Gemini 3.1 Pro 深堀分析レポート

エグゼクティブサマリー

2026年2月20日、Googleは**Gemini 3.1 Pro**を発表した。前世代のGemini 3 Pro（2025年11月リリース）から約3ヶ月でのアップデートであり、バージョン番号こそ「.1」だが、推論能力において劇的な向上を遂げている。最大の注目点は、抽象推論ベンチマーク**ARC-AGI-2で77.1%**（検証済みスコア）を達成し、Gemini 3 Proの31.1%から**+148%**という単一世代での史上最大の向上を記録したことである。[1][2]

---

1. モデル概要と技術仕様

Gemini 3.1 Proは、Gemini 3 Proをベースにコア推論能力を重点的に強化したモデルである。Googleは「単純な答えでは不十分なタスクのために設計された」と説明している。[3]

技術仕様

項目	仕様
モデルID	`gemini-3.1-pro-preview`
コンテキストウィンドウ	約100万トークン（入力上限 1,048,576）[4][3]
出力トークン上限	65,536トークン（約64K）[3]
知識カットオフ	2025年1月[3]
入力モダリティ	テキスト、画像、動画、音声、PDF[3]
出力モダリティ	テキストのみ[3]
対応機能	Thinking、Function Calling、構造化出力、検索グラウンディング[3]

Google DeepMindのモデルカードによれば、アーキテクチャはGemini 3 Proに基づいており、エージェント性能、高度なコーディング、長文コンテキスト・マルチモーダル理解、アルゴリズム開発に特に適したモデルと位置づけられている。[5]

提供プラットフォーム

Geminiアプリ（Pro/Ultraプラン）、Google AI Studio、Vertex AI、Gemini API、Google Antigravity、NotebookLM、Gemini CLI、Android Studioで利用可能。**Free tierでは利用不可**である点に注意が必要。[6][3]

---

2. ARC-AGI-2ベンチマークの意義 —— なぜ"AGIへの登竜門"と呼ばれるか

ARC-AGI-2とは何か

ARC-AGI-2（Abstraction and Reasoning Corpus for AGI 2）は、ARC Prize Foundationが2025年3月に公開したベンチマークで、François Chollet（Keras開発者）らが設計した。従来のAIベンチマークが暗記・パターンマッチングで高スコアを取れる問題を抱えていたのに対し、ARC-AGI-2は**「人間には簡単だがAIには困難な」抽象推論タスク**を集めた点が革新的である。[7][8][9]

具体的には、色付きセルのグリッド変換タスクを2〜5個の例示ペアから規則を推論し、未知のテストケースに適用するという形式をとる。タスクは以下の3つの難所を内包する：[9]

- **記号解釈（Symbolic Interpretation）**：表面的な対称性ではなく、記号の意味を理解する必要がある[7]
- **合成的推論（Compositional Reasoning）**：複数のルールを同時に適用・組み合わせる能力[7]
- **文脈依存的規則適用（Contextual Rule Application）**：グリッド内の文脈的手がかりに応じて規則の適用方法を変える[9]

なぜ"AGIへの登竜門"なのか

François Cholletは、ARC-AGI-2を「AGIへの進歩を厳格に測定する次世代ツール」と位置づけている。前世代のARC-AGI-1では、計算力による総当たり探索で約50%のタスクが解けてしまう問題があったが、ARC-AGI-2はこれを排除し、**効率的な抽象推論能力**そのものを測定する設計に進化した。[9][10][11][12]

評価プロトコルも厳格で、Kaggleのセキュアなサンドボックス環境（NVIDIA L4 GPU×4台、12時間制限、インターネット接続なし）で240の未知タスクを解かせる。高スコアを出すためには「暗記」ではなく、真の汎化能力が求められる。[11][9]

人間の平均スコアは約60%（公開評価セットでは66%）であり、Gemini 3.1 Proの77.1%は人間の平均を上回る水準に達している。[8]

---

3. ベンチマーク結果の詳細分析

ARC-AGI-2スコア比較

Gemini 3.1 Proの77.1%は、Claude Opus 4.6（68.8%）を8.3ポイント、GPT-5.2（52.9%）を24.2ポイント上回る。Gemini 3 Pro（31.1%）からの+46ポイントの向上は、**主要ベンチマークにおける単一世代での最大の改善幅**とされている。[1]

主要ベンチマーク横断比較

ベンチマーク	Gemini 3.1 Pro	Claude Opus 4.6	GPT-5.2	備考
ARC-AGI-2（抽象推論）	**77.1%**	68.8%	52.9%	Geminiが大差でリード[1]
GPQA Diamond（PhD級科学知識）	**94.3%**	91.3%	92.4%	全モデル最高[1]
SWE-Bench Verified（バグ修正）	80.6%	**80.8%**	–	Claude Opusが0.2pt差でリード[3]
HLE（ツールなし）	**44.4%**	40.0%	34.5%	総合知識でGeminiリード[1]
HLE（ツール付き）	51.4%	**53.1%**	–	ツール活用ではClaudeに劣後[1]
BrowseComp（エージェント検索）	**85.9%**	84.0%	–	Webブラウジング能力でトップ[1]
MMMLU（多言語QA）	**92.6%**	–	–	多言語理解でも高水準[3]
Terminal-Bench 2.0	68.5%	65.4%	54.0%	GPT-5.3 Codexの77.3%には及ばず[3]
LiveCodeBench Pro	**2887 Elo**	–	2393 Elo	約500 Elo差でGeminiリード[1]

APEX-Agents	**33.5%**	29.8%	23.0%	エージェントタスクでトップ[13]

総合評価： 比較可能な13ベンチマーク中10でClaude Opus 4.6を上回る。Claude Opusはツール活用型タスク（HLE with tools、SWE-Bench Verified）で僅差のリードを保つが、推論・エージェント系ではGeminiの優位が明確である。[1]

世代間の改善幅

Gemini 3 Pro → 3.1 Proの主な改善：

- ARC-AGI-2：31.1% → 77.1%（**+148%**）[1]
- BrowseComp：59.2% → 85.9%（**+45%**）[1]
- SWE-Bench Pro：43.3% → 54.2%（+25%）[1]
- LiveCodeBench Pro：2439 → 2887 Elo（+448 Elo）[1]
- Terminal-Bench 2.0：56.9% → 68.5%（+20%）[1]

---

4. 長文コンテキストの重大な課題

仕様上は100万トークンのコンテキストウィンドウに対応しているが、**超長文脈での情報抽出精度に深刻な課題**が確認されている。

MRCR v2（8-needle）テスト結果：[3]

- **128Kトークン**：84.9%（高精度を維持）
- **100万トークン**：**26.3%**（精度が急落）

これは、長大なドキュメントやコードベースを一括処理する実務用途において大きな制約となる。特許文書の一括分析や大規模コードベースのレビューなど、100万トークン級の入力を想定する用途では、この精度劣化を考慮した設計が不可欠である。[3]

---

5. 新機能と実用的改善点

コーディング能力の強化

Googleはこのモデルを「Vibe Codingとエージェント型コーディングのための最高のモデル」と位置づけている。主な改善点：[3]

- **ツール使用の精度向上**とマルチステップタスクの同時処理能力
- **指示追従性の改善**（プロンプトの意図をより正確に把握）
- ソフトウェアエンジニアリングにおける**マルチステップ実行の信頼性**向上
- テキストプロンプトからの**アニメーションSVG生成**能力[2]

出力完遂能力の改善

前モデルのGemini 3 Proは約21Kトークンで出力が途切れる傾向があったが、Gemini 3.1 Proでは約48,000トークンのコードベースに対して55,000トークン超の完全な出力が可能になった。出力上限値自体（64K〜65K）は変わっていないが、**上限近くまで実際に出力を完遂する能力**が向上した。[3]

ビジュアル・クリエイティブ領域

複雑なAPIとユーザーフレンドリーなUIの間を橋渡しするコード生成（例：ISS軌道可視化ダッシュボード）、3Dムクドリの群舞（マーマレーション）のインタラクティブ体験生成、文学作品のテーマからのポートフォリオサイト構築など、高度なクリエイティブコーディング能力が実証されている。[2]

---

6. API価格と競合比較

価格体系（1Mトークンあたり）

項目	Gemini 3.1 Pro	Claude Opus 4.6	GPT-5.2	Gemini 3 Flash
入力（≤200K）	$2.00	$5.00	–	$0.50
入力（>200K）	$4.00	–	–	–
出力（≤200K）	$12.00	$25.00	–	$3.00
出力（>200K）	$18.00	–	–	–

Claude Opus 4.6と比較すると、入力は**40%**、出力は**48%の価格で済む。月間10億トークン処理の試算では、Gemini 3.1 Proが約$14,000、Claude Opus 4.6が約$90,000と約6.4倍の差**がつく。[13][3]

ただし、Gemini 3.1 Proは**出力が冗長になる傾向**があるとの報告もあり、実コストはトークン単価だけでは測れない点に注意が必要である。[3]

バッチ利用は標準価格の50%、コンテキストキャッシングは$0.20〜$0.40/1Mトークン（ストレージ$4.50/時間）で利用可能。[3]

---

7. 実テストからの知見（4モデル比較検証）

ChatGPT Labが実施した4モデル同一プロンプト比較テスト（Gemini 3.1 Pro / Claude Opus 4.6 / GPT-5.3 Codex / GPT-5.2 xhigh）では、以下の傾向が確認された：[3]

テスト内容	Gemini 3.1 Proの評価
LPデザイン（和モダン茶室）	Claude Opus 4.6に匹敵、部分的に上回るUI生成
SVGアニメーション	動きの滑らかさとディテールで印象的
3Dゲーム生成	**エラー発生、動作せず**
クリエイティブライティング（日本語）	モデル間で個性が分かれる
哲学的思考	思考の深さと独創性が試される

ビジュアル面では一貫して強いが、**複雑なロジックを含むタスク（3Dゲーム等）では安定性に課題**がある。また、Gemini App（ウェブ版）ではGoogle AI Studioと比較して著しく品質が低下する場合がある点も注意すべきである。[3]

---

8. 市場の反応と開発者コミュニティの評価

ポジティブな評価

- ビジュアル・クリエイティブ領域での顕著な進化（アニメーション、UI設計）[3]
- one-shotでのWeb OS生成や都市計画アプリ構築など実デモの共有[3]
- Claude Opus 4.6の約半額で同等クラスの性能というコストパフォーマンス[1][13]

- 13ベンチマーク中10でClaude Opus 4.6を上回る総合性能[1]

懸念・注意点

- 同じプロンプトで公式デモを再現できないケースの報告[3]
- プレビュー版の「ナーフ」（正式リリース後の性能低下）への警戒感[3]
- エージェント型の長期ワークフローでは「まだClaudeの方が安定」との声[3]
- 100万トークンコンテキストでの精度急落（84.9% → 26.3%）[3]
- 出力の冗長化傾向[3]

---

9. フロンティアモデル競争の現在地

各社の強み分布

領域	リーダー	備考
抽象推論・汎化	**Gemini 3.1 Pro**	ARC-AGI-2で圧倒的リード[1]
科学的推論	**Gemini 3.1 Pro**	GPQA Diamond 94.3%[1]
コーディング（SWE-Bench）	Claude Opus 4.6 ≈ Gemini 3.1 Pro	0.2pt差で拮抗[3]
ターミナル操作	**GPT-5.3 Codex**	77.3%でリード[3]
エージェントタスク	**Gemini 3.1 Pro**	APEX-Agents 33.5%[13]
ツール活用型推論	**Claude Opus 4.6**	HLE with toolsで優位[1]
コスト効率	**Gemini 3.1 Pro**	Opus比で入力40%、出力48%[3]
マルチモーダル	**Gemini 3.1 Pro**	ネイティブマルチモーダルアーキテクチャ[14]

Gemini 3.1 Proの登場により、フロンティアモデル間の競争は「単一モデルの総合優位」から「領域別の強み分布」へとさらに複雑化している。推論・エージェント・コスト効率でGoogleが優位に立つ一方、実務的なソフトウェアエンジニアリングやツール活用ではAnthropicが僅差で粘り、ターミナル操作ではOpenAIが明確なリードを維持している。

---

10. 関連情報：Gemini 2.0 Flashの廃止予告

2026年2月18日、Googleは以下のGemini 2.0 Flashモデルを**2026年6月1日にシャットダウン**すると予告している：[3]

- `gemini-2.0-flash` / `gemini-2.0-flash-001`
- `gemini-2.0-flash-lite` / `gemini-2.0-flash-lite-001`

該当モデルを利用中のシステムは、Gemini 3 Flash Preview等への移行計画が必要である。

---

11. IP・技術戦略の観点からのインプリケーション

特許分析・知財戦略への示唆

1. **推論能力の飛躍的向上**：ARC-AGI-2でのスコア急上昇は、Googleが抽象推論のアーキテクチャ・学習手法に**根本的な技術革新**を実現した可能性を示唆する。単なるスケーリングではなく、新たな特許出願の対象となり得る独自の技術的アプローチが背景にあると考えられる。[1]

2. **「Vibe Coding」市場の形成**：自然言語からの直接的なアプリケーション構築は、ソフトウェア開発プロセスの根本的変革を意味し、関連する特許ポートフォリオの構築が各社で加速すると予想される。

3. **長文コンテキストの精度課題**：100万トークン対応を謳いながら精度が急落する点は、RAG（Retrieval Augmented Generation）技術の価値が引き続き高いことを意味する。長文コンテキスト処理に関連する特許技術の重要性は維持される。

4. **マルチモーダル推論の発展**：テキスト・画像・動画・音声・PDFの統合処理能力は、特許文書分析、先行技術調査、非侵害設計分析等のIP実務において活用可能性が広がる方向にある。

5. **Frontier Safety Framework**：DeepMindはCBRN、サイバー、有害操作、機械学習R&D、ミスアラインメントの5リスク領域で評価を実施し、CCL（Critical Capability Level）以下であることを確認している。AI安全性に関する規制・ガバナンスの枠組み構築において参照されるべき枠組みである。[5]

References

1. Gemini 3.1 Pro: 77% ARC-AGI-2 Score & Benchmarks (2026) - Google Gemini 3.1 Pro released today with 77.1% ARC-AGI-2 - doubling its predecessor. Full benchmark...

2. Gemini 3.1 Proを発表 - ARC-AGI-2で前世代比2倍超の推論性能を達成 - GoogleがARC-AGI-2で77.1%を達成した「Gemini 3.1 Pro」を発表。前世代比2倍超の推論性能向上が確認され、API料金は据え置きでGemini APIとVertex AI経由...

3. 【徹底解説】Gemini 3.1 Pro 登場。実力を4モデル比較で検証 - コンテキストウィンドウ（入力）: 約100万トークン · 出力トークン上限: 65,536トークン（約64K） · 知識カットオフ: 2025年1月 · 入力: テキスト、画像、動画、 ...

4. そっとリリース！Gemini 3.1 Proとは？強み（推論力・長文処理 ... - ... 100万）トークンのコンテキストに対応します。[S2] （APIのモデル仕様 ... （APIのモデル仕様では、入力トークン上限が 1,048,576、出力トークン上限が 65,536 ...

5. Gemini 3.1 Pro - Model Card - (Deep Think mode) The model shows gains on RE-Bench compared to Gemini 3 Pro, with a human-normalise...

6. Google、強化された推論機能を備えたGemini 3.1 Proを発表 - 新しい論理パターンを解決するモデルの能力をテストするARC-AGI-2ベンチマークにおいて、Gemini 3.1 Proは77.1%という検証済みスコアを達成し、前 ...

7. The Ultimate Benchmark for AI Intelligence ARC Prize ... - One of the defining features of ARC-AGI-2 is its ability to remain solvable by humans while posing s...

8. ARC-AGI 2 - AI Wiki - Artificial Intelligence Wiki

9. ARC-AGI-2: A New Challenge for Frontier AI Reasoning ...

10. [2505.11831] ARC-AGI-2: A New Challenge for Frontier AI ... - F Chollet 著 · 2025 · 被引用数: 63 - ARC-AGI-2 aims to serve as a next-generation tool for rigorously mea...

11. ARC-AGI In 2026: Why Frontier Models Still Don't Generalize - ARC-AGI-2 exposes the real gap: generalization efficiency under budget constraints, where refinement...

12. ARC-AGI-2 Overview With François Chollet - Effective Altruism Forum - A talk by AI researcher François Chollet. He discusses the ARC-AGI benchmark, which was created in 2...

13. Gemini 3.1 Pro vs Claude Opus 4.6 vs GPT-5.2 - Takeaway: Gemini 3.1 Pro leads on raw reasoning and scientific knowledge. Claude Opus 4.6 edges ahea...

14. GPT-5 vs Claude Opus 5 vs Gemini 3 Ultra - Humai.blog - Confused about GPT-5, Claude Opus 4.5, and Gemini 3 Pro? I spent 3 months testing all three AI model...