Google Gemini 3.1 Pro: A Deep Dive (Preview)

Noto Sans JP

推論能力の飛躍と実務実装への示唆

Noto Sans JP
社内検討用・技術評価レポート

エグゼクティブ・サマリー：誇大広告と実力の選別

発表 (LAUNCH)

2026年2月19日、Gemini 3系の改良版として「Preview」リリース。開発者およびVertex AI向けに展開開始。

性能の飛躍 (PERFORMANCE LEAP)

抽象推論ベンチマーク「ARC-AGI-2」でスコア **77.1%** を記録（前モデル31.1%から**2倍以上の向上**）。

注意点 (CAVEATS)

競合他社スコアは「**自己申告値**」が含まれる。また、**1Mトークンの長文脈処理は精度低下のリスクあり**(MRCR v2)。

推奨アクション (VERDICT)

R&Dや複雑なエージェント構築には最適。ただし「Preview」ステータスのため、SLAが厳格な本番環境への**投入にはガバナンスが必要**。

抽象推論における「2倍」の飛躍

ARC-AGI-2とは？: 知識の記憶ではなく、未知のロジックパターンに適応する能力を測定する「AGI判定のリトマス試験紙」に近い指標。

出典: Google DeepMind Model Card / ARC Prize Verified

ベンチマークの詳細と「自己申告」の罠

Category	Gemini 3.1 Pro	Top Competitor
Agentic Coding (SWE-Bench Verified)	80.6%	Claude Opus 4.6: **80.8%**
Science (GPQA Diamond)	**94.3%**	GPT-5.2: 92.4%
Math/Reasoning (HLE - No Tools)	**44.4%**	Claude Opus 4.6: 40.0%

Key Insight
コーディング（SWE-Bench）ではClaude Opus 4.6と拮抗。
科学・数学推論ではGeminiがリード。

※ 比較表の多くは他社の「最大 thinking 設定」や「自己申告値」に基づくため、自社データでの検証が不可欠。

Technical Editorial

AGIへの到達か、それとも「汎用推論」の進展か？

ARC Prizeの見解

ARC-AGI高得点は「AGI到達」の十分条件ではない。インターネット・計算資源の制約がない環境でのスコアであることに注意。

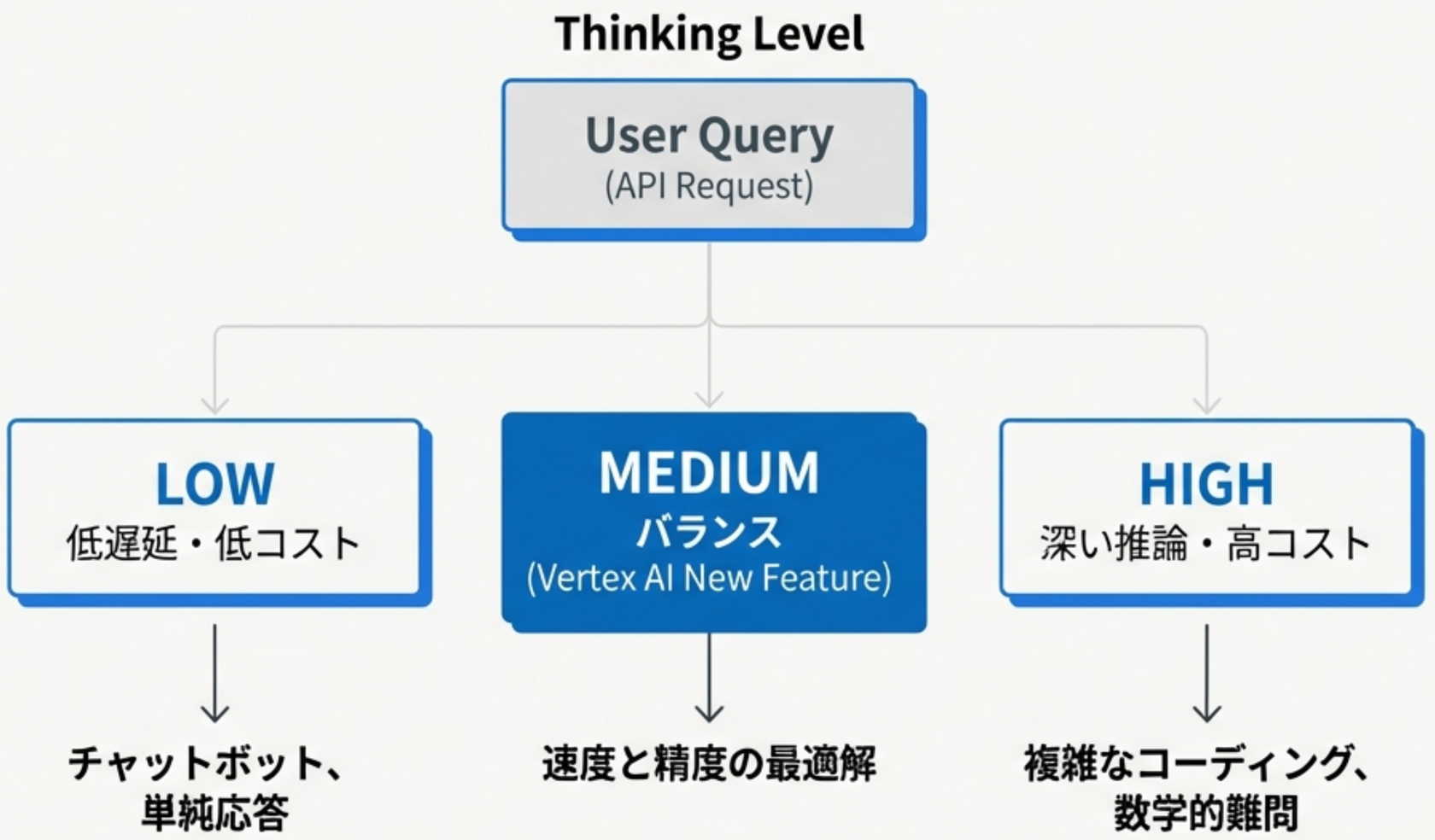
HLE（Humanity's Last Exam）

高得点は「閉じたタスクでの専門家レベル」を意味するが、自律的な研究能力を示すものではない。

結論: 「汎用推論」の強力なシグナルだが、完全自律型AGIへの到達と断定するには論拠不足。

Technical Editorial

「思考（Thinking）」を制御する：コストと精度のトレードオフ



Note

デフォルトで 'Dynamic Thinking' が適用されるが、API/Vertex AI側で明示的に制御可能。

Technical Editorial

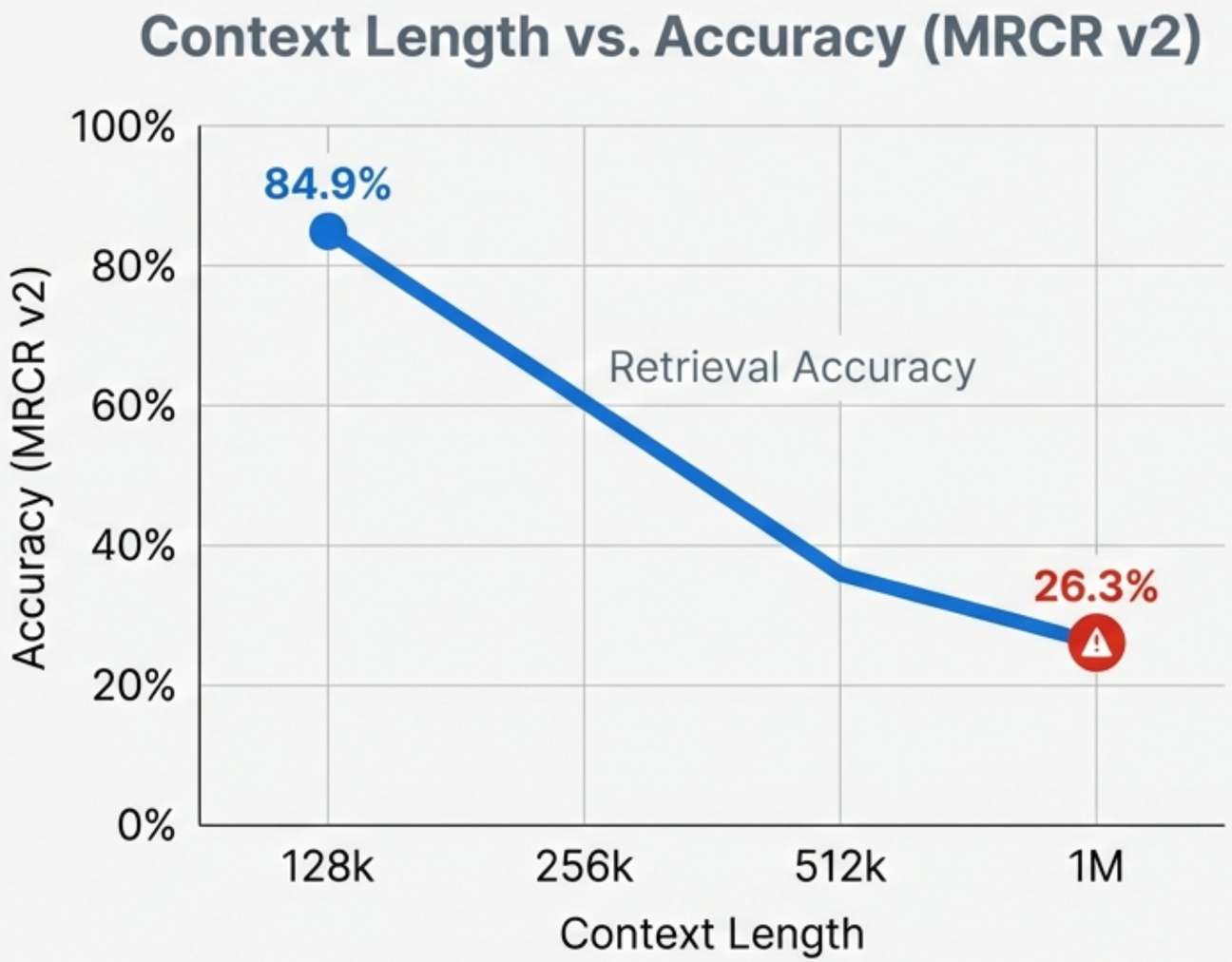
基本仕様と提供形態

Model ID	Input Modalities	Output Modalities
`gemini-3.1-pro-preview`	 Text, Image, Video, Audio, PDF	Text Only ※画像・音声生成は未サポート

Context Window	Max Output	Knowledge Cutoff
1,048,576 tokens (~1M)	65,536 tokens	2025年1月

ステータス: Public Preview (レート制限が厳格な可能性あり)

「100万トークン」の現実と限界

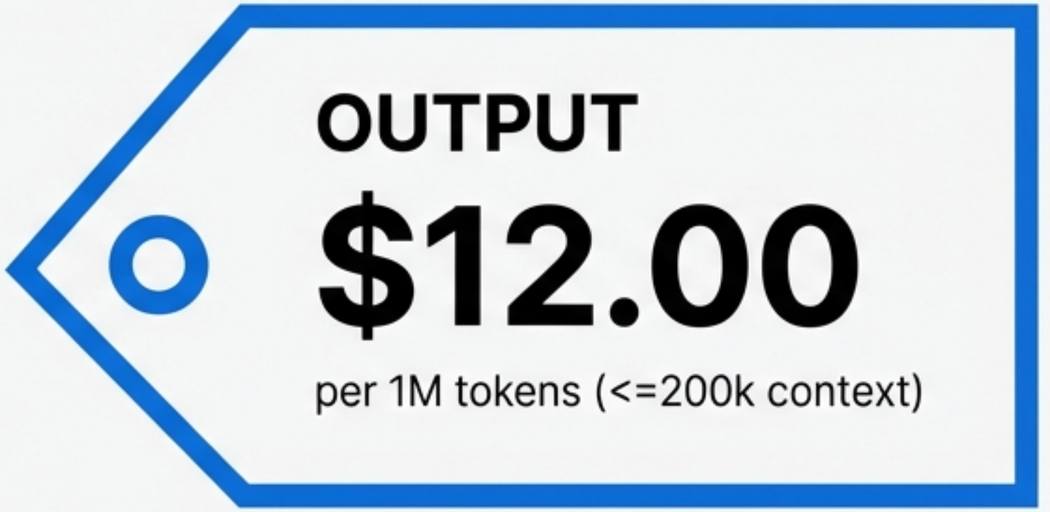


Key Takeaway: コンテキスト一杯にデータを詰め込むと精度が劇的に低下する。

Action: 1M入力を常時使用せず、RAG（検索拡張生成）、キャッシュ、データの段階分割を併用する設計が必須。

Technical Editorial

価格戦略とは、価格戦略とコスト対効果



- Batch API: 概ね半額水準 (approx. 50% off)
- **Value Proposition**: Gemini 3 Pro Previewと同水準の価格設定。つまり、「コスト増なし」で推論能力の向上が可能。

⚠ **※Thinkingトークン（思考プロセス）も出力課金に含まれるため、High設定時の総トークン量に注意。**

Technical Editorial

実務での推奨ユースケース

Dynamic Visualization

テキスト指示から SVG/ダッシュボードを即座に生成。

構造化出力と推論能力の結合。

Agentic Workflows

調査 -> 統合 -> 可視化の一連フローを自律実行。

APEX-Agentsベンチマークでのスコア向上 (**18.4% -> 33.5%**)。

Legacy Code Analysis

大規模リポジトリの解析とドキュメント化。

1Mコンテキスト**（注意して使用）**とネイティブ⚠マルチモーダルの活用。

Technical Editorial

競合ポジショニング

Gemini 3.1 Pro (The Specialist)	Claude Opus 4.6 (The Artisan)	GPT-Series (The Ecosystem)
• **推論コストパフォーマンス、ネイティブマルチモーダル入力。**	• コーディングの微細なニュアンス、人間らしいトーン。	• エコシステム統合、汎用性。
Best for: コスパ重視の推論エージェント、大量データの処理。	Best for: 複雑なソフトウェア開発、対話品質重視。	

結論: 「推論単価」と「入力の柔軟性」においてGemini 3.1が優位。

Technical Editorial

ガバナンスとリスク管理

規制環境 (Regulatory Context)

- **EU AI Act**: **GPAI**（汎用AI）義務が2025年8月から適用。システミックリスク評価が必要。
- **NIST AI RMF**: ベンチマークだけでなく、運用監視と説明責任の実装が推奨される。

運用リスク (Operational Risk)

- **Preview Status**: レート制限が厳しく、仕様変更の可能性があるため、クリティカルなシステムへの即時導入は**リスクあり**。
- **Human-in-the-Loop**: 重要な意思決定には必ず**人間による確認プロセスを挟むこと**。

Technical Editorial

導入に向けた3段階のロードマップ

1. Verify (検証)

公開ベンチマークを鵜呑みにせず、自社データと評価セットで性能を再検証する。特にSWE（コーディング）クンタスクにおいて。

2. Architect (設計)

コスト最適化のため、基本は「Medium Thinking」を採用し、難問のみ「High」に切り替えるルーティングを設計する。

3. Monitor (監視)

Previewモデル特有のレート制限（RPM/TPM）を監視し、フォールバック（予備）体制を構築する。

回避すべきアンチパターン

1Mコンテキストへの過信 (Over-reliance on 1M Context)

理由: 精度が26.3%まで低下するため、全データを投入しての「完璧な検索」を期待してはならない。

SLA厳守システムへのPreview投入 (Preview in Critical SLA)

理由: 安定性とレート制限が保証されていない。

完全自律化への期待 (Expecting Full Autonomy)

理由: HLEスコアはあくまで「閉じたタスク」での性能。自律的な研究や長期計画能力はまだ人間による補完が必要。

結論：実用的な革命

Conclusion: A Pragmatic Revolution

ARC-AGI-2の飛躍は本物だが、魔法ではない。成功の鍵は、APIを叩くだけでなく、ビジネス課題に合わせて「**思考レベル（Thinking Level）**」を適切に設計・実装することにある。