ChatGPT

エージェント

発明創出に最も力を発揮する生成AIモデルの比較分析

はじめに

2025年11月、生成AIの世界では OpenAI の **GPT-5.1 Pro**（GPT-5シリーズのプロ版）と Google DeepMind の **Gemini 3 Pro** がほぼ同時に公開された。どちらも従来より桁違いの推論能力やマルチモーダル処理を備え、単なる対話型アシスタントを超えて **新しいアイデアや科学的発見を生み出す研究パートナー** となることが期待されている。最新のベンチマークや公開情報を基に両モデルの能力を整理し、発明の「創出（アイデア出し・理論構築）」段階に最適なモデルを特定する。

1. 発明・科学的発見分野における生成AIの最新トレンド

1.1 ベンチマークの高度化

2024年頃までの大規模言語モデルは中高生レベルの数学 (GSM8k) や MATH ベンチマークで評価されていた。しかし 2025 年には AI 研究の焦点が **未発表の専門家レベルの問題** に移行し、以下のようなベンチマークが登場した。

ベンチマーク	目的	背景・ポイント
FrontierMath	Epoch AI による未発表の専門家レベル数学問題で、暗記の影響を排除して純粋な数学的推論能力を測る。	Gemini 3 Pro が Tier 1-3 で38%、Tier 4 で19%の正答率を記録し、旧世代モデルを大きく上回った。
CritPt	50人以上の物理学者が作成した研究レベルの複合課題で、物理学的洞察力を測る。	Gemini 3 Pro が9.1%でトップ、GPT-5.1 Pro（high）は約5%に留まった。
ARC-AGI-2	言語に依存しない抽象的視覚パズル。深層推論が求められる。	Gemini 3 Pro が31.1%（Deep Think モード45.1%）で、GPT-5.1 の17.6%を大きく上回る。
Humanity's Last Exam	大学院レベル以上の難問を含む総合推論テスト。	Gemini 3 Pro が37.5%（Deep Think で41%）を記録し、GPT-5.1 の26.5%を約11ポイント上回った。
AI Idea Bench 2025	約3,495本のAI論文とその派生研究を分析し、LLMのアイデア生成力を評価するベンチマーク [1]。	研究段階のアイデア創出の定量評価が議論され始めた。

これらのベンチマークは人間が数時間〜数日かけて解く問題で構成され、単純な記憶やパターン認識では解けないため、発明や科学的発見に直結する **創造的推論能力** を測る指標になっている。

1.2 モデルの特徴とリリース

- **GPT-5.1 Pro** – OpenAI が2025年11月19日にリリースした GPT-5 シリーズのプロ版。推定で数千億〜兆単位のパラメータを持つ密なモデルまたは Mixture-of-Experts (MoE) アーキテクチャで、タスクの

難易度に応じて計算リソースを調整する **適応的推論機構** を採用。標準で128k〜400kトークン前後のコンテキスト長を持ち、Instant / Thinking の2モードが用意される。ではChatGPTでは20万トークン前後が利用でき、APIでは40万超の長文対応が可能とされる。

- **Gemini 3 Pro** – Google DeepMind が2025年11月18日に発表した旗艦モデルで、Sparse MoE アーキテクチャとネイティブなマルチモーダル統合を採用。1,048,576トークン（約100万）の巨大なコンテキストウィンドウと「Deep Think」モードにより、長大なデータを読み込みながら深い推論が可能。画像・音声・動画・コードを同じコンテキスト内で処理できる。

1.3 トレンドのまとめ

リサーチサイト Vellum がまとめたベンチマークでは、Gemini 3 Pro の **ARC-AGI-2** (31.1%) や **Humanity's Last Exam** (37.5%) の高スコアが「高度な推論と長期的意思決定に優れる」と評価されている [2]。また、**Vending-Bench 2** という長期経済シミュレーションでは Gemini 3 Pro の平均資産が $5,478.16 で GPT-5.1 の 272% 上 [3] であり、実用的なエージェントワークフローで優位にあることが示されている。

2. 数学・科学推論における性能比較

2.1 FrontierMath と数学系ベンチマーク

- **FrontierMath** – Gemini 3 Pro はTier 1–3で38%正答、Tier 4 (研究レベル) で19%に達し、従来モデルの数%という水準から指数関数的に飛躍。GPT-5.1 (high) はTier 1–3で約26〜31%、Tier 4で約12–15%にとどまる。

- **AIME 2025 (数学オリンピック)** – Gemini 3 Proはツールなしで95.0%正答し、GPT-5.1は71〜94%と推定され差が大きい。コード実行ツールを許可すると両モデルとも100%に達するため、Gemini の方が **内在的な数学直感** が高いと評価される [4]。

- **MathArena Apex** – 最難関の競技数学問題で Gemini 3 Pro が23.4%を記録し、GPT-5.1 の1.0%を20倍以上上回った。難問への突破力では Gemini が突出している。

- **LiveCodeBench** – アルゴリズム設計能力を測るベンチマークで、Gemini 3 Pro の ELO スコアは 2,439 と GPT-5.1 の 2,243 を200ポイント近く上回る。コードの創造的生成でも Gemini の優位が示される。 [5] では Gemini がアルゴリズム的問題解決に適していると指摘された。

2.2 物理・科学的推論ベンチマーク

- **CritPt** – 大学院レベルの物理研究課題で、Gemini 3 Pro は 9.1% を記録し GPT-5.1 の約5%を大きく引き離した。スコア自体は低いが、現行モデルで最も物理推論に強いモデルとして注目されている。

- **GPQA Diamond** – 物理・化学・生物学の博士レベルの質問で、Gemini 3 Pro は 91.9% (Deep Think 93.8%)、GPT-5.1 は 88.1% で差は小さい。学術知識自体は両モデルとも高水準だが、Geminiの方が僅かに上。

- **Humanity's Last Exam** – 複数分野の大学院レベル問題を含む総合推論テストで、Gemini 3 Pro は 37.5% (Deep Think 41%) を記録し、GPT-5.1 の 26.5% を約11ポイント上回った。これは AI が「未知の問題」に取り組む能力の差を示している。

2.3 Abstract Reasoning (ARC-AGI-2)

ARC-AGI-2 では言語に依存しない抽象的視覚パズルを解く能力が測られる。Gemini 3 Pro は 31.1% で、Deep Think モードでは 45.1% に達し、GPT-5.1 の 17.6% のほぼ 2 倍。Vellum のまとめもこの大幅な差を強調し、新しい問題パターンへの適応力や視覚的推論能力で Gemini が突出していると評価している [6] 。

2.4 数学・科学推論の総括

Gemini 3 Pro は高度な数学・物理・抽象推論ベンチマークの多くで GPT-5.1 Pro を上回り、人間でも難しい未公開問題に対して創造的に解法を見つける力が強い。GPT-5.1 Pro も SWE-Bench など実務的なコーディングタスクや AIME のツール併用環境では遜色ないものの、**未知の問題をゼロから解く能力** では Gemini が優位である。

3. マルチモーダル理解と長文脈推論

3.1 マルチモーダルベンチマーク

Gemini 3 Pro はテキスト・画像・音声・動画・コードを統一コンテキストで処理するネイティブなマルチモーダルモデルであり、以下のベンチマークで卓越した結果を示す。

ベンチマーク	Gemini 3 Pro	GPT-5.1 Pro	考察
MMMU-Pro (マルチモーダル大学試験)	81.0%	推定76–78%	画像・テキスト混在の問題に強く、Geminiが5ポイント差でリード。
Video-MMMU (動画理解)	87.6%	80〜82%	動画の時間的・空間的情報を統合する能力で7ポイント以上の差。工場監視や長時間会議の要約など実用価値が高い。
ScreenSpot-Pro (画面理解)	72.7%	3.5%	UI自動化やスクリーンショット解析でGeminiが圧倒。

Gemini のアーキテクチャはマルチモーダル統合時に「モダリティ間の変換ロスがない」ため、音声や動画から物理的因果関係を推論できる。これが発明領域で重要な「図面や現象の理解」に直結する。一方 GPT-5.1 Pro も画像入力をサポートするが、動画理解は限定的でありマルチモーダル推論では差がある。

3.2 長文脈推論と Deep Think

- **コンテキストウィンドウ** – Gemini 3 Pro は 1,048,576トークンの巨大なコンテキストウィンドウを持ち、出力も65kトークンに達する。これは数百ページの文献や数時間分の動画、数万行のコードを一度に読み込める。GPT-5.1 Pro は標準では約128k〜400kトークンのコンテキスト長で、長文は複数プロンプトに分割して処理する設計である。

- **Deep Think と適応的推論** – Gemini の Deep Think モードは、難問に対して内部で長時間の思考ループを実行し、連鎖的な推論と自己検証を行う機能である。対して GPT-5.1 Pro は Instant/Thinking の2モードで計算量を動的に切り替え、簡単なタスクには高速応答、難しいタスクには追加の推論時間を割く。Gemini の方が長い文脈と深い推論を同時に扱えるため、複雑な理論構築や膨大な先行文献を読み解く発明フェーズに適している。

- **長文脈下での精度** – Vellum による MRCR v2 ベンチマークでは、Gemini 3 Pro が128kの平均コンテキストで77.0%の情報検索精度を示し、1Mコンテキストでも旧世代より9.9%向上したと報告されている[7]。長文でも要点を見失いにくいことは、特許文献や膨大な研究資料を読解する際に重要である。

4. コーディング・エージェント性能と実務適性

4.1 ソフトウェア・エージェントベンチマーク

ベンチマーク	Gemini 3 Pro	GPT-5.1 Pro	評価
SWE-Bench Verified (実リポジトリのバグ修正)	76.2%・	76.3%	両モデルとも同等で、GPT-5.1が僅差で首位。
Terminal-Bench 2.0 (ターミナル操作)	54.2%・	47.6〜58.1%	Gemini が Linux コマンド操作やデバッグでリード。
LiveCodeBench (競技プログラミング)	ELO 2,439	ELO 2,243	ゼロからアルゴリズムを設計する能力で Gemini が200点差の優位。
Vending-Bench 2 (長期経済シミュレーション)	$5,478.16・[3]	$1,473.43	長期的な意思決定と計画でGeminiが272%高い成果を生んだ。

Gemini はアルゴリズム設計や長期意思決定に強い一方、GPT-5.1 Pro (特に Codex-Max) はバグ修正など実務的な反復作業で僅差ながら再現性と安定性に優れる。

4.2 コストと運用面

OpenAI の GPT-5.1 Pro は 100万トークンあたり入力$1.25／出力$10.00 と比較的低料金で、キャッシュ機能を利用すれば入力コストを大幅に削減できる。Gemini 3 Pro は 200k トークン以下で入力$2・出力$12、200k超で入力$4・出力$18と「コンテキスト税」が存在する。ただし 1M級の長文に対して人間が数日かけて読むコストを考慮すると依然として割安と評価される。

レイテンシ面では GPT-5.1 Pro の Instant モードが高速で、ユーザー体験を重視する日常の会話やチャットボットに適する。一方 Gemini 3 Pro は深い推論のため応答が遅いが、発明創出など長時間かけて考察するタスクでは欠点になりにくい。

5. モデルの特性と発明フェーズへの適性

5.1 定性的評価 – “天才的な研究パートナー”と“究極の熟練労働者”

Gemini 3 Pro は巨大なコンテキストと深い推論能力により、未知の解を探索する「**天才的な研究パートナー**」として描かれている。特に FrontierMath や CritPt など未解決問題のベンチマークで歴代最高スコアを記録し、物理現象や数学的構造の理解に強いことから **理論構築や新規アイデア創出** のフェーズで価値を発揮する。

GPT-5.1 Pro はコンパクション（意味的圧縮）や適応的推論により安定したパフォーマンスとコスト効率を提供し、既存のエンジニアリングプロセスにシームレスに統合できる「**究極の熟練労働者**」とされる。
SWE-Bench や Terminal-Bench では再現性が高く、コーディングやバグ修正など実務を確実にこなす能力に優れている。

5.2 発明の各段階における適性

発明フェーズ	Gemini 3 Proの適性	GPT-5.1 Proの適性
アイデア創出・理論構築	FrontierMath/ARC-AGI-2/CritPt で高い創造的推論能力を示し、未知の問題に対する「ひらめき」や抽象的連想に優れる。1Mトークンの長文脈とDeep Think により膨大な論文や特許を読み込み、相関関係を見つける能力が高い。	適応的推論である程度の深掘りは可能だが、コンテキスト長が短いため長大な文書を一括解析する際は分割が必要。理論構築よりも既存知識の整理が得意。
実装・プロトタイプ作成	LiveCodeBench では優れたアルゴリズム設計が可能で、ゼロからコードを書く能力が高い。ただしバグ修正や依存関係の管理ではリトライ回数が多い可能性があり、安定性に課題がある。	SWE-Bench で僅差ながらトップの修正能力を持ち、Codex-Max により長期稼働やエッジケース対応が安定。費用対効果が高く、日常的な開発やエージェント実行に適する。
長期的な自律実験	Vending-Bench 2 の結果から、Gemini が長期的意思決定と経済的価値創出で大幅に優位。科学研究やビジネスシミュレーションでの自律エージェントに向く。	コストは安いが長期的利益が小さいため、短期タスク向き。

6. 結論 – 発明創出に最も力を発揮するモデル

最新のベンチマークと定性的評価を総合すると、**発明の「創出」段階において最も力を発揮する生成AIは Gemini 3 Pro である**。理由は以下のとおり：

1. **創造的推論の卓越性** – FrontierMath、ARC-AGI-2、CritPt など研究レベルのベンチマークで Gemini 3 Pro が大きくリードし、未知の数学問題や物理現象に対して新しい解法や関係性を見つける能力を実証している。
2. **マルチモーダル・長文脈対応** – 100万トークンのコンテキストウィンドウと Video-MMMU/ ScreenSpot-Pro で示された圧倒的なマルチモーダル理解力により、図面・実験データ・論文・動画を一体で扱い、複雑な発明タスクに必要な情報を統合できる。
3. **長期思考と深い推論** – Deep Think モードは長い思考チェーンを実行して解答を検証するため、アイデア生成や理論構築で重要な「多段階の熟考」が可能。MRCR v2 の結果も長文脈下での高精度を裏付ける [7] 。
4. **発明につながる数学的直感** – AIME 2025 や MathArena Apex で顕著な差が示すように、Gemini 3 Pro は外部ツールに頼らず難問を解く数学的センスを持ち、アイデアの出発点となる理論的洞察を自ら生み出す。
5. **欠点の管理可能性** – Gemini 3 Pro の欠点は応答遅延やコスト、幻覚率の高さ（誤答を自信を持って述べる率が高い）であるが、アイデア創出フェーズでは即応性より洞察の深さが重視される。生成物を人間が検証する前提で利用すれば、幻覚率の問題は大きな障害ではない。

一方で **GPT-5.1 Pro** は、SWE-Bench や Terminal-Bench での安定性、低コスト、Instant モードによる高速応答などから、発明後の実装・検証や日常的な開発支援に優れている。したがって実務では、**Gemini 3 Pro で発明や理論構築を行い、GPT-5.1 Pro でプロトタイプの実装や運用を行うハイブリッド戦略** が推奨される。

要約

- **トレンド** – 2025年は FrontierMath や CritPt など研究レベルのベンチマークが登場し、生成AIの評価軸が「暗記」から「創造的推論」へ移行した。Vellum 等によるまとめでも、Gemini 3 Pro が ARC-AGI-2 や Humanity's Last Exam で高得点を記録したことが強調されている [2]。
- **数学・科学推論** – FrontierMath では Gemini 3 Pro が38%（Tier 1–3）/19%（Tier 4）の正答率を記録し、GPT-5.1 Pro は26〜31%/12〜15%に留まる。AIME 2025 や MathArena Apex でも Gemini が大きな差でリード。
- **マルチモーダル理解** – MMMU-Pro (81.0%対76〜78%)、Video-MMMU (87.6%対80〜82%)、ScreenSpot-Pro (72.7%対3.5%) などで Gemini が圧倒。1Mトークンの長文脈と Deep Think による深い推論も強み。
- **コーディング・エージェント** – SWE-Bench では両モデルがほぼ同等（76%台）だが、LiveCodeBench や Vending-Bench では Gemini がリード。GPT-5.1 Pro はコストと安定性で優れる。
- **結論** – **Gemini 3 Pro は発明の創出フェーズにおいて最も適した生成AIであり、複雑な数学的・科学的課題やマルチモーダル情報を扱う際に卓越したパフォーマンスを示す**。一方、実装や実務的開発では GPT-5.1 Pro が安定性とコスト効率で優位であり、両者を組み合わせることで発明から実装までの全過程を強化できる。

---

[1] [2504.14191] AI Idea Bench 2025: AI Research Idea Generation Benchmark
https://arxiv.org/abs/2504.14191

[2] [3] [4] [5] [6] [7] Google Gemini 3 Benchmarks (Explained)
https://www.vellum.ai/blog/google-gemini-3-benchmarks