

Gemini 3 Deep ThinkのAGIレベル4到達性を検証

AGIレベル4「AI自体による発明」とは何か

AGIレベル4とは、ソフトバンク孫正義氏やOpenAIが提唱するAI進化段階で「AI自身が新しい発明や発見を行う能力」を指します¹。レベル2は複数の博士号レベルの知性、レベル3は自律エージェント機能、そしてレベル4になるとAI自ら新たなアイデアや解決策を創出し、人類未踏の問題に独創的なアプローチで挑む段階です²。つまり知識や推論能力の枠を超え、AIが自主的に仮説提案・検証し発明を成し遂げることが求められます²。孫氏はこのレベル4の実現が「今後2年以内」に起こる可能性が高いと予測しており³、AIが人類の知らない革新的発明を行う「知のゴールドラッシュ」が始まりつつあると述べています³。

Gemini 3 Deep Thinkとは

Gemini 3 Deep Thinkは2025年12月4日に提供開始された、Google DeepMindの最新モデルGemini 3における強化推論モードです⁴。Gemini 3自体はマルチモーダル理解と最先端の推論能力を備え「あらゆるアイデアを現実にする」ことを目指した高性能AIモデルであり⁵⁶、Sundar Pichai氏も「AGI実現に向けての一步となるGemini 3を公開します」と位置付けています⁷。その中でDeep Thinkモードは複数の仮説を同時並行で検討する高度な並列推論を可能にし、難問解決力を飛躍的に向上させるオプション機能です⁸。通常のモデルが一連の推論を順次行うのに対し、Deep Thinkでは複数の推論経路を並行探索することで見落としを減らし、より正確な回答にたどり着きやすくなっています⁸。このモードは回答に数分かける代わりに、コード生成や図表作成の精度も高める「じっくり考えるAI」として位置付けられています⁹¹⁰。Gemini 3 Deep Thinkは当初安全性テスト目的で限定提供され、Google AI Ultraサブスクリューザ向けにリリースされました⁵¹¹。

最先端ベンチマークでの性能比較（定量評価）

Gemini 3 Deep Thinkが本当に“発明級”の知性に到達したかを測るため、最新の難関ベンチマークでの成績を確認します。FrontierMathやCritPt、ARC-AGI-2、Humanity's Last Exam (HLE) といった試験で、Gemini 3 Deep Thinkは現行モデル中トップクラスのスコアを記録しています。

Gemini 3 Deep Thinkモードの評価結果（Google公式ブログより）¹²。従来のGemini 3 Proを上回り、複数のベンチマークで新記録を樹立している。

- **Humanity's Last Exam (HLE)** – 人類最終試験とも呼ばれる難問集合：Gemini 3 Proは**37.5%**（ツール未使用）正答でしたが、Deep Thinkは**41.0%**まで向上し、**博士号レベルの推論力**を示しました¹³¹²。わずかながらも4割を超えたのは画期的で、GPT-5.1などOpenAI系モデルを含め最高水準です（GPT-5系は3割台前半と報じられます¹⁴）。
- **FrontierMath** – 未公開の超難度数学問題集：Gemini 3 Proは**37.6%**（290題中109問正解）でしたが、Deep Thinkでは**40%以上**に達し¹⁴¹⁵、**高級GPT-5.1（約32.4%）を大きく上回る記録**となりました¹⁵。これにより「真に難解な数学パズルでGeminiが先頭に立った」と評価されています¹⁴。特に未解決級のTier 4問題ではDeep Thinkが**18~19%**程度正解し、GPT-5.1 Proの約12~15%を大幅に凌駕したとの分析があります¹⁶¹⁵。
- **CritPt** – 博士課程レベルの物理研究課題71問からなる新ベンチマーク：最新評価によれば、Gemini 3 Pro (Preview版)が**9.1%**の正解率でトップ、OpenAIのGPT-5.1（high設定）は**4.9%**で2位でした¹⁷。**GeminiがGPT-5.1の約2倍に迫るスコアを記録したものの、いずれも90%以上の問題に失**

敗しており、「自律的な科学者」には程遠い結果と報告されています¹⁷¹⁸。この低スコアは**現行のAIにはオープンエンドな物理問題を解く創造性・精密さが欠けている**ことを示すと同時に、Gemini系モデルが部分的なサブタスクでは改善傾向を示すことも確認されました¹⁹。

- **ARC-AGI-2** – 汎用推論テスト（Adaptive Reasoning Challenge第2版）：論理パズルや新規問題への汎化能力を測るもので、Gemini 3 Proは約**31%**、Deep Thinkでは**45.1%**という**前例のない高スコア**を達成しました²⁰。コード実行などツール使用込みの評価ですが、従来モデルが20%未満に留まっていた中で突出した値であり、**未知の課題への解決力が飛躍した**ことを示しています²⁰。OpenAI側のGPT-5.1 Proは約17.6%程度とされ、Deep Thinkは**+27ポイント以上**も上回った計算になります（Gemini 3世代での大幅な躍進）²¹²²。
- **GPQA Diamond** – 大学院課程レベルの科学QA: Gemini 3 Proは**91.9%**正答と既に非常に高精度ですが、Deep Thinkでは**93.8%**まで僅かに向上しました¹⁶。科学知識の応用においてGPT-5.1を含め群を抜く**正確さ**を示しています¹⁶。
- **その他**：マルチモーダル推論では、Gemini 3 Proが画像・表理解テスト**MMMU-Pro**で**81%**、動画推論**Video-MMMU**で**87.6%**という画期的スコアを記録し²³、Deep Thinkでもこの分野のさらなる向上が期待されています。事実検証の**SimpleQA Verified**でも**72.1%**と従来を上回る正確性を達成しました²³。

以上のように、**Deep Thinkモードは主要ベンチマークで軒並みGemini 3 Proを上回り、GPT-5.1など他モデルに対しても優位に立っています**。特に数学・論理・コード推論といった**発明の土台となる能力**で最先端の結果を出しており、「**博士号レベルの推論力**」を**明確に実証した**と言えます²⁴¹³。一方で、CritPtのような真に創造性を要する未知問題では依然正解率1割前後に留まるなど、**人間研究者レベルには遠い現実も浮き彫り**です¹⁷²⁵。

創造性・推論力・多モーダル理解の進化（定性的評価）

Gemini 3 Deep Thinkは定量的なスコアだけでなく、**質的な思考能力の深化**でも注目されています。まず**推論の深みとニュアンス理解**が飛躍的に向上しました。短い質問からでも背後にある文脈や意図を読み取り、ありきたりな回答ではなく「**聞きたいことではなく、聞くべきこと**」を**提案できる真の洞察**を示すとされています²⁶。これは過剰なリップサービスではなく簡潔で的確な応答を返すという質的向上であり、**ユーザの曖昧な要求から本質を汲み取る能力**は「その場の空気を読むAI」にまで進化したと評されています²⁷。

また**創造性**の面でも、Gemini 3はユニークな成果を見せています。例えば**核融合実験装置トカマク内部のプラズマ流動を可視化するコードを自動生成**しつつ、それを詩的表現で説明することまで可能です²⁸。高度な科学概念をコードと自然言語の両面から理解・表現できる点は、**理系研究と文学的創造の両能力を併せ持つ**ような柔軟性と言えます。Deep Thinkモードではさらに出力の正確性・緻密さが増し、生成するコードや図表の品質が一段と高まると報告されています⁹。

Gemini 3は**マルチモーダル理解**でも世界最高水準です。テキスト・画像・動画・音声・コードといった**異なるモダリティの情報を1つのコンテキストに統合し、最大100万トークンにも及ぶ長大な入力にも対応**します²⁹。その結果、例えば手書きの多言語レシピ画像を読み取って翻訳・書き起こしたり、学術論文や長時間の講義動画から知識カードやビジュアル解説を生成したり、スポーツ試合の動画を分析して訓練プランを作成したりできます³⁰。これは**視覚・空間認識と高度言語推論の統合**を意味し、人間が視覚資料と文章を行き来しながら学習・発明するプロセスをAIがある程度再現できている証左です。

総じて**推論力（Reasoning）・創造力（Creativity）・多モーダル理解力**のいずれもが前世代を凌ぐレベルに達しており、**未知の問題に柔軟に取り組む基礎的素養**が備わりつつあると評価できます。特にDeep Thinkでは**時間をかけて自己検証・自己訂正するプロセス**が実装されており、難題に対する**思考の持続力と探索力**が格段に強化されています⁸¹⁰。これは人間で言えば「熟考しながら様々な可能性を試す研究者の思考法」に近づいており、創造的ブレークスルーへの片鱗を見せる部分と言えるでしょう。

Gemini 3 Proとの能力差と「Deep Think」強化点

Gemini 3 Pro（通常モード）とDeep Thinkモードの違いを技術的に整理すると、以下のようなポイントがあります：

- **推論アプローチ**: Proは**高速応答**を重視し、単一路線の思考で素早く答える設計ですが、Deep Thinkは**並列かつ継続的な探索**を行います⁸。複数の仮説を同時に検証することで、**複雑な問題でも見落としを減らす狙い**があります³¹。
- **性能向上領域**: Deep Thinkは特に**高度な数学・科学・論理問題**や**コーディング**で威力を発揮します³²。前述の通りベンチマークでも数学推論・抽象推論でProを凌駕する成績を収めており、**難解タスク専門のブーストモード**と位置付けられます²⁰。
- **リソースと応答時間**: Deep Thinkは**追加の計算資源**と**長い思考時間**を要します。回答生成に数分かけるため、Proのような瞬時回答はできませんが、その**分精度重視**です³³。実運用では**コストや待ち時間**とのトレードオフがあり、必要な場面でDeep Thinkを有効化する形になります³⁴³⁵。一部分析では、このモードは大幅にトークン消費が増えるため**実用上のコスト課題**も指摘されています³⁵。
- **利用シーン**: Proは汎用対話や日常タスク向けで、**リアルタイム性**が求められる用途に適します。一方Deep Thinkは研究開発や専門分析の場で、**正確さ・網羅性が優先されるタスク**に適しています³⁶³⁷。「すぐ答えるAIから、じっくり考えるパートナーへ」というコンセプトで、転換点となるアップデートと位置付けられました³⁸³⁹。

以上のように、Gemini 3 Deep Thinkは**Gemini 3 Proを土台に推論深度を深めた上位モード**であり、**大規模モデルにおける“二段構えの知性”**の実例です⁴⁰。簡易な問いには迅速なPro、難問には思考拡張したDeep Thinkという**使い分け**が可能になったことで、AIの実用範囲が拡大しました。技術的にも、OpenAIやAnthropicなど競合に対し「**高精度だがコスト高の特化モード**」でリードする戦略であり⁴¹、今後このような**モード切替型インテリジェンス**が主流になる可能性も示唆されています⁴⁰。

AI自身の発見・発明の事例はあるか？

肝心の「**AI自ら新発見・新理論を生み出した例**」については、現時点で明確なブレイクスルーは報告されていません。Gemini 3 Deep Think自体が**独立に科学的発見を成し遂げた**という公式事例は無し、というのが実情です。例えば、CritPtベンチマークは「AIが未知の物理課題を解けるか」を測りましたが、最高のGemini 3 Proでも1割しか解けず、**自主的に新知見を獲得するには程遠い結果**でした¹⁷²⁵。このことは**AIが独力で科学者となるには依然大きな壁がある**ことを示唆します。

しかし一方で、AIが**研究プロセスの自動化**に寄与し始めている例はいくつかあります。日本のSakana AI社が公開した「**AI Scientist**」システムは、研究アイデア生成から実験実行、論文執筆・査読対応までを自動化する試みです⁴²。2025年には**完全AI生成の論文**が機械学習分野の国際会議(ICLR 2025)ワークショップで**査読通過・採択**される歴史的成果も報告されました⁴³。その論文は平均スコア6.33を獲得し、**初の「AIだけで書かれた査読論文」**となったのです⁴³。ただしこれはメイン会議ではなく採択率60-70%のワークショップであり、人間専門家による引用ミス指摘等を受け**最終的には撤回**されています⁴⁴。品質的にも「急いで書かれた学部生の論文程度」と評され、**約42%の実験が失敗**するなど課題も多く報告されています⁴⁵⁴⁶。つまり**AIが完全に独創的かつ信頼性ある研究を成し遂げるには尚改善が必要**であり、現状は**人間研究者の強力な助手**として部分的工程を自動化する段階といえます⁴⁷⁴⁸。

他にも、DeepMindによる**AlphaDev**が独自に高速な新アルゴリズム（ソート算法）を発見した例や、創薬分野でAIが**新薬候補分子**をデザインした事例など、**限定的ながら「AI発明」の先行例**は存在します。これらは特化システムや人間の審査を伴うもので、汎用モデルGemini 3 Deep Think自身が自主的に発明者となったとは言えません。しかし**Gemini 3の高い推論・探索能力**は、今後こうした自律的発見を後押しする土台になり得ます。実際、スタンフォードや清華大学の研究者は「**AIエージェントがトップ研究者のように大胆な仮説飛躍を再現できるか**」を模索しており、Gemini級モデルにより**科学研究の自動化が現実味を帯びてきた**との見

方もあります⁴⁹。OpenAIも2026年にAI研究インターン、2028年までに完全自律研究者を目指す計画を公表しており⁵⁰、業界全体でレベル4（発明AI）実現へのロードマップが描かれ始めています。

総じて、現時点でGemini 3 Deep Thinkが人間の発明家のように自律的イノベーションを起こした例は確認されていません。しかし、論文執筆の自動化やアルゴリズム発見など端緒となる事例は増えており、Gemini 3の持つ高い知性がその延長線上にあることは間違いありません。今後、安全性と信頼性を確保しつつ、いかにこの芽生えた創造力を実用的な発明に結び付けられるかが鍵となるでしょう。

専門家・企業による評価と公式声明

孫正義氏（ソフトバンクG）はGemini 3の登場を踏まえ、「AGIのレベル4到達が視野に入った」と楽観的な見解を示しています。彼は2024年末時点で「レベル4（AI自ら発明）は2～3年で達成される見込み」と語り¹、レベル5（AIが組織的活動を開始）を経て10年以内に超知性（ASI）時代が来ると予測しました¹。孫氏いわくGemini 3 Proは既にレベル3（エージェント）相当の能力を持ち、**レベル4への境界領域に位置している**との指摘もあります²¹。実際、ソフトバンクは生成AIの研究開発投資を加速しており、「知のゴールドラッシュ」を迎える準備を進めているようです。

OpenAI側の視点では、対抗モデルGPT-5.1の登場もあって熾烈な競争が展開されています。OpenAIは公式に「現時点のGPT-5系列は万能の科学者には程遠いが、既に研究者の時間を節約し有用なリサーチアシスタントとなりつつある」と述べています⁵⁰。同社は2026年までに“AI研究インターン”システムをリリースし、**2028年頃までに“完全自律の研究者AI”を実現**する計画を明かしており⁵⁰、Gemini 3 Deep Thinkの成果もそのロードマップ上の刺激と捉えているようです。OpenAI関係者からは「Gemini 3は確かに素晴らしいが、依然として**発明の核心は人間がリードしている**」といった慎重な意見も報じられています（裏付けとして、CritPt結果などからも現行モデルの限界が示されています¹⁸¹⁷）。

Google DeepMindはGemini 3の公式発表において、「あらゆるアイデアを現実にする最も高性能なモデル」と謳い⁵¹、深層推論による新たな知性の時代を強調しました。CEOのDemis Hassabis氏やKoray Kavukcuoglu氏も、Gemini 3 Deep Thinkが**高度な洞察とニュアンス理解でこれまでにない思考力**をもたらすと自負しています²⁶。一方で、研究者らは「依然ほとんどの難題は解決できない」という現実を直視すべきと指摘しており¹⁸²⁵、Google内部でも安全性検証を慎重に行いながら能力拡張を図る方針です⁵¹⁰。総じて、**業界有力者はGemini 3 Deep Thinkのポテンシャルを認めつつも「完全な発明AI」にはなお距離がある**との認識で一致しています。創造性・自律性の向上は目覚ましいものの、**人類の科学者を代替できるレベルには達していない**という慎重な評価が多い印象です⁴⁷⁴⁹。

結論：AGIレベル4に到達したと言えるか？

以上の分析から、Gemini 3 Deep ThinkはAGIレベル4（AI自身による発明）の入り口に立ちつつあるものの、**完全到達と断言するには時期尚早**と判断されます。⁵²でも指摘されるように、現段階ではレベル3（自律エージェント）とレベル4（イノベーター）の境界領域に位置し、いくつかの重要な能力で**レベル4への片鱗を示している**状況です²¹。

まず定量面では、Deep Thinkが難関ベンチマークで軒並み最高記録を更新し、**博士レベルの推論力**や限定的ながら**未知課題への適応力**を証明しました¹²¹⁷。これはレベル4に必要な「高度な知識と推論」を満たす基盤と言えます。しかし一方で、その**発明性を示す質的跳躍**（まったく新しい概念の創出や独創的仮説の検証）はまだ限定的です。自律モードで長時間思考させても、CritPtのような本物の研究課題では**大半が解決不能**であり¹⁷、人間が思いつかないような斬新な発明・発見が生まれたとは報告されていません。

とはいえ、Gemini 3 Deep Thinkは**明確に前世代を超える創造的知能の兆候**を示しています。並列推論による粘り強い問題探索や⁸、モダリティ横断の理解力、背景意図を汲む洞察など²⁶、「**機械が自ら考え、新**

しい答えを出す」方向への進化は確実です。このような能力に、人間研究者の監督や適切な強化学習を組み合わせれば、近い将来に部分的な発明（例：新材料の発見、新アルゴリズムの創出）が実現する可能性は十分あります。実際、専門家も「今後数年でエージェントから発明者へとAIは進化する」と期待を寄せています³。

結論として、Gemini 3 Deep ThinkはAGIレベル4に「あと一歩」の段階です。現時点では自律的发明を成し遂げたとは言えないものの、創造性・推論力・自律性の飛躍的向上によりレベル4要件の多くを備え始めています²¹。今後、実際にAIが人類未踏の発見を生み出す具体例（たとえば新理論の提唱やノーベル賞級の成果）が出てくれば、真にレベル4到達と評価できるでしょう。その日が訪れるまで、Gemini 3 Deep ThinkはAGIレベル4への有望な橋頭堡であり、引き続き技術的・倫理的検証を重ねつつ発展していくものと考えられます。現状の判断材料としては、「完全なAGIレベル4には未達だが、その片鱗と前提条件を十分に備えている」というバランスの取れた評価が妥当です⁵²⁴⁷。

参考文献・出典: Sundar Pichai他「Gemini 3 発表ブログ」¹³¹²、天秤AIニュース²⁰、i10x解析記事¹⁵、The Decoder記事¹⁷⁴⁹、孫正義氏講演資料¹³、包括的レビュー資料²¹、Sakana AI関連報告⁴³ など。

¹ 「超知性は10年以内」 孫正義「AIの王道」を語る SoftBank World 2024 - Impress Watch
<https://www.watch.impress.co.jp/docs/news/1628277.html>

² ³ yorozuipsc.com
<https://yorozuipsc.com/uploads/1/3/2/5/132566344/9c63e2ecf1b77892bc3b.pdf>

⁴ ⁸ ⁹ ¹⁰ ¹¹ ²⁰ ³¹ ³² ³³ ³⁶ ³⁸ ³⁹ Google、最上位AIモデル「Gemini 3 Deep Think」を提供開始 | 天秤AIメディア byGMO
https://tenbin.ai/media/generative_ai/AI_NEWS_251205

⁵ ⁶ ⁷ ¹² ¹³ ²³ ²⁶ ²⁷ ²⁸ ²⁹ ³⁰ ⁵¹ Gemini 3: Introducing the latest Gemini AI model from Google
<https://blog.google/intl/ja-jp/company-news/technology/gemini-3/>

¹⁴ ¹⁵ ³⁴ ³⁵ ³⁷ ⁴⁰ ⁴¹ Gemini 3 Tops FrontierMath: AI Math Record & Costs
<https://i10x.ai/news/gemini-3-frontiermath-benchmark-analysis>

¹⁶ ²¹ ²² ²⁴ ⁵² yorozuipsc.com
<https://yorozuipsc.com/uploads/1/3/2/5/132566344/76341f1f00c54baecbd0.pdf>

¹⁷ ¹⁸ ¹⁹ ²⁵ ⁴⁷ ⁴⁸ ⁴⁹ ⁵⁰ Gemini 3 Pro and GPT-5 still fail at complex physics tasks designed for real scientific research
<https://the-decoder.com/gemini-3-pro-and-gpt-5-still-fail-at-complex-physics-tasks-designed-for-real-scientific-research/>

⁴² ⁴³ ⁴⁴ ⁴⁵ ⁴⁶ Sakana AIのAI Scientist：自動科学研究の革命と課題 - 包括的レビュー - ニュースポータル
https://hatohato.jp/ai/article/20250608_sakana_ai_scientist.php