2025年、発明創出AIの頂上決戦：Gemini 3 Pro vs GPT-5.1 Pro

どちらが真の「研究パートナー」となり得るのか？

2025年11月、OpenAIとGoogle DeepMindは次世代の巨大モデルをほぼ同時に公開した。これらは単なるアシスタントではない。科学的発見と技術革新のプロセスそのものを変革する可能性を秘めている。本分析では、特に「発明創出」という最も創造的なタスクにおいて、どちらのモデルが優れているかを徹底的に解剖する。

AI評価の新時代：知識の記憶から「創造的推論」へ

> 2025年、AIの評価軸は、既知の問題を解く能力から、人間でも数日を要する「未公開の専門家レベル問題」への挑戦へとシフトした。これは、発明や科学的発見に直結する能力を測るための必然的な進化である。

~2024

GSM8k, MATH
（中高生レベル）

2025+

FrontierMath, CritPt, ARC-AGI-2
（専門家レベル、未知の問題）

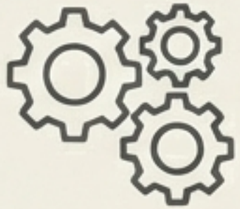
FrontierMath	CritPt	ARC-AGI-2	Humanity’s Last Exam
未発表の専門家レベル数学問題。純粋な数学的推論能力を測る。	物理学者が作成した研究レベルの複合課題。物理学的洞察力を測る。	言語に依存しない抽象的視覚パズル。深層推論が求められる。	大学院レベル以上の難問を含む総合推論テスト。

対峙する二つの知性：モデルの特性と設計思想

Gemini 3 Pro
(Google DeepMind)

天才的な研究パートナー

- ネイティブなマルチモーダル統合（画像、音声、動画、コード）
- 100万トークンの巨大コンテキストウィンドウ
- 「Deep Think」モードによる深い思考と自己検証
- Sparse MoE アーキテクチャ



GPT-5.1 Pro
(OpenAI)

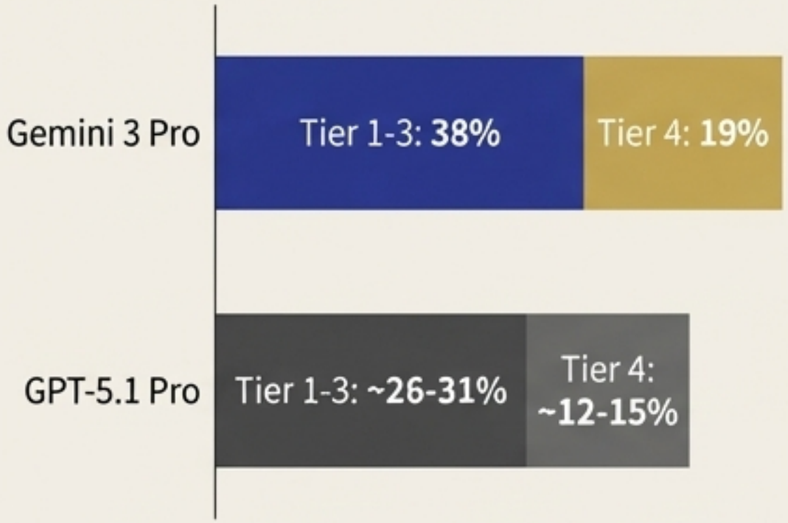
究極の熟練労働者

- 適応的推論機構（Instant/Thinkingモード）
- コンパクション技術による効率的な長文脈処理
- 12.8万〜40万トークンの標準コンテキスト長
- 安定性と再現性に優れたパフォーマンス

ROUND 1: 純粋な科学的推論能力 – Gemini、未知の問題で圧勝

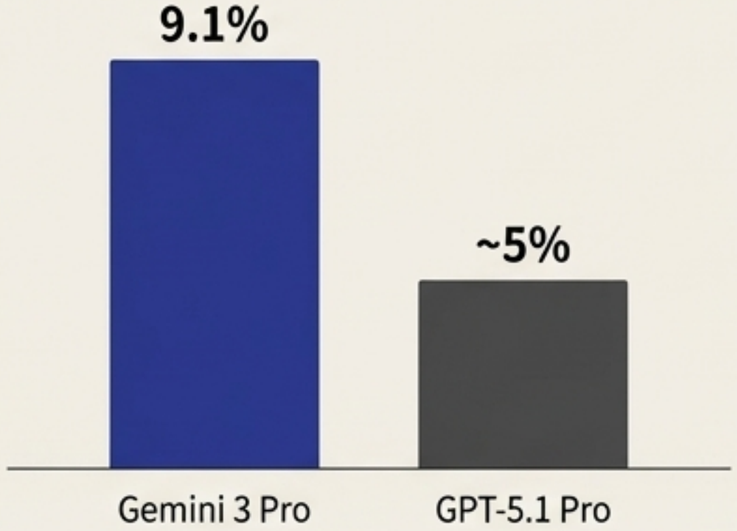
人類未踏の数学・物理学の領域において、Gemini 3 ProはGPT-5.1 Proを有意に上回る突破力を示す。

FrontierMath（最先端数学）



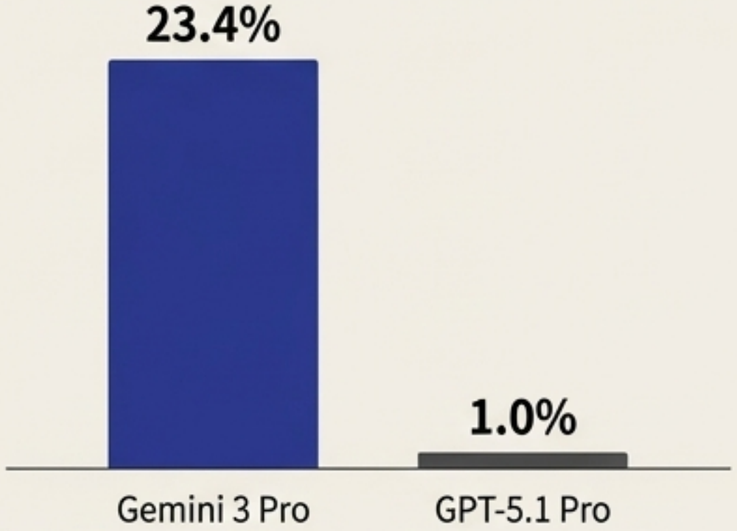
指数関数的な飛躍。未知の数学問題への対応力でGeminiが一世代先行。

CritPt（大学院レベル物理学）



多くのモデルが0%に沈む中、Geminiが約2倍のスコアを記録。限界状況での物理推論力に明確な差。

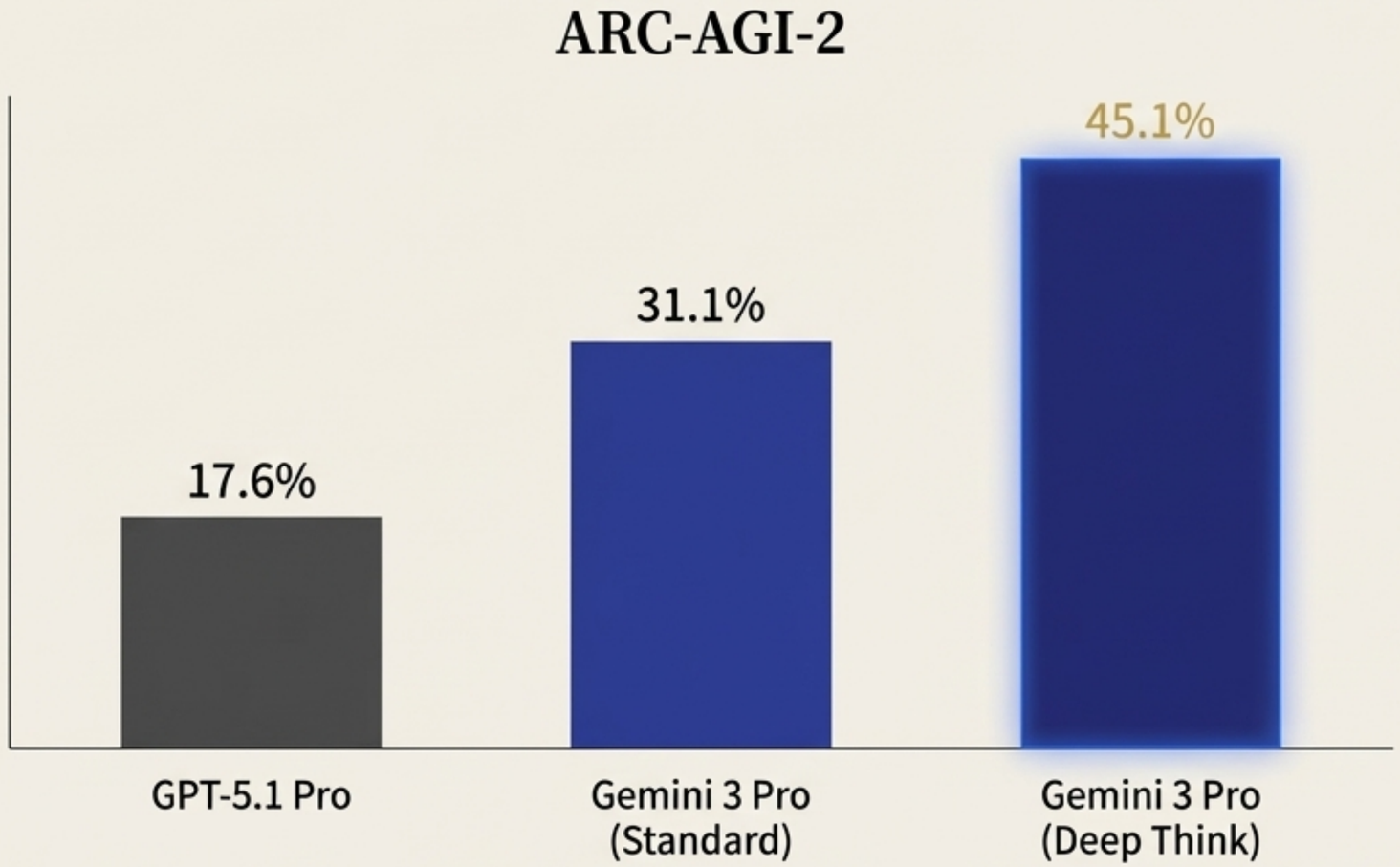
MathArena Apex（最難関競技数学）



難問への突破力で20倍以上の差。Geminiの内在的な数学的直感が突出。

ROUND 2: 抽象的推論 – 言語と思考の壁を超える

言語に依存しない視覚パズルにおいて、Gemini 3 ProはGPT-5.1 Proの2倍近いスコアを記録。
新しい問題パターンへの適応力と、真の汎用推論能力で他を圧倒する。



What it means

ARC-AGI-2は、未知の課題に対するゼロショット推論能力を測る。このスコアの差は、単なる知識量ではなく、新しい法則や関係性を自ら見出す「ひらめき」の能力差を示唆する。

Deep Thinkの威力

Deep Thinkモードとコード実行を組み合わせることで、Geminiは前例のないスコアを達成。これは、複雑な問題に対して深く、多角的に思考する能力の証明である。

この一世代でのスコアの進歩は「革命的」と評価されている。

ROUND 3: マルチモーダル理解 – 図面、動画、UIをネイティブに解釈

Gemini 3 Proは、テキスト、画像、動画、音声を統一的に処理する設計思想により、マルチモーダルタスクでGPT-5.1 Proを圧倒。これは発明に不可欠な「現象の理解」に直結する。

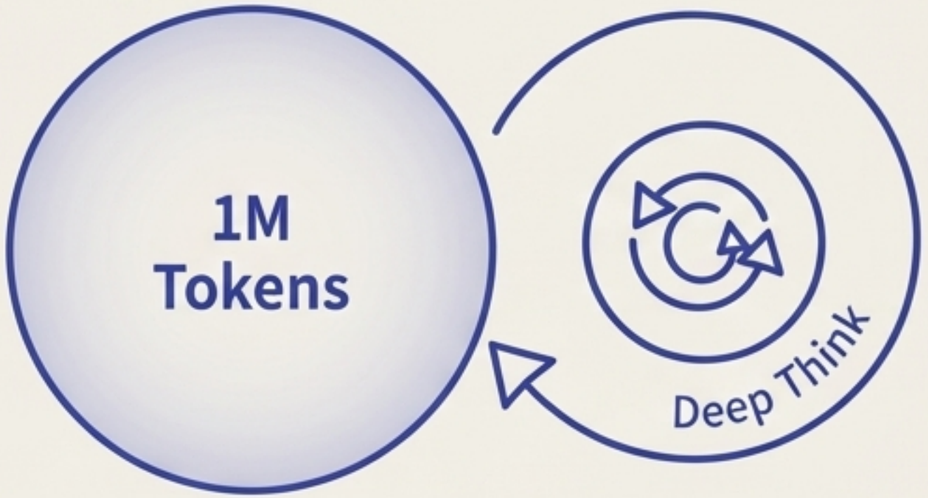
ベンチマーク	Gemini 3 Pro	GPT-5.1 Pro	考察
MMMU-Pro (画像+テキスト)	**81.0%**	~76-78%	視覚情報を含む総合推論でGeminiが5ポイントリード。
Video-MMMU (動画理解)	**87.6%**	~80-82%	時間的・空間的情報の統合能力で7ポイント以上の差。
ScreenSpot-Pro (画面理解)	**72.7%**	3.5%	UI自動化やスクリーンショット解析でGeminiが圧倒的優位。

Architectural Insight

Geminiのアーキテクチャは「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直接推論できる。GPT-51はプラグイン等で対応するが、ネイティブな統合力で差が生まれる。

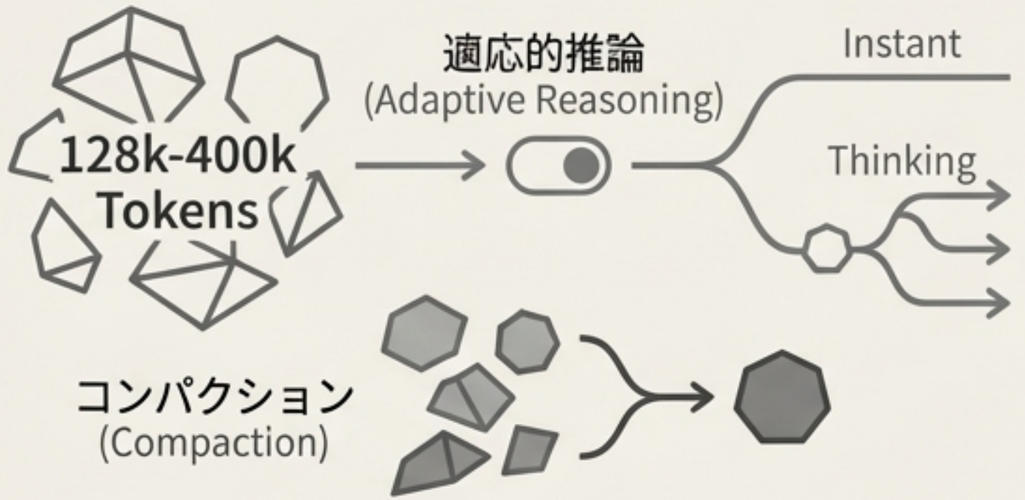
ROUND 4: 長文脈処理と思考深度 – 膨大な知識をどう扱うか

Gemini 3 Pro: 深さと広さの両立



- **Capacity:** 1,048,576トークンの巨大なコンテキストで、数百ページの文献や数万行のコードを一括処理。
- **Method:** Deep Thinkモードで、難問に対して内部で長時間の思考ループを実行し、連鎖的推論と自己検証を行う。MRCR v2ベンチマークでは長文脈下でも高い情報検索精度(77.0% @128k)を維持。

GPT-5.1 Pro: 効率性と安定性



- **Capacity:** 標準で12.8万〜40万トークン。コンパクション技術で対話履歴を要約・圧縮し、実質的な無限長コンテキストを目指す。
- **Method:** 適応的推論により、タスク難易度に応じて思考モードを自動で切り替え、計算リソースを最適化。応答の安定性と制御性に優れる。

ROUND 5: 実装と自律性 – コード生成とエージェント性能

実務的なバグ修正では両者互角。しかし、ゼロからのアルゴリズム設計や長期的な自律タスクでは、Geminiの創造性と計画能力が再び輝きを放つ。

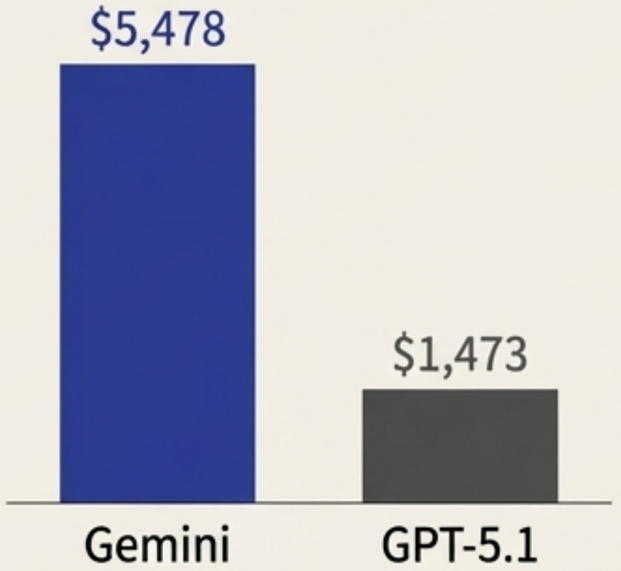
実務的コーディング (SWE-Bench Verified)

GPT-5.1はバグ修正など、既存プロセスに統合する実務で僅差ながら再現性と安定性に優れる。「熟練労働者」としての強みを発揮。

創造的コーディング (LiveCodeBench)

ゼロからアルゴリズムを設計する能力でGeminiが200ポイント近い差をつける。「天才パートナー」のひらめきがコード生成にも現れる。

長期的エージェント性能 (Vending-Bench 2)



長期的な意思決定と経済的価値創出でGeminiが**272%高い成果**。自律エージェントとして、より優れた戦略を立て実行できることを示唆。

評価の統合：二つのAIペルソナ – “天才”か、“熟練”か

ベンチマークの結果は、単なる性能差ではなく、根本的な思想と能力の方向性の違いを浮き彫りにする。

Gemini 3 Pro – “天才的な研究パートナー”

- 未知の問題に対する**創造的・飛躍的**な発想
- 抽象的・視覚的なパターン認識
- マルチモーダル情報を統合した**大局的**な理解
- 深い思考と長期的な意思決定

最適用途: 理論構築、新規アイデア創出、分野横断的な関連性の発見、自律的な科学研究

GPT-5.1 Pro – “究極の熟練労働者”

- **安定的・再現性**の高いパフォーマンス
- 既存のワークフローへのシームレスな統合
- コスト効率と高速な応答性
- 論理的な整合性と着実なタスク遂行

最適用途: 実装、プロトタイピング、バグ修正、日常的な開発支援、既存知識の整理

結論：発明の「創出」段階に最も力を発揮するのは Gemini 3 Pro

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（アイデア出し・理論構築）」フェーズにおいて、Gemini 3 ProはGPT-5.1 Proを凌駕し、現時点で最適なモデルである。

1. **卓越した創造的推論**: FrontierMath, ARC-AGI-2, CritPtで実証された、未知の問題から新しい解法を見つけ出す能力。
2. **統合されたマルチモーダル理解**: 100万トークンの文脈で図面・実験データ・動画を一体で扱い、複雑な情報を統合する力。
3. **深い思考と長期的視点**: Deep ThinkモードとVending-Bench 2が示す、多段階の熟考と戦略的計画能力。
4. **内在的な数学的直感**: AIME 2025やMathArena Apexが示す、ツールに頼らず難問を解く数学的センス。

実践的戦略：発明プロセス全体を最大化するハイブリッドアプローチ

どちらか一方を選ぶのではなく、発明のフェーズに応じて両モデルの強みを使い分けることで、アイデア創出から実装までの全過程を加速できる。

Stage 1: アイデア創出・理論構築
(Ideation & Theory)

Recommended Model:
Gemini 3 Pro

膨大な文献の読解、異分野の知見の結合、マルチモーダル情報の解析、創造的な仮説立案。

→

Stage 2: 実装・プロトタイプ作成
(Implementation & Prototyping)

Recommended Model:
GPT-5.1 Pro

安定したコード生成、バグ修正、仕様書に基づいた着実な実装、高いコスト効率。

→

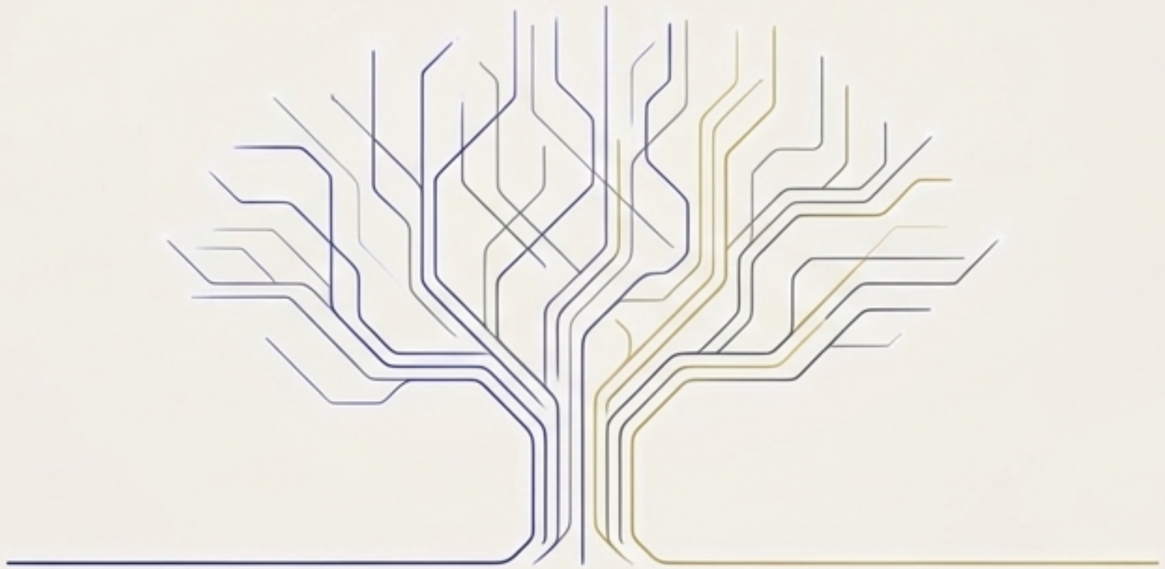
Stage 3: 長期的な自律実験・検証
(Long-term Experimentation)

Recommended Model:
Gemini 3 Pro

長期的な意思決定能力、シミュレーションでの価値創出、自律エージェントとしての運用。

競争の先にある未来：AI自身が発明する時代へ

GeminiとGPT-5.1の競争は、生成AIが単なるツールから、自律的に発見・発明を行う
「パートナー」へと進化する大きなトレンドの一部である。



特化モデルの台頭

数学の圏論に基づき、自らアルゴリズムを発明し特許を取得した
「**Categorical AI**」のような、特定領域で人間を超える能力を持つ
モデルが登場している。

閉ループ型システムへ

AIが仮説を立て、自律的に実験・検証し、知識ベースを更新する
「創発的エージェント」が科学研究の未来像として示されている。

**今日の選択は、単一モデルの優劣を決めることではない。来るべき「AIによる発明創出」の時代に向けて、
いかにして最高の「知のチーム」を構築するか、という戦略的問いかけである。**