利用も開始。
PLaMo Prime（PFN）	公称スコア非公開も、日本語数学問題は概ね良好に解決。コード含む学習で**推論力高**。トップモデルには一歩譲るか。	広範な科学知識を学習済み。**官公庁翻訳AIにも採用**され信頼性高 [53]。物理現象の説明など安定。	**長文32k対応**かつ追加学習で**多ターン専門QAも一貫** [48]。推論安定性高く実務利用に耐える。	コード生成・専門フォーマット対応良し。**出力形式整然**。計算誤差などは残るが、Markdownや数式も崩さず処理。	**商用API提供中**（利用クレジット制度あり） [56]。オンプレ提供は未明。小モデルはOSS公開で研究利用可。
CyberAgentLM3 (22B, サイバーエージェント)	**国内オープンLLM最大級**。中級程度の数学問題は正確。超難問は不得手だが、日常算数は安定解答。	物理含む百科事典的知識は広い。高校レベルまでなら概ね対応。高度応用は苦手だが、一般説明は流暢。	**日本語対話調整良好**で手順説明は上手。だが長い推論は規模相応に限界。複雑推論では誤答増加。	**基本的な数式・コード出力可**。フォーマットも比較的整う。長い論証や大規模コードでは抜け漏れあり。	**モデルを一般公開** [60]（商用利用OK）。誰でもダウンロード・ローカル実行可。APIは無し（自前で環境構築）。

モデル（提供企業）	数学推論能力	物理知識・応用問題	ステップバイステップ推論	数式/論理記述出力	API提供・商用利用
Gemini 3 Pro (Google DeepMind)	**数学推論で世界最高水準**。FrontierMath難問で38%正解 [70] 。高度な定理証明や難問解決に卓越。	**科学全般に極めて強い**。物理の最難関評価でもトップ級 [74] 。未知の物理仮説への洞察力も示す。	**思考連鎖を的確に保持**し超複雑な推論も正確 [71] 。長大タスクやツール併用エージェントでも抜群の安定性。	**数式・グラフなどリッチ出力**が可能 [68] 。コード生成精度も飛躍的向上（2.5比+50%課題解決） [51] 。	**Googleクラウド経由で提供**。一般公開無し。BardやDuet AI等サービスに統合。商用ライセンスはGoogle管理下。
ChatGPT-5.2 (OpenAI)	**数学難問で記録更新**（FrontierMath40.3%） [83] 。専門分野含め人間専門家級の解答力。	**科学QAで最高精度**（大学院レベル93%） [87] 。物理計算もPythonツール併用で正確。研究支援に最適。	**256k長文でも破綻しない推論力** [82] 。マルチステップ対話・ツール操作も98%超の成功率 [90] 。	**数式・コードの精密出力**に長け80%超のコーディング正答率 [91] 。誤情報もGPT-5.1比で大幅減 [95] 。	**OpenAIのAPI/サービスで提供**（Plus/Enterpriseで利用可） [81] 。モデル重み非公開。多数のアプリに組込み展開中。
Claude Opus 4.5 (Anthropic)	**MATHベンチ85%** [100] と非常に高精度。Gemini/GPTに次ぐ実力。誤りを指摘すれば自己修正も可能 [64] 。	**科学知識は豊富**でMMLU上位。物理の前線課題でも善戦 [74] 。対話的に問題を深堀りし解決に導くのが得意。	**長文説明・内省に強み**。一度の推論ではミスも、人のフィードバックで改善 [64] 。通常対話では一貫性高い。	**コーディング世界一**との評価 [98] 。長文コードも高品質。数式も正確だが、ごく稀に要件見落としあり。	**API提供**（低価格） [98] 。商用利用広範。Slack等に組込み済。重み非公開だが提携通じ利用可能性拡大。

※表中の評価は各モデルの公開情報【7】【12】【18】【21】【29】【30】【31】に基づく相対比較です（2025年末時点）。日本のモデルはいずれも**日本語能力に優れる一方、超高度な数学・物理推論では最大規模の海外モデルに一歩譲る**傾向があります。しかし、用途次第では十分実用に耐える性能を示しており、特に**ELYZAやPFNのモデルはGPT-4クラスの実力**を発揮しています。一方、GeminiやChatGPT-5.2は桁違いの推論

力で、**人間専門家レベルの問題解決**が見えてきています [73]。Claude 4.5も含め海外先端モデルは数学・物理のみならず**総合知性の面で飛躍的進歩**を遂げており、国内勢も今後これらに対抗すべくモデル規模拡大やデータ特化で追随していくものと思われます。

日本国内のLLMは、国産ならではの**データ主権やセキュリティ面のメリット**があり [5]、企業が扱う専門文書の理解や日本語ならではのニュアンス理解では強みを発揮します [9]。一方、最先端のChatGPTやGeminiは**科学技術計算から創造的課題までこなす汎用性**を備えており、用途によって使い分けるのが現実的です。例えば、「高度な数学モデル構築にはChatGPT-5.2をAPI利用し、自社データを扱う日本語対話システムはNTTやPFNのLLMをオンプレ運用する」といったハイブリッド活用も考えられます。

今後、国内LLM各社もモデルの大型化やマルチモーダル対応を進めることで、数学・物理推論能力のさらなる向上が期待されます。また、**研究開発段階の超大規模モデル（例：ソフトバンクのSarashina　4.6Tなど）**が実用化すれば、海外モデルに匹敵する性能を純国産で達成する可能性もあります。その際には、日本語環境で高度な科学技術計算や専門課題を自己完結的に処理できるようになり、国内のDX・研究開発を強力に後押しするでしょう。

まとめとして、2025年以降の最新LLMにおける数学・物理推論対応状況は以下の通りです：

- **国内LLM（NTT・NEC・SB・ELYZA・PFN・CA）**は、日本語テキストの理解・要約・対話に非常に優れており、ビジネス文書や教科書レベルの数学・物理問題には十分対応可能。ただし、世界最先端の難問を解く推論力では、より巨大なパラメータを持つ海外モデルがリードしている。 [3] [70]

- **Gemini 3 Pro・ChatGPT-5.2・Claude Opus 4.5**といった海外モデルは、**数学定理の証明や未解決問題への挑戦、物理理論の深い理解**といった領域で突出した性能を示している [73] [83]。特にChatGPT-5.2は科学研究支援で卓越し [87]、Gemini 3 Proはマルチモーダルな推論出力で新次元を開いています [68]。

- **ステップバイステップの推論**では、国内勢もエージェント機能やReasoningモデルで対応を強化中だが [14]、**長い思考の一貫性ではGeminiやGPT-5系が依然トップ**です [71]。もっとも、国内モデルも実用上問題ない範囲での逐次推論は可能であり、特定ドメイン知識と組み合わせることで高い有用性を発揮します。

- **数式や論理的な記述の精度**について、ELYZAやPFNのモデルは日本語環境向けに最適化されており既に高精度な出力が可能です [34]。海外モデルはさらに上を行く精度で、コード生成やフォーマット整形ではClaudeやGPT-5.2が突出しています [100] [91]。利用シナリオに応じて、必要な精度とコストを見極めモデルを選定することが重要でしょう。

- **API提供と制約**に関しては、国産LLMの多くが企業向けソリューション内で提供され、オープンに使えるAPIは限定的です（CyberAgentLM等一部除く） [60]。一方、海外モデルは商用APIが整備されていますが、その使用には個人情報や機密データの取り扱い制約・利用料金など考慮事項があります。日本企業にとっては、**機密データは国内LLMで処理し、汎用タスクは精度重視で海外LLMを使う**といった使い分けも選択肢となります。

以上のように、2025年以降のLLMは国内外それぞれ長所を持っており、**日本語環境や用途適合性を踏まえてハイブリッドに活用すること**が望ましい状況です [104]。今後もモデル間競争により性能向上が続くため、最新動向を追い適切なモデルを選択していくことが重要です。

参考文献・情報源：（各モデルの性能・特徴に言及した公式発表やニュース記事より） - NTT「tsuzumi 2」提供開始ニュースリリース（2025年10月） [5] [3] - NEC公式発表「cotomi」性能強化プレスリリース（2025年7月） [14] [15] - SB Intuitions技術ブログ「Sarashina2.2 数学・コーディング性能向上」（2025年3

月） 103 24 - ELYZA公式サイト「ELYZA LLMデモ版」（2025年7月） 34 35 - PFNプレスリリース「PLaMo 2.0 Prime最優秀賞受賞」（2026年1月） 39 48 - サイバーエージェントニュース「独自日本語LLM ver3公開」（2024年7月） 61 60 - Google DeepMind Gemini 3 Pro関連情報（Zhihu記事、DeepMind公式） 70 74 - OpenAI ChatGPT-5.2に関する報道（量子位ニュース，2025年12月） 83 87 - Anthropic Claude Opus 4.5に関する情報（Zhihu記事, Tencent報道） 100 74

3 : よろず知財コンサルブログ『進化したNTTの純国産LLM「tsuzumi 2」』(2025年10月) – 「物理、数学、コーディング等の性能ではChatGPT-5より劣るようですが…」 4

9 : debono.jp『NECのLLMを技術的優位性から導入まで徹底解説！』(2025年8月) – 「NECのLLM『cotomi』は…日本語理解ベンチマークJGLUEで…海外製LLMを大幅に上回る世界トップクラス…」 9

34 : ELYZA公式『ELYZA LLM（デモ版)』(2025年7月) – 「数学能力を測る『MATH-500』『MATH-500(邦訳版)』…でOpenAI社のGPT-4oに匹敵するスコアを獲得。」 34

39 : PFNプレスリリース (2026年1月5日) – 「高い日本語性能、長文処理(32kコンテキスト対応)、文脈理解、検索拡張生成(RAG)など、日本語の実務に求められる精度と機能を安定して発揮」 39

70 : Zhihu質問「谷歌Gemini 3.0 Pro評価」回答 (2025年11月) – 「在MathArena测试中达到了惊人的23%…虽然SWE Bench略逊於Sonnet 4.5…」 70

83 : 量子位QbitAI『GPT-5.2反超谷歌Gemini 3 Pro』(2025年12月) – 「专家级数学评测FrontierMath(Tier 1-3)上，GPT-5.2 Thinking以40.3%的解题率创下新纪录。」 83

100 : Automatio AI『Claude Opus 4.5 LLMベンチデータ』(2025年) – 「来自AIME考试的竞赛级数学问题。Claude Opus 4.5在此基准测试中得分37%。MATH:85%。」 100

14 : NECプレスリリース (2025年7月8日) – 「問題解決の過程に着目した学習を強化…推論性能を高めた結果、タスクプランニングやタスク遂行における適切なツール選択能力が強化され…回答品質と速度の両面で向上」 14

61 : サイバーエージェントニュース (2024年7月9日) – 「スクラッチで開発を行なった225億パラメータ…CyberAgentLM3-22B-Chat…Nejumi LLMリーダーボード3において70BのMeta-Llama-3-70B-Instructと同等の性能」 61

51 : Google DeepMind公式『Gemini 3』(2025年11月) – 「新しいGemini 3 Proモデルは…開発者ツールにおけるAIの推論深度と信頼性を向上させ、Gemini 2.5 Proより50%以上多くのベンチマーク課題を解決」 51

---

1 〖レポート〗NTT R&D FORUM 2025「IOWN∴Quantum Leap」量子コンピューターとの出会いで進化するIOWNとAI | NTT STORY | NTT
https://group.ntt/jp/magazine/blog/rd-forum2025_report/

2 3 4 5 6 7 進化したNTTの純国産LLM「tsuzumi 2」
https://yorozuipsc.com/blog/nttllmtsuzumi-2

8 9 104 NECのLLMを技術的優位性から導入まで徹底解説！ - debono
https://debono.jp/7148

10 NEC開発の生成AI 「cotomi」
https://jpn.nec.com/LLM/cotomi.html

11 12 13 14 15 16 17 NEC、生成AI「cotomi」の性能強化でAIエージェントの活用を加速〜業務の更なる高度化を目指す〜 | 日本電気株式会社のプレスリリース
https://prtimes.jp/main/html/rd/p/000000989.000078149.html

18 19 20 26 27 28 国産生成AI 開発責任者インタビュー - 統合報告書 2025 | 企業・IR | ソフトバンク
https://www.softbank.jp/corp/ir/documents/integrated_reports/fy2025/tamba/

21 22 23 24 25 103 Sarashina2.2：数学・コーディングタスクの性能を向上させた日本語言語モデル - SB Intuitions TECH BLOG
https://www.sbintuitions.co.jp/blog/entry/2025/03/06/112144

29 30 31 32 33 34 35 36 ELYZA LLM (デモ版) | 株式会社ELYZA
https://elyza.ai/lp/elyza-llm

37 国産の日本語版”医療”特化LLM基盤「ELYZA-LLM-Med」を開発
https://prtimes.jp/main/html/rd/p/000000061.000047565.html

38 39 40 41 44 47 48 52 53 55 56 57 58 PFNの国産大規模言語モデルPLaMo 2.0 Prime、2025年日経優秀製品・サービス賞にて最優秀賞を受賞 | 株式会社Preferred Networksのプレスリリース
https://prtimes.jp/main/html/rd/p/000000020.000156310.html

42 43 49 50 59 大規模言語モデルの次期バージョン PLaMo 3 シリーズにおける8B, 31Bの小規模モデルによる事前学習の検証 - Preferred Networks Tech Blog
https://tech.preferred.jp/ja/blog/plamo_3_8b_31b/

45 46 GENIAC第3期で自律稼働デバイス向けの軽量な大規模視覚言語モデルPLaMo 2.1-8B-VLを開発 - 株式会社Preferred Networks
https://www.preferred.jp/ja/news/pr20251216

51 66 67 69 76 77 78 79 89 Gemini 3 - Google DeepMind
https://deepmind.google/models/gemini/

54 自社開発した大規模言語モデルをどうプロダクションに乗せて運用 ...
https://speakerdeck.com/pfn/20240906-cloud-operator-days-2024-pfn

60 61 独自の日本語LLM（大規模言語モデル）のバージョン3を一般公開 ―225億パラメータの商用利用可能なモデルを提供― | 株式会社サイバーエージェント
https://www.cyberagent.co.jp/news/detail/id=30463

62 独自の日本語LLM「CyberAgentLM2」に視覚を付与したVLM（大 ...
https://www.cyberagent.co.jp/news/detail/id=30344

63 国産お手軽ローカルLLMのSarashina2.2-3Bをご家庭パソコンで動かす
https://note.com/thediaryof/n/n2530090c219f

64 别再看榜单了！普通人也可以测出了各大编程模型真实差距 - 53AI
https://www.53ai.com/news/LargeLanguageModel/2026010710432.html

65 Gemini - 深獲學生喜愛的Google AI 學習好夥伴
https://gemini.google/tw/students/?hl=zh-TW

68 3. Gemini 3 Deep Think（更深入推理能力）： 類似OpenAI o1 的 ...
https://www.threads.com/@future.ai.tw/post/DRN96X0gVzS/3-gemini-3-deep-think%E6%9B%B4%E6%B7%B1%E5%85%A5%E6%8E%A8%E7%90%8
%E7%9A%84%E6%85%A2%E6%80%9D%E8%80%83%E6%A8%A1%E5%BC%8F%E9%87%9D%E5%B0%8D%E6%95%B8%E5%AD%B8%E7%A8%8B
%E5%85%B7%E5%82%99%E4%BA%86-sys

70 谷歌Gemini 3发布：AI数学物理双霸榜，Nano Banana Pro开启AI生 ...
https://aistudio.baidu.com/blog/detail/744879285998917

71 72 如何评价谷歌在2025年11月18日凌晨发布的gemini3.0pro模型？ - 知乎
https://www.zhihu.com/question/1974268445404177395

73 2025年AIモデル徹底分析：次世代言語・マルチモーダルモデルの展望とチャンピオン選定
https://mgx.dev/insights/
2025%E5%B9%B4ai%E3%83%A2%E3%83%86%E3%83%AB%E5%BE%B9%E5%BA%95%E5%88%86%E6%9E%90%E6%AC%A1%E4%B8%96%E4%B
9aede7cd1fc14e81b09fb1a40c44d893

74 Claude Opus 4.5 來了 網友：太神啦(懶人包)
https://tenten.co/learning/claude-opus-45/

75 Google 剛剛投下了震撼彈。Gemini 3 Pro 數據釋出，在MathArena ...
https://www.facebook.com/tentencreative/posts/google-%E5%89%9B%E5%89%9B%E6%8A%95%E4%B8%8B%E4%BA%86%E9%9C%87%E6%92%
%E6%95%B8%E6%93%9A%E9%87%8B%E5%87%BA%E5%9C%A8-matharena-
%E6%B8%AC%E8%A9%A6%E4%B8%AD%E9%81%94%E5%88%B0%E4%BA%86%E9%A9%9A%E4%BA%BA%E7%9A%84-23%E5%85%B6%E4%BB%
ui-%E7%90%86%E8%A7%A3%E8%83%BD%E5%8A%9B%E4%B8%8A%E9%81%99%E9%81%99%E9%A0%98%E5%85%88%E9%9B%96/131207736

80 81 82 83 84 85 86 87 88 90 91 92 93 94 95 GPT-5.2果然反超谷歌Gemini 3 Pro！北大数院校友核心贡献 – 量子位
https://www.qbitai.com/2025/12/360439.html

96 98 Introducing Claude Opus 4.5 - Anthropic
https://www.anthropic.com/news/claude-opus-4-5

97 降价升级！Opus 4.5 发布：依旧最强编码，超大杯现在只比Sonnet 贵 ...
https://cloud.tencent.com/developer/article/2595764

99 性能仍是天花板？Claude Opus 4.5 这一波“降价打击”让谁慌了？
https://zhuanlan.zhihu.com/p/1977775099995263303

100 102 Claude Opus 4.5 - LLM benchmark 数据| 免费试用 - Automatio AI
https://automatio.ai/zh/models/claude-opus-4-5

101 「北京版幻方」冷不丁开源SOTA代码大模型！一张3090就能跑 - 量子位
https://www.qbitai.com/2026/01/366408.html