Google I/O 2026発表「Gemini 3.5 Flash」調査まとめ

作成者: **Manus AI**

作成日: 2026年5月20日

要旨

Googleは2026年5月19日、Google I/O 2026に合わせて**Gemini 3.5シリーズ**を発表し、その最初のモデルとして**Gemini 3.5 Flash**を一般提供しました。Googleの位置づけは明確で、3.5 Flashは従来の「高速・廉価なFlash」という枠を超え、**エージェント実行、コーディング、長期タスク**を主戦場にしたモデルです。公式ブログでは、Gemini 3.1 Proを複数のコーディングおよびエージェント系ベンチマークで上回り、出力速度では他のフロンティアモデルの4倍に達すると説明されています。[1]

評判は、発表直後としては**性能と速度への期待が大きい一方、価格上昇と総コスト、安全性への懸念が目立つ**という構図です。Ars TechnicaやEngadgetは、フロンティア級に近い性能を高速に出せる点を肯定的に報じています。[5] [6] 一方、Simon Willison氏やReddit上の議論では、Flash系としては価格が大きく上がり、推論トークンや出力量を含めた実運用コストがGemini 3.1 Proより高くなる場面がある点が批判されています。[7] [8] Artificial Analysisも、3.5 Flashは知能指数で上位かつ極めて高速だが、出力がやや冗長で、評価実行コストが1,551.60ドルに達したと整理しています。[9]

発表内容と位置づけ

Googleの公式発表では、Gemini 3.5は「frontier intelligence with action」、すなわち**高度な知能を実行能力に結びつけるモデル群**として説明されています。3.5 Flashはその先行モデルであり、Geminiアプリ、Google SearchのAI Mode、Google Antigravity、Gemini API、Google AI Studio、Android Studio、Gemini Enterprise Agent Platform、Gemini Enterpriseで利用可能です。[1]

Google Cloud側の発表も、3.5 Flashを企業向けエージェント基盤の中核モデルとして扱っています。特に、Gemini Enterprise Agent Platform、Google AI Studio、Antigravityで開発者が利用できる点を強調し、長期的なエージェントタスクにおいて同等モデル比で半分未満のコストになり得ると説明しています。[2]

項目	内容
発表日	2026年5月19日、Google I/O 2026
モデル名	Gemini 3.5 Flash

位置づけ	Gemini 3.5シリーズの先行モデル、Flash系の高性能版
主用途	エージェント実行、コーディング、長期ワークフロー、マルチモーダル理解
主な提供先	Geminiアプリ、Search AI Mode、Google Antigravity、Gemini API、Google AI Studio、Android Studio、Gemini Enterprise系
APIモデルID	`gemini-3.5-flash`
コンテキスト	1,048,576入力トークン、最大65,536出力トークン
料金	有料APIでは入力100万トークンあたり1.50ドル、出力100万トークンあたり9.00ドル、キャッシュ入力100万トークンあたり0.15ドル

技術的特徴

Gemini 3.5 Flashの最大の特徴は、**高速性とエージェント能力の両立**です。Googleは、Terminal-Bench 2.1、GDPval-AA、MCP Atlas、CharXiv Reasoningなどのベンチマークを挙げ、Gemini 3.1 Proを上回る領域があると説明しています。[1] Google DeepMindのモデルカードでは、入力としてテキスト、画像、音声、動画を扱い、最大1Mトークンのコンテキストを持つネイティブマルチモーダル推論モデルとされています。[3]

APIドキュメントでは、Gemini 3.5 Flashは**GA、つまり一般提供済みで、本番利用向けに安定している**と明記されています。また、サブエージェント展開、問題解決、反復的なコーディング、長期ツール利用、マルチターン会話における中間推論の保持などが新機能として挙げられています。ただし、現時点でComputer Useはサポートされていないとされています。[4]

> Google AI for Developersは、Gemini 3.5 Flashについて「generally available (GA), stable, and ready for scaled production use」と説明している。[4]

指標・仕様	Gemini 3.5 Flashの値・説明	補足
Terminal-Bench 2.1	76.2%	Google公式およびモデルカードが提示
SWE-Bench Pro	55.1%	モデルカード上ではClaude Opus 4.7などが上位
MCP Atlas	83.6%	モデルカード上で比較対象中トップ

Toolathlon	56.5%	実世界ツール利用の評価
OSWorld-Verified	78.4%	GPT-5.5の78.7%に近い水準
GDPval-AA	1656 Elo	GPT-5.5やClaude Opus 4.7が上位
CharXiv Reasoning	84.2%	モデルカード上で非常に高い値
Artificial Analysis速度	277.0出力トークン/秒	147モデル中2位

重要なのは、3.5 Flashが**すべてのベンチマークで首位ではない**ことです。Googleのモデルカードでも、Terminal-Bench 2.1ではGPT-5.5、SWE-Bench ProではClaude Opus 4.7、GDPval-AAではGPT-5.5、長文コンテキストのMRCR v2 128kではGPT-5.5が上回っています。[3] したがって、「最強」というよりは、**エージェント・コーディング・速度を重視した総合バランス型の上位モデル**と見るのが妥当です。

肯定的な評判

肯定的評価の中心は、**速度が非常に速く、実用的なエージェント処理に向く**という点です。Ars Technicaは、3.5 Flashが約300トークン/秒に近い出力速度を持ち、より大きなフロンティアモデルに近いスコアを4分の1程度の時間で出せると報じました。[5] Artificial Analysisでも、出力速度277.0トークン/秒、147モデル中2位とされており、速度面の評価はかなり高いです。[9]

第二に、**コーディングとサブエージェント運用**への期待があります。TechCrunchは、Googleがチャットボットではなくエージェントを次のAI波として打ち出していると分析し、Antigravity 2.0で複数エージェントが個別コンポーネントを担当してOSを構築するデモを紹介しました。[10] Engadgetも、3.5 Flashが長期エージェントタスクに適し、銀行やフィンテック企業が複数週のワークフロー自動化に活用しているというGoogleの説明を取り上げています。[6]

第三に、**Google製品への即時・広範な投入**が信頼感を生んでいます。GeminiアプリとSearch AI Modeでデフォルトモデルとして提供され、Google Antigravity、Gemini Enterprise系、Gemini APIにも同時展開されているため、Googleが3.5 Flashを単なる実験モデルではなく中核モデルとして扱っていることが分かります。[1] [7]

肯定的論点	根拠	代表的な見方
高速性	Googleは他のフロンティアモデル比4倍、Artificial Analysisは277トークン/秒と評価	エージェント処理では反復回数が多いため速度の価値が大きい
コーディング性能	Terminal-Bench 2.1、SWE-Bench Proなどで旧FlashやGemini 3.1 Proを上回る領域	開発支援、レガシーコード移行、プロトタイピングに有望

エージェント向け設計	Antigravity、サブエージェント、長期ワークフローを前面に出している	チャット中心から実行中心への転換を象徴
本番利用可能	APIドキュメントでGA、安定、本番スケール向けと説明	開発者がすぐ導入しやすい
マルチモーダル・長文対応	1M入力トークン、テキスト・画像・音声・動画入力	ドキュメント処理や複合的な業務タスクに向く

否定的な評判・懸念

否定的評価で最も大きいのは、**Flash系としては価格が高くなった**という点です。公式料金では、3.5 Flashの有料API価格は入力100万トークンあたり1.50ドル、出力100万トークンあたり9.00ドルです。[11] Simon Willison氏は、3.5 Flashが3 Flash Previewの3倍、3.1 Flash-Liteの6倍の価格で、Gemini 3.1 Proの2ドル/12ドルに近づいたと指摘しています。[7]

さらに、単価だけでなく**総コスト**への懸念があります。Artificial Analysisは、3.5 FlashがIntelligence Index評価で73Mトークンを出力し、平均36Mよりかなり冗長だったとしています。その結果、評価コストは1,551.60ドルに達しました。[9] Redditでも、この数値をもとに、3.5 FlashのIntelligence scoreが55でGemini 3.1 Pro Previewの57を下回る一方、評価コストは3.1 Pro Previewの892ドルより高いという批判が投稿されています。[8]

安全性・倫理面の懸念もあります。TechCrunchは、強力な自律エージェントを一般消費者向けに広く提供することは、誤用や有害な出力、依存的利用などへの監視を強めると論じています。[10] GoogleはFrontier Safety Framework、サイバーおよびCBRN対策、解釈可能性ツールの活用を掲げていますが、モデルカード上の自動安全評価ではText to Text SafetyがGemini 3 Flash比で-3.9%、Multilingual Safetyが-2.6%と、改善と後退が混在しています。[3]

否定的論点	内容	影響
価格上昇	Flash系としては入力1.50ドル、出力9.00ドルと高め	大量利用する開発者には負担増
総コストの不透明性	出力が多く、推論トークンを含めると実費が膨らむ可能性	単価だけで比較すると判断を誤る
全ベンチマーク首位ではない	GPT-5.5やClaude Opus 4.7が上回る領域あり	深い推論や特定専門タスクではPro級・競合上位モデルが優位な場合がある
Computer Use未対応	APIドキュメントでは現時点で非対応	汎用PC操作エージェント用途では制約

安全性・倫理	自律エージェントの一般展開には誤用・有害利用リスクが伴う	企業導入ではガードレール設計が必須

総合評価

Gemini 3.5 Flashは、従来のFlash系モデルのイメージである「安い・速い・軽量」から、**速いが高性能、ただし必ずしも安くはない**モデルへ移行した存在です。特に、コーディング、ツール利用、サブエージェント、長期ワークフローでは強い魅力があり、GoogleがGeminiアプリやSearch AI Modeの標準モデルとして投入したことからも、同社のAI戦略の中核に置かれていると考えられます。

一方で、開発者や企業が導入する際は、ベンチマーク上の高評価だけでなく、**タスクごとの出力量、推論トークン、キャッシュ利用、バッチ利用、失敗時の再試行回数**を含めた総費用を試算する必要があります。価格面ではFlash-Liteや旧Flash系より明らかに高く、Gemini 3.1 Proに近い価格帯に入ってきています。したがって、3.5 Flashは低コスト万能モデルというより、**速度が収益性やUXに直結するエージェント処理向けの高性能モデル**として選定するのが適切です。

現時点の結論として、Gemini 3.5 Flashの評判は「非常に速く、エージェント時代に適した有力モデル」という肯定と、「Flashという名前に反して高価で、総コストと安全性を慎重に見るべき」という否定が併存しています。特にAPI利用者にとっては、単価ではなく実タスクでの**総コスト対効果**を測ることが、導入判断の中心になるでしょう。

References

[1] Google Blog - Gemini 3.5: frontier intelligence with action
[2] Google Cloud Blog - Innovations from Google I/O 26 on Google Cloud
[3] Google DeepMind - Gemini 3.5 Flash Model Card
[4] Google AI for Developers - What's new in Gemini 3.5 Flash
[5] Ars Technica - Gemini 3.5 Flash might be fast enough for gen AI to make sense
[6] Engadget - Google says Gemini 3.5 Flash rivals large flagship models for coding and agentic tasks
[7] Simon Willison - Gemini 3.5 Flash: more expensive, but Google plan to use it for everything
[8] Reddit - Gemini 3.5 Flash looks worse than it seems on Artificial Analysis
[9] Artificial Analysis - Gemini 3.5 Flash Intelligence, Performance & Price Analysis
[10] TechCrunch - With Gemini 3.5 Flash, Google bets its next AI wave on agents, not chatbots
[11] Google AI for Developers - Gemini Developer API pricing