Google I/O 2026：Gemini 3.5 FlashとAntigravity 2.0が切り拓く「エージェンティックAI」の深層と市場評価

Gemini 3.1 pro

1. 序論：Google I/O 2026が示すAIパラダイムの歴史的転換

2026年5月19日、Googleはカリフォルニア州マウンテンビューで開催された年次開発者会議「Google I/O 2026」において、同社の人工知能戦略における極めて重要なマイルストーンとなる最新鋭モデル群「Gemini 3.5」ファミリー、および物理世界をシミュレートする次世代マルチモーダル基盤「Gemini Omni」を大々的に発表した[1]。このカンファレンス全体を貫く最大の焦点は、AIとのインタラクションを、従来の「プロンプト入力とテキスト応答」という受動的な対話モデルから、AIが自律的に計画を立て、ツールを駆使し、非同期で複雑なタスクを完遂する「エージェンティック・ワークフロー（自律的行動プロセス）」へと完全に移行させるという、明確なビジョンの提示であった[1]。

Google CEOのSundar Pichai氏は基調講演の中で、このエージェンティックな未来の先陣を切る存在として「Gemini 3.5 Flash」の即日提供開始を宣言した[5]。これまで「Flash」という名称は、フラッグシップモデルである「Pro」シリーズに対して、知能や推論能力をある程度妥協する代わりに、圧倒的な生成速度と低コストを提供する軽量モデルの代名詞として機能してきた。しかし、今回発表されたGemini 3.5 Flashは、複雑なエージェント・タスクやコーディング・ベンチマークにおいて、前世代の最上位モデルである「Gemini 3.1 Pro」を凌駕する高度な知能を獲得しながら、他社の最先端フロンティアモデルの4倍という驚異的な処理速度を実現するという、これまでの常識を覆す進化を遂げている[2]。

また、来る6月には、さらに高度な推論能力を備えた「Gemini 3.5 Pro」のリリースも予告されており、現在社内でテスト中であることが明かされた[5]。本レポートでは、このAI業界の特異点とも言えるGemini 3.5 Flashの技術的仕様、推定されるアーキテクチャ、APIの詳細な変更点、そして同時発表された革新的なエージェント開発プラットフォーム「Antigravity 2.0」の全貌を徹底的に解き明かす。さらに、市場における肯定的な評価と、大幅な価格改定およびトークン消費量の増加に伴う開発者コミュニティからの厳しい批判的考察の双方を包括的に分析し、同モデルがAI開発エコシステム全体に与える波及効果を提示する。

2. AIアーキテクチャの進化：Gemini Omniによる「世界モデル」の確立

Gemini 3.5 Flashの分析に入る前に、Google I/O 2026におけるもう一つの巨大なブレイクスルーである「Gemini Omni」の存在に触れる必要がある[1]。DeepMindのCEOであるDemis Hassabis氏によって披露されたこの新モデルは、汎用人工知能（AGI）に向けた「世界モデル（World Model）」としての側面を強く打ち出している[3]。

2.1. 推論と生成の統合メカニズム

Gemini Omniは、Googleがこれまでに開発してきた「Veo（動画生成）」「Nano Banana」「Genie」と

いった多様な生成メディアモデルの推論能力を、一つの巨大なニューラルネットワークへと統合したモデルファミリーである[3]。Hassabis氏の解説によれば、Omniは単にテキストと画像を結びつけるだけでなく、運動エネルギー（Kinetic energy）や重力（Gravity）といった物理世界の概念を深く理解し、シミュレートする能力を備えている[3]。これにより、複雑なアイデアを極めて正確な視覚的表現へと翻訳することが可能となった。Googleは初期段階からマルチモーダル基盤としてGeminiを設計してきたが、その「より困難な道（Harder path）」が結実したのがこのOmniであると説明されている[3]。

2.2. 生成能力と編集能力のデモンストレーション

基調講演で行われたデモンストレーションは、Omniの能力の異次元さを示すものであった。Hassabis氏が「タンパク質の折りたたみ構造（Protein folding）についてのクレイアニメーションによる解説動画を作成せよ」とプロンプトを入力すると、AIは科学的に正確でありながら、指定されたストップモーション・スタイルのリアルな教育ビデオを即座に生成した[3]。

さらに、Omniは自然言語による高度な動画編集機能も有している[3]。ユーザーが自撮り動画（Selfie video）をアップロードし、テキストで指示を与えることで、シーンの芸術的スタイルを変更したり、環境内に全く新しい要素を追加したりすることが可能である。デモでは、動画内の単純な円を「ブラックホール」に変換したり、静寂な夕暮れの散歩風景に生命を吹き込んだりする様子が披露され、「あらゆるものが全く新しい現実を創り出すキャンバスになる」と強調された[3]。このOmniファミリーの最初のモデルとして、「Gemini Omni Flash」が本日よりGoogle製品スイート全体で利用可能となっている[3]。

また、Googleのコアとなるエコシステムもこの AI技術によって大幅な刷新を受けている。イベントでは、Google検索にとって過去25年間で最大級のアップグレードとなるAIベースの検索機能の強化、エコシステム全体にAIショッピング体験をもたらす「Universal Cart」、さらには再びフォーカスが当てられたスマートグラスおよびAndroid XRの展開、生成AIコンテンツの透かし技術「SynthID」の全製品への拡張、そしてスマートウォッチ向けの「Wear OS 7」の発表など、ハードウェアとソフトウェアの両面でAIの統合が極限まで推し進められていることが示された[1]。

3. 先駆的モデル「Gemini 3.5 Flash」の基盤技術と推定アーキテクチャ

Googleの「知能とアクションの融合」戦略を牽引する中核エンジンが、今回一般公開（GA）されたGemini 3.5 Flashである[8]。このモデルはプレビュー期間をスキップして即座にスケール可能な本番環境用として展開され、一般ユーザーからエンタープライズの高度な開発環境まで、あらゆるレイヤーのデフォルトモデルに指定された[8]。

3.1. 推定されるハードウェア構成とパラメータ規模

GoogleはGemini 3.5 Flashの正確な内部パラメータ数を公式には開示していない。しかし、技術者コミュニティ（Hacker Newsなど）において、リバースエンジニアリングによる推計が行われている[9]。ユーザーeasygenesの計算によれば、同モデルのフットプリントは公開されているハードウェア性能データとGoogleのサービングパラメータからある程度予測可能である[9]。

Gemini 3.5 Flashは、Googleの最新アクセラレータである「TPU 8i」ハードウェア上でサービングされていると推測されている[9]。TPU 8i（VRAM 288GBを想定）のメモリ帯域幅、演算性能、および約280トークン/秒というGoogleの出力速度から逆算すると、同モデルの総パラメータ数は約2,500億（

250B）から3,000億（300B）に達すると推定される[9]。さらに、TurboQuantのような高度な圧縮および量子化技術が適用されている場合、推論品質を損なうことなく総パラメータ数は最大で4,000億（400B）規模に達する可能性がある[9]。
TPU 8iにおけるメモリフットプリントの割り当て推計は以下の通りである。

- 静的モデル重み（**Static Model Weights**）: 110GB～150GBと推定[9]。
- 動的割り当ておよび圧縮**KV**キャッシュ: 138GB～178GBと推定[9]。KVキャッシュはAIが過去の文脈を効率的に参照するための中間データを保持する領域であり、後述する長大なコンテキストウィンドウを処理する上で極めて重要な役割を果たす。

特筆すべきは、総パラメータ数が数千億規模であるにもかかわらず、推論時に実際に「アクティブ」になるパラメータ数が約100億（10B）から160億（16B）程度に抑えられている点である[9]。これは、Gemini 3.5 Flashが、入力内容に応じて特定の専門ネットワーク（エキスパート）のみを駆動させる高度な「Mixture of Experts（MoE）」アーキテクチャを採用していることを強く示唆している。この極端に低いアクティブ・パラメータ比率こそが、フラグシップ級の推論能力を保ちながら、前例のない生成速度を実現している根幹のメカニズムである。

3.2. 入出力仕様とコンテキスト処理能力

Gemini 3.5 Flashは、完全なマルチモーダルモデルとして設計されており、長平滑（Long Horizon）なエージェンティック・タスクを遂行するための長大なコンテキストウィンドウを備えている。

項目	詳細仕様
モデル**ID**	gemini-3.5-flash [8]
入力モダリティ	テキスト、画像、動画、音声、コード、PDF [10]
出力モダリティ	テキスト（推論の思考プロセスも含む） [10]
最大入力トークン数	1,048,576トークン（約100万トークン） [8]
最大出力トークン数	65,535トークン [8]
ナレッジカットオフ	2025年1月 [10]

約100万トークンという長大なコンテキストウィンドウは、開発者が長大なソースコードのリポジトリ、膨大なシステムログ、あるいは数十ページに及ぶAPIドキュメントを丸ごとプロンプトに含めることを可能にする。これにより、エージェントは外部ツールへの不要なアクセスを減らし、文脈を見失うことなく多段階のタスクを連続的に処理できる。なお、現時点では「コンピュータの直接操作（Computer Use）」機能はサポートされていない[8]。

4. 高度な推論制御メカニズムと開発者向け**API**の刷新

Gemini 3.5 Flashは、出力の品質とコストを最適化するために、APIの挙動とパラメータ制御に関する

いくつかの重要なパラダイムシフトをもたらした。

4.1. 「Dynamic Thinking」機能と推論レベルの制御

最大の変化は、モデルの推論深度（思考時間）をユーザー側で明示的に制御できる「Dynamic Thinking」の導入である[8]。Gemini 3.x標準に準拠する形で提供されるこの機能は、タスクの難易度に応じて4段階のthinking_level（思考レベル）を設定できる。旧プレビュー版ではデフォルトがhighであったが、Gemini 3.5 Flashでは速度とコストのバランスに優れたmediumが新たなデフォルト設定となった[8]。

- **MINIMAL**（最小）: 推論プロセスをほぼ完全に無効化する。チャットボットのようなリアルタイム性が求められるユースケース、単純な事実確認、容易なツール呼び出しなど、出力のスピードとスループットが最優先される場面に最適化されている[8]。
- **LOW**（低）: 短いレイテンシを維持しつつ、一定の推論を行う。単純なコーディングや少ないステップで完了するエージェントタスク、または一定の思考を要する文章作成やデータ分析に適している。Googleによれば、Gemini 3.5 Flashにおいてはこの「低思考レベル」の品質が飛躍的に向上しており、コストを抑えつつ高いパフォーマンスを発揮するという[8]。
- **MEDIUM**（中・デフォルト）: 複雑なコーディングや標準的なエージェント機能など、一般的な開発ユースケースにおいて最適な結果をもたらすデフォルト設定。処理時間、コスト、推論品質のバランスが取れている[8]。
- **HIGH**（高）: モデルの論理的推論能力とツール使用能力を最大まで引き上げる。複雑な数学的問題の解決、高度なアルゴリズムの設計、多段階にわたる長期的な計画立案など、難易度の高いタスクに不可欠である（一部のコンソール等ではこれがデフォルトになっている場合もある）[8]。

なお、以前のバージョンで使用されていた数値によるthinking_budgetの指定は非推奨となり、上記の文字列表現（enum）を使用することが強く推奨されている[8]。

4.2. 推論コンテキストの保存と関数呼び出しの厳格化

Gemini 3.5 Flashは、マルチターンの対話において「中間的な推論コンテキスト」を自動的に保存し、次のターンへと引き継ぐ能力（Thought Preservation）を備えている[8]。反復的なデバッグ作業などにおいて、モデルが過去に「どのような思考過程を経てその結論に至ったか」を維持できるため、コンテキストの断絶を防ぐことができる。Interactions APIを使用する場合は自動的に処理されるが、GenerateContent APIを使用する場合は、思考シグネチャを含む全会話履歴をcontentsパラメータに未変更のまま渡す必要がある[8]。

さらに、システム設計における極めて重要な変更として、temperature、top_p、top_kといった従来のサンプリング・パラメータの変更が「非推奨」となった[8]。Gemini 3シリーズの高度な推論とエージェント機能は、デフォルトの設定値に深く最適化されており、これらを変更すると予期せぬ性能低下を招く恐れがある。決定論的な出力を得たい場合は、パラメータ操作ではなく、明確な「システムプロンプト（System Instructions）」を用いてルールを明示することが推奨されている[8]。

また、ツールを統合する際の「関数呼び出し（Function Calling）」においても、エラーや空の応答を防ぐため、厳格なマッチングが要求されるようになった。GenerateContentおよびInteractions APIでは、すべてのFunctionResponseに、対応するFunctionCallのidが含まれていること、nameが完全に一致していること、そして1つの呼び出しに対して正確に1つの応答が返されることが必須となってい

る[8]。画像などのマルチモーダルコンテンツを関数応答に含める場合は、モデルへの思考漏れ（Thought leakage）を防ぐため、応答パーツの「内側」に含める必要がある[8]。

4.3. エンタープライズ向け機能と強固なセキュリティ

Gemini Enterprise Agent Platformを通じて提供されるGemini 3.5 Flashは、企業の厳しいコンプライアンスを満たすための高度な機能とセキュリティ制御を備えている[10]。
サポートされる主要機能：

- **Google Search Grounding**: 最新情報を取得するため、ライブのGoogle検索結果をモデルの推論に統合する機能[10]。
- **Code Execution**: モデルが生成したコードを安全なサンドボックス環境内で直接実行・検証する機能[10]。
- **Structured Output**: 確実なシステム統合のため、応答をJSONなどの特定のフォーマットに強制する機能[10]。
- コンテキスト・キャッシング: 長大なプロンプトを入力する際、コストを削減するために暗黙的・明示的に入力トークンをキャッシュする機能[10]。
- **OpenAI互換API**: 既存システムからのスムーズな移行を促すためのチャットコンプリーション互換エンドポイント[10]。
- ※なお、Gemini Live API（リアルタイム音声/動画処理）およびコンテンツ認証技術（C2PA）は、現時点ではサポートされていない[10]。

エンタープライズ・セキュリティ制御（オンライン予測およびコンテキスト・キャッシングにて対応）[10]：

- **Data Residency**（データ・レジデンシ）: データの保存および処理を行う地理的リージョンを指定し、各国の法規制に対応する。
- **CMEK**（**Customer-Managed Encryption Keys**）: 顧客自身が管理する暗号鍵を用いてデータを保護する。
- **VPC Service Controls**: 仮想プライベートクラウド（VPC）の境界内でデータを処理し、外部への情報漏洩を防止する。
- **AXT**（**Access Transparency**）: Googleの管理者によるアクセスを監査するためのアクセストランスペアレンシ機能（バッチ予測やチューニングでは非対応）。

5. ベンチマークが示す知能と処理速度のブレイクスルー

Gemini 3.5 Flashは、これまでの「軽量・安価なモデルは推論能力で劣る」という業界の常識を根本から覆すパフォーマンス指標を叩き出している。Googleは、同モデルがエージェンティック・タスクにおいて最適な「速度と知能のバランス」を提供すると強調している[2]。

5.1. 圧倒的な生成スピードとスループット

モデルの実用性を決定づける出力スピードにおいて、Gemini 3.5 Flashは1秒あたりおよそ277〜289トークンという驚異的なスループットを記録している[9]。これは、前世代のハイエンドモデルである「Gemini 3.1 Pro」の1秒あたり約135トークンと比較して2倍以上の高速化であり、競合他社の同等フロンティアモデルと比較した場合は実に4倍の速度に達する[4]。
この速度向上は、チャットボットとしての体感レスポンスを「スナッピー（俊敏）」にするだけでなく、後述するAntigravityプラットフォームにおける「複数エージェントによる非同期・並列処理」をボトルネッ

クなしで実行するための必須条件となっている[6]。エージェントが自律的に多数の判断を下し、ツールを連続して呼び出す環境下において、推論レイテンシの低さは直接的にソフトウェア開発のベロシティ向上に直結する。

5.2. ベンチマーク指標におけるProモデル超え

独立評価機関およびGoogle独自の測定によれば、Gemini 3.5 Flashは主要なコーディング、エージェント能力、およびマルチモーダル・ベンチマークにおいて、フラグシップであるGemini 3.1 Proを凌駕する成績を収めている[2]。

ベンチマーク指標	スコア / 評価	概要
Terminal-Bench 2.1	76.2%	CLI環境における複雑なエージェンティック・タスクの実行能力[2]
GDPval-AA	1656 Elo	汎用的な推論および問題解決能力の相対評価[2]
MCP Atlas	83.6%	複数コンテキストやツールの使用を伴う複雑なワークフロー処理[2]
CharXiv	84.2%	チャート、グラフ、科学的図表を含む高度なマルチモーダル理解力[2]
Artificial Analysis Intelligence Index	55	同等クラスモデルの平均スコア（36）を大幅に上回る。全147モデル中7位の高評価[12]

これらの数値は、AIの進化が新たな局面を迎えたことを示している。「計算資源を大量に投下した巨大で遅いモデル（Pro/Ultraクラス）」でなければ解決不可能だと考えられていた複雑な論理展開やゼロベースでのコード生成が、高度に最適化された中規模の高速モデル（Flashクラス）によって十分に代替可能になったのである。

6. 新たな開発者エコシステムの誕生：Google Antigravity 2.0の全貌

Google I/O 2026において、開発者コミュニティに最大の衝撃を与えた発表の一つが、エージェント・ファーストの開発プラットフォーム「Google Antigravity 2.0」の登場である[4]。昨年「バイブコーディング（Vibecoding）」の台頭とともに発表された初代Antigravityは、MicrosoftのVS Codeに酷似したIDE（統合開発環境）の拡張機能としての性質が強かった[15]。しかし今回のバージョン2.0で、Antigravityは完全に独立したスタンドアロンのデスクトップ・アプリケーション（macOS, Linux,

Windows対応）として再設計された[4]。
これは、単なるエディタの進化ではない。人間の開発者がコードを書くための環境から、AIエージェントの作業を「オーケストレーション（指揮・統括）」し、監視するための中央司令室（Unified Platform）へのUI/UXのパラダイムシフトである[15]。

6.1. OSコア生成デモが示す「ソフトウェア工場の自動化」

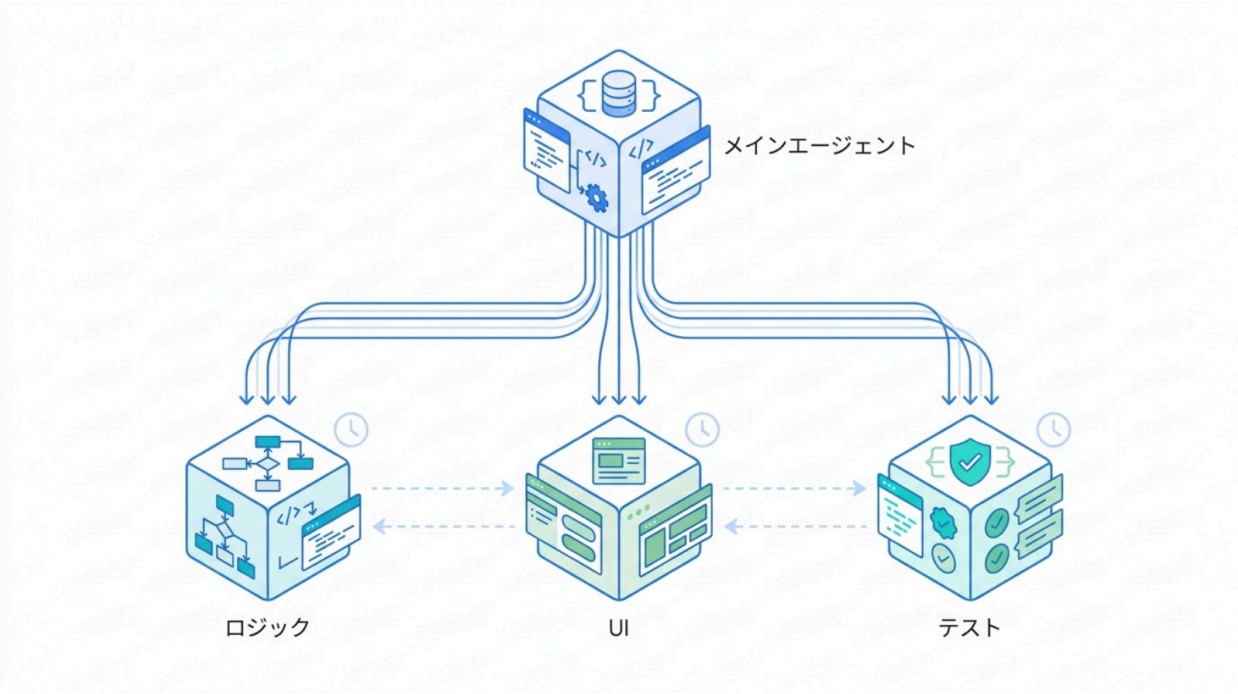
このプラットフォームの威力を如実に示したのが、I/Oのプレゼンテーション内で行われたデモンストレーションである。Varun Mohan氏による発表では、Antigravity 2.0が93個のAIサブエージェントを連携させ、わずか12時間で機能するオペレーティングシステム（OS）のコアを構築することに成功した[14]。
この前代未聞のプロセスにおいて、システムはエージェント間で26億トークンという膨大な通信を行ったが、そのAI処理コスト（APIの利用料）は1,000ドル未満に収まったと報告されている[18]。高度なソフトウェア・エンジニアリング・チームが数週間から数ヶ月かけて行う基盤開発を、AIエージェント群が並列処理を行うことで半日に短縮し、かつ経済的コストを劇的に引き下げられるという「ソフトウェア工場の完全自動化」の概念実証（PoC）である。

6.2. 並列実行を可能にするアーキテクチャの革新

Antigravity 2.0がこれほどの生産性を発揮できる背景には、Gemini 3.5 Flashの高速推論を前提としたいくつかのアーキテクチャ上の革新が存在する。

1. 動的サブエージェント（**Dynamic Subagents**）の自律生成: メインとなるエージェントが複雑な課題に直面した際、自ら判断して特定領域に特化したサブエージェントを動的に定義・生成し、タスクを分割して並列実行させる機能である[16]。これにより、メインエージェントのコンテキストウィンドウが不要なタスク情報で汚染されることを防ぎ、システム全体の並列処理能力と開発速度を劇的に向上させる[15]。
2. 非同期タスク管理とスケジュール実行（**Scheduled Tasks**）: タスクやコマンドは非同期で管理され、メインプロセスの作業をブロックすることなくバックグラウンドで進行する[16]。さらに、Cronスケジュールを定義することで、人間の介入を一切必要とせずにエージェントを定期的に起動し、コードのテスト、システムの監視ルーチン、定期的なリファクタリングなどを自律的に実行させることが可能となった[16]。
3. **JSON**フックと**Live Voice Transcription**による高度な制御: 開発者は単純なJSONフォーマットでフックを定義し、エージェントの挙動を直接制御・傍受することができる[16]。また、インターフェースに最新のGemini Audioモデルが統合されており、リアルタイムの音声対話（Live Voice Transcription）によって、口頭での指示を明確なプロンプトに変換してエージェントに与える機能も搭載されている[17]。
4. プロジェクト管理とアーティファクト: 従来のようにリポジトリとエージェントの会話が密結合する制約を取り払い、複数のフォルダやワークスペースを跨いでエージェントを稼働させる「プロジェクト（Projects）」単位の管理が可能となった[16]。エージェントの進捗は「アーティファクト（Artifacts）」として可視化され、ユーザーは出力物に対して直接フィードバックを与え、軌道修正を行うことができる[16]。

Antigravity 2.0における動的サブエージェントの並列オーケストレーション



Antigravity 2.0では、メインエージェントがコンテキストを汚染することなくタスクを自律的に分割し、複数のサブエージェントをバックグラウンドで並列実行させることで、劇的な開発速度の向上を実現する。

6.3. エコシステムの拡張：CLI、SDK、そしてシームレスな統合

Googleはグラフィカルなユーザーインターフェースを持たない環境や、独自インフラでの運用を望む企業向けにもエコシステムを拡張している。
新たな「Antigravity CLI」は、ターミナル環境に留まることを好む開発者に対し、GUIなしで瞬時にエージェントを生成・操作できる軽量かつ高ベロシティなインターフェースを提供する[4]。これはAntigravity 2.0と同一のエージェント・ハーネスを共有しており、コアエージェントの改善が自動的に適用される[19]。注意点として、この新しいCLIは既存の「Gemini CLI」を完全に置き換えるものであり、ユーザーは2026年6月18日までに新環境へのワークフロー移行を求められている[15]。
また「Antigravity SDK」が提供されたことで、開発者はGoogle自身のプロダクトを駆動しているのと同じ強力なエージェント基盤にプログラムから直接アクセスし、マークダウンファイルを使用して独自のスキルやカスタム指示を定義し、任意のインフラ上に独自のエージェントをホストすることが可能となった[4]。
さらに、Google AI Studio、Firebase、Android環境との統合も深化している。例えば、出先で思いついたアイデアを「Android Studio」の専用アプリやGoogle AI StudioのPlaygroundで初期テストし、そのプロジェクトをすべてのコンテキストとファイル状態を保持したまま、ローカルのAntigravityアプリにエクスポートして開発を再開する、といったシームレスなマルチターン・セッションが実現している[4]。

7. 消費者向けおよびエンタープライズ製品への統合

Gemini 3.5 Flashの影響は、開発者ツールだけにとどまらない。同モデルは即座にGoogleの消費者向けサービスの根幹に組み込まれた。

発表同日より、数十億人のユーザーが利用する無料版の「Geminiアプリ」およびGoogle検索の「AI Mode」のデフォルトモデルが、Gemini 3.5 Flashへと切り替わった[1]。特にGeminiアプリは、Android、iOS、macOS全体で大規模な刷新を受けており、単なるテキスト応答型のチャットボットから、ユーザーの要求を先回りして処理する「プロアクティブな24/7（年中無休）のデジタル・アシスタント」への進化を目指して6つの重要な新機能が導入されている[20]。

さらに、今後リリースが予定されているパーソナルAIエージェント「Gemini Spark」も、このGemini 3.5 Flashによって駆動されることが明かされた[6]。Googleはこれらの展開において、強力なサイバー保護機能、AIの透明性向上、そして厳格な安全保護対策（Safety protections）が組み込まれていることを強調している[7]。

8. 市場の評価：肯定的な視点と現場の熱狂

Gemini 3.5 Flashのローンチは、AI業界全体やソフトウェア開発者コミュニティにおいて即座に大きな議論を呼び起こした。その圧倒的な性能と仕様変更を巡って、市場の評価は賞賛と批判に二分されている。

肯定派の評価の大部分は、その「速度の劇的な向上」と「エージェントとしての即戦力」に集中している。一般ユーザーからは、Google検索やGeminiアプリのレスポンスが、モデルの切り替えによって即座に「スナッピー（俊敏）」に体感できるようになったことが高く評価されている[6]。

開発者コミュニティからは、特にコーディング能力に対する称賛の声が上がっている。ゼロからのコード生成や、ワンショット（一回のプロンプト入力）での複雑な推論能力が非常に高く、「最先端のフロンティアモデルに近い」知能を持っていると実証されている[9]。前述のAntigravity 2.0プラットフォームとの親和性も相まって、自律的なエージェント・ワークフローやリアルタイムの意思決定が求められるアプリケーションを構築するためのエンジンとして、現時点で最高峰の選択肢の一つであると見なされている[10]。

9. 市場の評価：批判的な視点とコスト構造の矛盾

しかしその一方で、APIを利用して独自プロダクトを開発する企業やプロフェッショナルの層からは、Gemini 3.5 Flashに対して極めて厳格かつ痛烈な批判が巻き起こっている[8]。批判の核心は、モデルの「コスト構造の破綻」と、出力の「冗長性（Verboseness）」に起因する予測不可能な費用の増大にある。

9.1. API利用価格の劇的な上昇と複雑な料金体系

最も強い不満の対象となっているのが、API単価の劇的な値上げである。Gemini 3.5 FlashのAPI利用価格は、入力が100万トークンあたり1.50ドル（コンテキストキャッシュ利用時は0.15ドル）、出力が100万トークンあたり9.00ドルに設定されている[8]。

前モデルである「Gemini 3.0 Flash Preview」の価格（入力0.50ドル / 出力3.00ドル）と比較すると、ユーザーの負担額は単価ベースで正確に**3**倍に跳ね上がっている[8]。さらに、超軽量モデルである「Gemini 3.1 Flash-Lite」（入力0.25ドル / 出力1.50ドル）と比較すると実に6倍の価格差がある[8]。この価格は、Googleの上位モデルである「Gemini 3.1 Pro」（入力2.00ドル / 出力12.00ドル）の価格帯に

肉薄しており、「Flash＝安価」というこれまでのブランドイメージを完全に破壊する設定となっている[8]。
さらに、Geminiファミリーの料金体系は「コンテキストウィンドウの閾値」によって価格が2倍に跳ね上がる複雑な構造を持っている。
Geminiモデル価格比較表（**100**万トークンあたり・**USD**）
[8]

モデル名	コンテキスト閾値条件	入力価格	出力価格
Gemini 2.0 Flash Lite	条件なし	$0.075	$0.30
Gemini 2.5 Flash	≤200k	$0.30	$2.50
Gemini 3.0 Flash Preview	条件なし	$0.50	$3.00
Gemini 3.5 Flash	条件なし（**1M**まで）	**$1.50**	**$9.00**
Gemini 3.1 Pro	≤200k	$2.00	$12.00
Gemini 3.1 Pro	>200k（超過時）	$4.00	$18.00

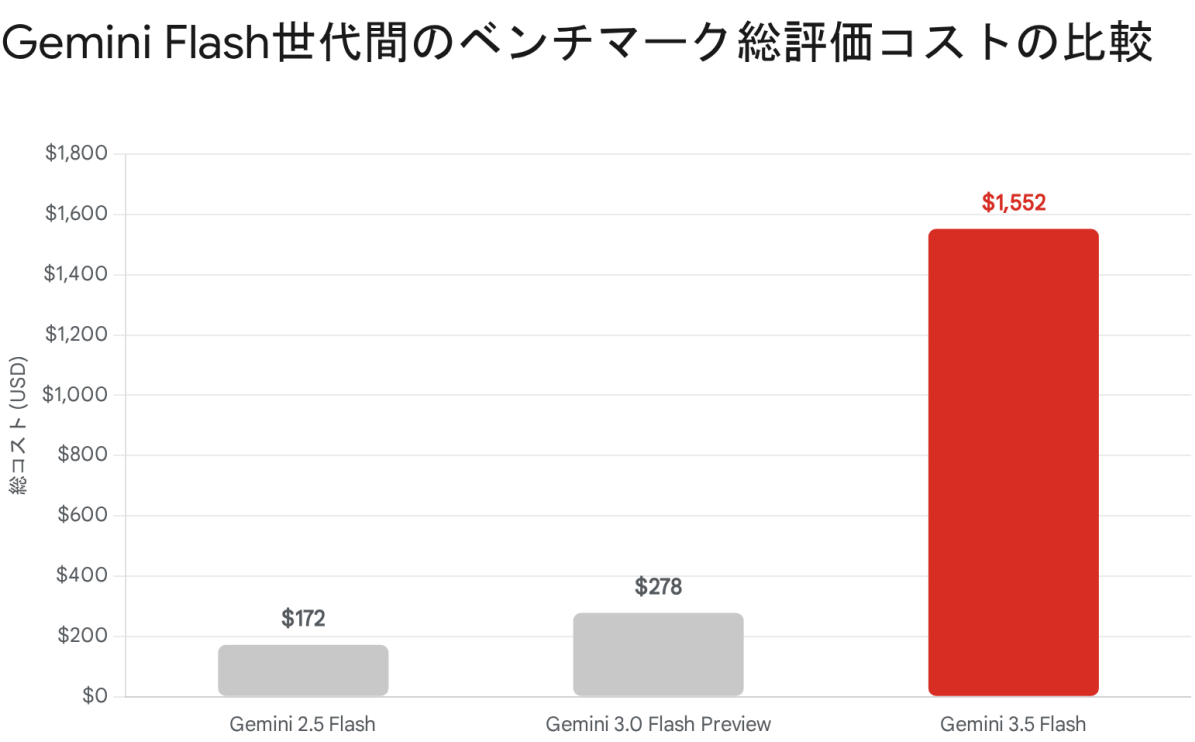
9.2. 出力の冗長性と隠れたコスト増（トークンの肥大化）

単価の引き上げ以上に開発者を悩ませている深刻な問題が、Gemini 3.5 Flashの出力傾向（Verbosity）である。Artificial Analysisが実施したIntelligence Indexの評価ベンチマークにおいて、Gemini 3.5 Flashは評価の全過程で**7,300**万トークンという途方もない量を出力した[9]。これは、同等クラスの他モデルの平均出力トークン数（3,600万トークン）の2倍以上に相当する「冗長な」数値である[12]。
この「出力トークン数の肥大化」と「出力単価の3倍化」が掛け合わさった結果、実際の運用時（ベンチマーク実行時）にかかる総コストは破壊的な上昇を記録している[8]。

ベンチマーク総評価コストの比較（**Intelligence Index**）
[9]

モデル名	**Intelligence Index** スコア	ベンチマーク総評価コスト	**2.5 Flash**比のコスト増
Gemini 2.5 Flash	27	$172	1.0x
Gemini 3.0 Flash Preview	46	$278	1.6x
Gemini 3.1 Pro Preview	-	$892	5.1x
Gemini 3.5 Flash	**55**	**$1,551**	**9.0x**

Gemini 2.5 Flashと比較すると、スコア（知能指標）が約2倍になった代償として、コストは約9倍（日本円で約2万7,000円から約25万円への増加）に膨れ上がっている[9]。



Gemini 3.5 Flashは性能指標（スコア）を向上させた一方で、出力トークンの冗長化と単価引き上げにより、実運用時のコストが前世代のプレビュー版から5倍以上、2.5世代からは約9倍に高騰している。

Data sources: Reddit (r/BetterOffline), Gigazine

9.3.「過剰装飾」と構造的デバッグの弱点（Pelican SVGテスト）

この出力の冗長性は、単なるコスト増にとどまらず、モデルの推論特性における致命的な弱点としても指摘されている。開発者からの報告によれば、Gemini 3.5 Flashは、SVGコードやウェブサイトの修正を要求された際、コアとなる構造的なエラーをピンポイントで修正するのではなく、不要な背景やボタン、過剰な装飾要素（Superficial fluff）を勝手に追加して「肥大化（Bloated）」した結果を出力する強い傾向がある[9]。

この傾向を如実に表したのが、著名な開発者Simon Willison氏が実施した「The Pelican SVG Test（自転車に乗るペリカンのSVGテスト）」である[9]。このプロンプトに対してモデルは、非常に詳細で情報密度の高い装飾的な画像を生成したものの、ペダルと後輪を繋ぐメカニズムという最も重要な「根本的な構造コンポーネント」を描き損ねた[9]。しかも、この過剰に装飾された無用な1つの出力に、約13セントという高額なAPIコストが消費された[9]。

ユーザーは、Antigravityのような環境でモデルに任意のツールの使用を許可し、長期的なエージェントタスクを行わせた場合、モデルが根本的な解決に至らないまま冗長な出力を繰り返し、無用なトークンを大量消費してループに陥る危険性を強く懸念している。そのため、コストパフォーマンスの観点から「Flashという名称は相応しくなく、実態はあらゆる面で安価なオープンモデル（Gemma4 26Bなど）に劣る高コストモデルである」と切り捨てる開発者すら存在している[9]。

10. 波及効果と業界全体の潮流：AIラボの価格戦略

市場のアナリストや専門家は、Gemini 3.5 Flashの強気な価格設定を「Google単独の問題」ではなく、AI業界全体の潮目の変化として捉えている[8]。これまで、主要なAI開発企業（AIラボ）は市場シェアを獲得するために熾烈な価格競争を繰り広げ、モデルの知能向上と並行してAPIの利用料を継続的に引き下げてきた。

しかし、Simon Willison氏の分析によれば、今回の値上げは、OpenAIの「GPT-5.5」やAnthropicの「Claude Opus 4.7」に見られる価格上昇の傾向と軌を一にしており、大手AIラボが一斉に「API顧客の価格許容度（Price Tolerance）を探り始めている」明白な兆候である[8]。つまり、AI技術のエコシステムがコモディティ化による低価格競争のフェーズを終え、エージェンティックな高度な推論能力を提供する対価として、実質的な値上げによる収益化フェーズへと突入したことを意味している。

11. 結論：エージェンティックAIの夜明けと持続可能な運用の条件

Google I/O 2026におけるGemini 3.5 Flash、Gemini Omni、およびAntigravity 2.0プラットフォームの発表は、AIの活用パラダイムを「人間を補助する便利なツール」から「自律的にタスクを完遂するソフトウェア工場」へと不可逆的に推し進めるGoogleの野心と自信の表れである。

Antigravityプラットフォームと、それに最適化された超高速推論エンジンであるGemini 3.5 Flashの組み合わせは、エンタープライズにおける開発プロセスやビジネスワークフローを根本から破壊するポテンシャルを秘めている。「12時間、1,000ドル未満のAPIコストでOSのコアモジュールを構築できる」という事実は、従来のシステムインテグレーションやソフトウェア開発における下請け構造の経済的前提を崩壊させる強力な初期シグナルである。

しかし同時に、Gemini 3.5 Flashが引き起こした「コストの高騰と冗長性の問題」は、エージェンティックAI時代に特有の新たな経営課題を浮き彫りにした。自律的なAIエージェントは、人間の監視を離れてバックグラウンドで無数のプロンプトとレスポンスのループを繰り返す。モデルが1回の出力で冗長な情報を付加する（過剰に思考しすぎる）特性を持っている場合、並列化されたサブエージェント環境下では、そのトークン消費は雪だるま式に膨れ上がり、企業に予期せぬ巨額のクラウドリソース請求をもたらすリスクがある。

Googleは現在、さらなる上位モデルである「Gemini 3.5 Pro」のリリースを来月に控えている。Flashモデルでさえ旧Proモデルと同等の価格帯に到達した現在、次期Proモデルがどのような価格設定と性能曲線を描くのかは業界最大の関心事となっている。

総括すると、Gemini 3.5 FlashはAIの「知能」と「速度」のトレードオフを過去のものにした技術的ブレイクスルーであることは疑いない。プレビュー期間をスキップし即日一般公開に踏み切り、無料版のコンシューマー向け製品にまで惜しげもなく投入したGoogleの決断もそれを裏付けている。しかし、AIの知能が向上しシステムの自律性が高まるほど、モデルの「思考の効率性（少ないトークンで的確な構造的課題を突く力）」と「経済的な予測可能性」が、今後のエンタープライズ市場への本格導入における最大の勝負の分かれ目となることが、今回の市場の反響から如実に示されている。企業や開発者は今後、AIモデルを単なる「便利なAPI」としてではなく、多大なコストを消費しながら自律稼働する「仮想的な従業員」として、そのパフォーマンスとROIを厳密に評価・管理する高度な運用フェーズへと移行せざるを得ないだろう。

引用文献

1. Everything Google announced at I/O 2026: Biggest upgrade to Search in 25 years, new Gemini 3.5 Flash and Gemini Omni AI model, redesigned Gemini app, and more, 5月 20, 2026にアクセス、https://timesofindia.indiatimes.com/technology/tech-news/everything-google-announced-at-i/o-2026-biggest-upgrade-to-search-in-25-years-new-gemini-3-5-flash-and-gemini-omni-ai-model-redesigned-gemini-app-and-more/articleshow/131217138.cms
2. Innovations from Google I/O 26 on Google Cloud | Google Cloud Blog, 5月 20, 2026にアクセス、https://cloud.google.com/blog/products/ai-machine-learning/innovations-from-google-io-26-on-google-cloud
3. Google I/O 2026: Google reveals Gemini Omni, Gemini 3.5 Flash with faster AI performance, 5月 20, 2026にアクセス、https://www.livemint.com/technology/tech-news/google-i-o-2026-google-reveals-gemini-omni-gemini-3-5-flash-with-faster-ai-performance-11779211490497.html
4. I/O 2026 developer highlights: Antigravity, Gemini API, AI Studio - Google Blog, 5月 20, 2026にアクセス、https://blog.google/innovation-and-ai/technology/developers-tools/google-io-2026-developer-highlights/
5. Google launches Gemini 3.5 Flash model. How to try it for free now. - Mashable, 5月 20, 2026にアクセス、https://mashable.com/article/google-io-2026-gemini-35-flash
6. Everything Google announced at I/O 2026: Gemini 3.5, Omni, Spark, and the Search that's changed forever, 5月 20, 2026にアクセス、https://timesofindia.indiatimes.com/technology/tech-news/everything-google-announced-at-i/o-2026-gemini-3-5-omni-spark-and-the-search-thats-changed-forever/articleshow/131218550.cms
7. Google I/O 2026: From AI agents to smart glasses, here are the biggest announcements, 5月 20, 2026にアクセス、https://indianexpress.com/article/technology/tech-news-technology/google-i-o-2026-gemini-biggest-annoucements-10698450/
8. Gemini 3.5 Flash: more expensive, but Google plan to use it for ..., 5月 20, 2026にアクセス、https://simonwillison.net/2026/May/19/gemini-35-flash/
9. Gemini 3.5 Flashの総パラメータ数は2500億〜3000億か、Hacker ..., 5月 20, 2026にアクセス、https://gigazine.net/news/20260520-gemini-3-5-flash-parameter/
10. 【アップデート】Gemini 3.5 Flashがリリースされました ..., 5月 20, 2026にアクセス、https://dev.classmethod.jp/articles/gemini-35-flash-released/
11. Agent Platform Pricing | Google Cloud, 5月 20, 2026にアクセス、https://cloud.google.com/gemini-enterprise-agent-platform/generative-ai/pricing
12. Gemini 3.5 Flash (high) Intelligence, Performance & Price Analysis, 5月 20, 2026にアクセス、https://artificialanalysis.ai/models/gemini-3-5-flash
13. Googleが「Gemini 3.5 Flash」発表 - ケータイ Watch, 5月 20, 2026にアクセス、https://k-tai.watch.impress.co.jp/docs/news/2110005.html
14. Google announces Gemini 3.5 Flash and Antigravity 2.0, claims they made full OS

in 12 hours, 5月 20, 2026にアクセス、https://www.indiatoday.in/technology/news/story/google-io-2026-antigravity-2-0-builds-os-core-in-12-hours-gemini-3-5-debuts-2914194-2026-05-19
15. Google flips Antigravity into an agentic dev suite, AI Studio app lands on Android, 5月 20, 2026にアクセス、https://9to5google.com/2026/05/19/google-antigravity-agentic-developer-suite/
16. Introducing Google Antigravity 2.0, 5月 20, 2026にアクセス、https://antigravity.google/blog/introducing-google-antigravity-2-0
17. Antigravity 2.0, 5月 20, 2026にアクセス、https://www.antigravity.google/product/antigravity-2
18. Google I/O 2026: Google unveils Gemini 3.5 Flash with faster AI performance along with Antigravity 2.0, 5月 20, 2026にアクセス、https://www.financialexpress.com/life/technology-google-io-2026-google-unveils-gemini-3-5-flash-and-antigravity-20-at-major-ai-showcase-4245321/
19. An important update: Transitioning Gemini CLI to Antigravity CLI, 5月 20, 2026にアクセス、https://developers.googleblog.com/an-important-update-transitioning-gemini-cli-to-antigravity-cli/
20. Gemini 3.5 Flash, Gemini Spark and 4 other features coming to Google Gemini app, 5月 20, 2026にアクセス、https://timesofindia.indiatimes.com/technology/tech-news/gemini-3-5-flash-gemini-spark-and-4-other-features-coming-to-google-gemini-app/articleshow/131209863.cms
21. Gemini 3.5 Flash costs are ugly : r/BetterOffline - Reddit, 5月 20, 2026にアクセス、https://www.reddit.com/r/BetterOffline/comments/1ti2qou/gemini_35_flash_costs_are_ugly/
22. Google: Gemini 3.5 Flash - API Pricing & Providers - OpenRouter, 5月 20, 2026にアクセス、https://openrouter.ai/google/gemini-3.5-flash