ChatGPT

Google I/O 2026におけるGemini 3.5 Flash発表の分析

エグゼクティブサマリー

Googleは現地時間2026年5月19日のGoogle I/Oで、新シリーズ「Gemini 3.5」を発表し、その第一弾として「Gemini 3.5 Flash」を即日展開した。今回の実質的な主役は未公開の3.5 Proではなく、**アプリ・検索・API・Enterpriseを同日に横断配備した3.5 Flash**であり、GoogleがAIを「会話」から「行動」へ移す戦略転換を鮮明にした点が重要である。基調講演では、3.5 FlashはGemini 3.1 Proをほぼ全面的に上回り、他の最先端モデルより**出力速度で4倍高速**、しかも**比較対象の最先端モデルの半額未満**だと位置づけられた。Gemini 3.5 Proは「来月」とだけ予告され、詳細は未公表のままだ。 [1]

技術的には、3.5 Flashは**100万トークン入力、約65k出力、テキスト・コード・画像・音声・動画・PDF入力、テキスト出力**のネイティブ・マルチモーダルモデルで、Gemini API/AI StudioではStable、Google CloudのAgent PlatformではGAとして公開された。一方でDeepMindの紹介ページではPreview表記が残っており、公開面によってステータス表記が揺れている。公開資料はパラメータ数やMoE構成を開示しておらず、モデルサイズは**未公表**である。 [2]

総合評価としては、**「近Pro品質をFlash価格帯と速度で出す」ことが今回の中核価値**であり、特にエージェント実行、コーディング、マルチステップ業務自動化で強い。しかし、長文推論や一部の学術推論ではGPT-5.5やClaude Opus 4.7に劣後する指標もあり、モデルカードでは自動安全評価の一部でGemini 3 Flash比の軽微な悪化も示されている。したがって、短期的には「全面置き換え」よりも、**高難度オーケストレーションを上位モデル、並列サブエージェントを3.5 Flash**に割り当てる混成運用が最も合理的である。 [3]

仕様と位置づけ

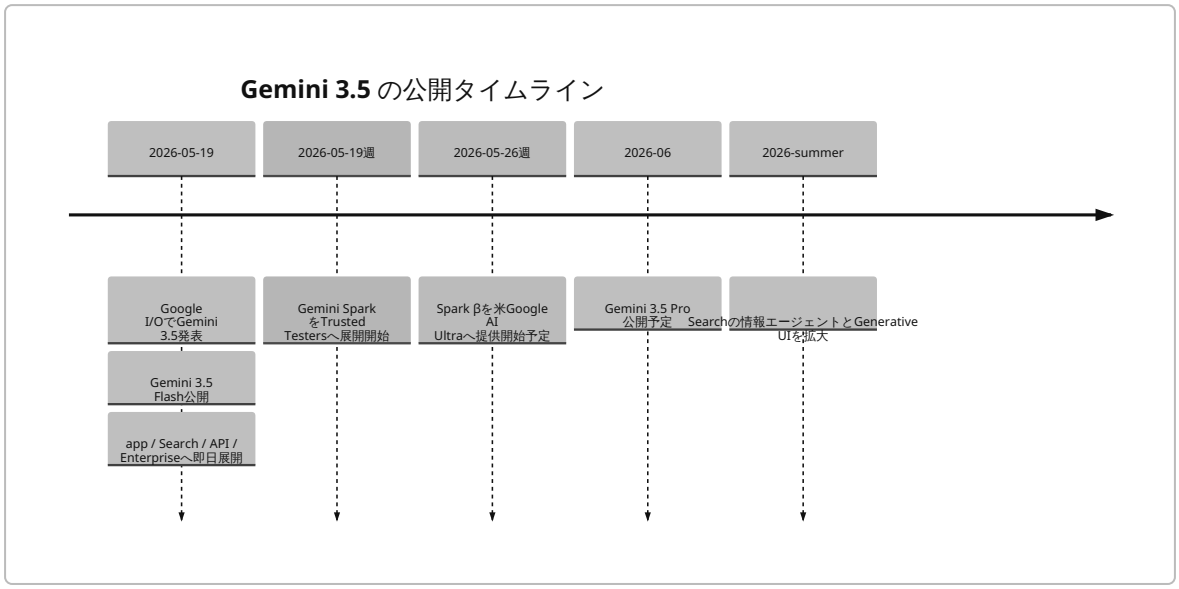
項目	Gemini 3.5 系列	Gemini 3.5 Flash
公開状況	3.5系列を発表。公開済みはFlashのみ。3.5 Proは「来月」予定、詳細未公表。 [4]	`gemini-3.5-flash` は2026-05-19公開。Gemini APIではStable、Cloud Agent PlatformではGA。 [5]
位置づけ	“frontier intelligence with action” がシリーズ全体のメッセージ。 [6]	近Proの知能をFlash級の速度・価格で提供、エージェントとコーディングを主戦場に設定。 [7]
アーキテクチャ注記	系列全体の詳細構成・パラメータ数は未公表。 [8]	Gemini 3 Flash reasoning foundationベース、thinking levelsで品質/コスト/遅延を調整。詳細アーキテクチャとパラメータ数は未公表。 [9]
マルチモーダル	系列としてエージェント・マルチモーダル・長期タスクを強調。 [10]	入力: text/code/image/audio/video/PDF、出力: text。100万入力/65,535出力。 [11]

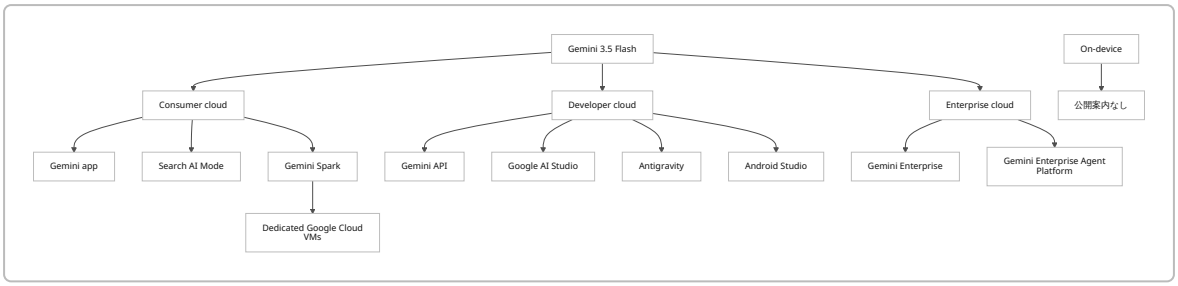
項目	Gemini 3.5 系列	Gemini 3.5 Flash
速度/スループット	公式比較は3.5 Flash中心。	公式に「他の最先端モデルより4倍高速」。Antigravity向け最適版は12倍高速。独立計測ではGoogle AI Studio提供で約277 tok/s、初回応答18.55秒。 12
API/価格	3.5 Proは未公表。	Gemini Developer API標準料金は**入力$1.50 / 出力$9.00**、Batch/Flexは半額水準。3.1 Pro Previewより安いが、3.1 Flash-Liteより大幅に高い。 13
配備形態	系列全体はクラウド中心。	Gemini app、Search AI Mode、Gemini API、AI Studio、Gemini Enterprise、Agent Platform、Antigravity、Android Studio。**オンデバイス版の公開案内はなし**。SparkはGoogle Cloud上の専用VMで稼働。 14
言語	広範な多言語対応。日本語を含む。 15	日本語を含む多言語対応。日本ではGeminiアプリFreeプランから3.5 Flashにアクセス可能。 16

補足すると、Gemini APIでは**Grounding with Google Search/Maps、context caching、function calling、structured output、code execution**が使える一方、Live APIは3.5 Flashでは未対応である。さらにGemini API側では2025年以降**チューニング機能が停止**しており、現時点の3.5 Flash活用は、微調整よりもプロンプト設計、RAG、キャッシング、Managed Agents、Provisioned Throughput前提のLLMOpsに寄る。 17

展開スケジュールと提供形態

Googleのロールアウトは異例に速く、**発表当日から**Gemini app、Search AI Mode、Gemini API、AI Studio、Antigravity、Gemini Enterprise、Agent Platformへ広げられた。日本語公式ブログでは、3.5 Flashが**日本を含む各国**でGeminiアプリと検索AI Modeの標準モデルになったと明記されている。Sparkは同週にTrusted Testers向け配布開始、翌週に米国Google AI Ultra向けβ、3.5 Proは6月予定である。 18





既存Geminiと競合比較

Googleの主張は「**3.1 Pro超えのFlash**」であり、これは単なる速度向上ではない。モデルカードではTerminal-Bench 2.1で76.2%、MCP Atlasで83.6%、MMMU-Proで83.6%と、3 Flashや3.1 Proを上回る。とはいえ、MRCR v2 128kやARC-AGI-2では3.1 ProやGPT-5.5、Claude Opus 4.7がまだ強い領域もある。つまり3.5 Flashは、**万能首位ではなく、エージェント/コード/速度の交点で最適化されたモデル**と見るのが正確だ。[19]

モデル	発表/公開	代表的特徴	文脈長/モーダル	価格・配備の要点
Gemini 3.5 Flash	2026-05-19	近Pro知能、4倍速度、エージェント・コーディング重視。[20]	1M、音声/画像/動画/PDF入力、テキスト出力。[11]	$1.50 / $9、app・Search・API・Enterprise即日展開。[21]
Gemini 3.1 Pro	2026-02-19	複雑問題・高度推論向け。[22]	1M、マルチモーダル。[23]	≤200kで$2 / $12、>200kで$4 / $18。[24]
Gemini 3 Flash	2025-12-17	3 Pro基盤の高速版、3.5 Flashの土台。[25]	1M、マルチモーダル入力。[25]	Preview開始。3.5 Flashはここから安全性/性能を上積み。[26]
Gemini 3.1 Flash-Lite	2026-05-07	高ボリューム・低コスト向け。[27]	1M、マルチモーダル入力。[28]	$0.25/$1.50でFlashより大幅に安い。[29]
OpenAI GPT-5.5	2026-04-23	最高クラスの総合知能と長文推論。[30]	約1.05M、128k出力、text+vision。[31]	$5 / $30、より高価。[30]
Anthropic Claude Sonnet 4.6	2026-02-17	高効率のエージェント/企業ワークフロー向け。[32]	1M、text+image。[33]	$3 / $15、Vertex AI等でも提供。[34]
Anthropic Claude Opus 4.7	2026-04-16	高精度な長時間コーディングと高度推論。[35]	1M、128k出力。[36]	$5 / $25、Google Cloudでも提供。[37]
Meta Llama 4 Maverick	2025-04-05	open-weight、MoE、自己運用可能。[38]	ネイティブmultimodal。17B active / 128 experts / 400B total。[39]	Metaの従量API価格は前面に出ず、主にダウンロード/自己運用。[40]

主要な反応

区分	出典・日付	トーン	要旨
公式	Google基調講演、2026-05-19 [41]	Positive	「高知能と高速性の両立」「比較対象の半額未満」「今日から提供」を強く訴求。
公式	Gemini 3.5 Flash model card、2026-05-19 [42]	Neutral	性能向上を示す一方、自動安全評価ではtext safety-multilingual safetyがわずかに後退。
英語メディア	TechCrunch、2026-05-19 [43]	Positive	Googleがチャットボット企業ではなく「エージェント企業」へ軸足を移す象徴的リリースと評価。
英語メディア	Financial Times、2026-05-20 [44]	Neutral	OpenAI・Anthropic優位のコーディング/業務自動化で巻き返しを狙う布石として報道。
英語メディア	Business Insider、2026-05-19 [45]	Negative	3.5 Pro未公開に会場から落胆が出た点を強調。
日本語メディア	ITmedia、2026-05-20 [46]	Neutral	3.1 Pro超えを認めつつ、Flashとしては価格が上がった点を指摘。
日本語メディア	ケータイWatch、2026-05-20 [47]	Positive	公式比較図を基に、3.1 Pro比で大幅な速度向上を強調。
専門分析	Artificial Analysis、2026-05-19 [48]	Positive/Neutral	速度×知能のPareto frontierで首位級だが、Gemini 3 Flash比ではコスト増と整理。
開発者	Simon Willison on X、2026-05-19 [49]	Positive	“previewなしでGA直行”を、実運用志向の強いリリースと受け止めた。
開発者	Reddit r/Bard、2026-05-20 [50]	Positive	体感速度が極めて高く、日常ワークフローを変えるとの反応。
開発者	Reddit r/singularity、2026-05-20 [51]	Negative	ベンチマーク偏重や3.5 Pro待ちの声が目立つ。
開発者	Hacker News、2026-05-20 [52]	Negative	Gemini 3 Flash比の価格上昇、長期エージェント性能への懐疑が見られる。
企業	Box CTO、2026-05-19 [53]	Positive	Box独自評価で3 Flash比19.6%改善、業界別タスク精度も大幅向上。
企業	JetBrains Junie、2026-05-19 [53]	Positive	Proに近い品質をFlashの速度/コストで得られると評価。

リスクと事業インパクト

技術面の懸念は四つある。第一に、安全性は改善一色ではない。モデルカードでは3 Flash比で**text safety -3.9%、multilingual safety -2.6%**と自動評価上の悪化が記載されている。第二に、長文・抽象推論は依然ムラがあり、MRCR v2やARC-AGI-2では上位競合に及ばない。第三に、プライバシー/データ所在では、Sparkが**Google Cloud上の専用VM**で動作し、Cloud文書も**グローバルエンドポイントはデータ所在地やリージョン**

内ML処理を保証しないと明示する。第四に、AI Studio無料枠ではデータが製品改善に利用されうる一方、有料枠では利用しないため、利用形態でガバナンスが変わる。 54

事業面では、Googleは3.5 Flashを**検索・Gemini app・AI Studio・Antigravity・Enterprise**へ同時展開し、単一モデルではなく**単一エコシステム**として売り込んでいる。これは価格競争だけでなく、Search grounding、Maps grounding、Managed Agents、Workspace連携、Enterprise Agent Platformまで含めたロックイン戦略である。他方、Gemini APIでの公開チューニング停止や、3.5 Pro詳細未公表は、厳格なモデル最適化を求める企業にとっては採用判断を遅らせる要因にもなる。 55

短期的には、**コーディング支援、検索エージェント、マルチエージェント試作**で採用が進む公算が大きい。中期的には、3.5 Proが強ければGoogleは「検索→アプリ→開発→企業基盤」まで一気通貫の優位を固めうるが、安全/価格/データ管理の説明不足が残れば、企業はOpenAI・Anthropic・Metaを併用するマルチモデル戦略を維持するだろう。 56

推奨事項と一次情報

対象	主な利点	主な難点	推奨アクション
開発者	高速・1M文脈・関数呼び出し・Search grounding・Batch/Flex対応。 57	Flash-Liteより高コスト、Live API非対応、微調整余地が限定的。 58	まず3.5 Flashを**サブエージェント/コード生成/並列処理**に投入し、最難タスクは他の上位モデルとAB比較。
企業	app・Search・Enterprise・Agent Platform一体運用、Provisioned Throughputや高度なセキュリティ統制。 59	データ所在と無料/有料でのデータ利用ポリシー差が大きい。 60	**Paid/Enterprise前提**でPoCし、法務・情報システムと一緒にデータ分類、ログ保持、地域制約を先に設計。
研究者	エージェント実行、MCP、実務ベンチでの進歩が観測しやすい。 61	公開アーキテクチャ/パラメータ情報が乏しく、再現性は限定的。 9	公式ベンチだけでなく、**独立eval・長文・安全・多言語**で追加検証し、特に日本語の業務タスクで再測定する。

一次情報として最優先すべきなのは、Google公式の**Gemini 3.5発表記事**、**Google I/O基調講演トランスクリプト**、**Gemini 3.5 Flash model card**、**Gemini API pricing/models docs**、**Google Cloud Agent Platform docs**である。日本語一次情報はGoogle Japan Blogの「Gemini 3.5：行動を起こす最先端の知能」と開発者向けハイライトが最も有用だった。 62

navlist関連主要報道turn21news35,turn21news36,turn24news40

---

1 4 10 12 20 41 56 https://blog.google/innovation-and-ai/sundar-pichai-io-2026/
https://blog.google/innovation-and-ai/sundar-pichai-io-2026/

2 7 11 17 57 59 https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-5-flash
https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-5-flash

3 9 14 19 42 54 61 https://deepmind.google/models/model-cards/gemini-3-5-flash/
https://deepmind.google/models/model-cards/gemini-3-5-flash/

5 https://ai.google.dev/gemini-api/docs/models
https://ai.google.dev/gemini-api/docs/models

6 8 https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/

13 21 24 29 58 https://ai.google.dev/gemini-api/docs/pricing
https://ai.google.dev/gemini-api/docs/pricing

15 16 https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/google-models
https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/google-models

18 55 https://cloud.google.com/blog/products/ai-machine-learning/innovations-from-google-io-26-on-google-cloud
https://cloud.google.com/blog/products/ai-machine-learning/innovations-from-google-io-26-on-google-cloud

22 23 https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-1-pro
https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-1-pro

25 26 https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-flash
https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-flash

27 28 https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-1-flash-lite
https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-1-flash-lite

30 https://openai.com/index/introducing-gpt-5-5/
https://openai.com/index/introducing-gpt-5-5/

31 https://developers.openai.com/api/docs/models/gpt-5.5
https://developers.openai.com/api/docs/models/gpt-5.5

32 https://www.anthropic.com/news/claude-sonnet-4-6
https://www.anthropic.com/news/claude-sonnet-4-6

33 34 https://www.anthropic.com/claude/sonnet
https://www.anthropic.com/claude/sonnet

35 37 https://www.anthropic.com/news/claude-opus-4-7
https://www.anthropic.com/news/claude-opus-4-7

36 https://docs.anthropic.com/en/docs/build-with-claude/context-windows
https://docs.anthropic.com/en/docs/build-with-claude/context-windows

38 39 40 https://ai.meta.com/blog/llama-4-multimodal-intelligence/
https://ai.meta.com/blog/llama-4-multimodal-intelligence/

43 https://techcrunch.com/2026/05/19/with-gemini-3-5-flash-google-bets-its-next-ai-wave-on-agents-not-chatbots/
https://techcrunch.com/2026/05/19/with-gemini-3-5-flash-google-bets-its-next-ai-wave-on-agents-not-chatbots/

44 https://www.ft.com/content/c47ab51e-2521-4ccb-9de5-a2b03791981a
https://www.ft.com/content/c47ab51e-2521-4ccb-9de5-a2b03791981a

45 https://www.businessinsider.com/google-io-2026-gemini-3-5-pro-2026-5
https://www.businessinsider.com/google-io-2026-gemini-3-5-pro-2026-5

46 https://www.itmedia.co.jp/aiplus/article/2605/20/2000000010/
https://www.itmedia.co.jp/aiplus/article/2605/20/2000000010/

47 https://k-tai.watch.impress.co.jp/docs/news/2110005.html
https://k-tai.watch.impress.co.jp/docs/news/2110005.html

48 https://artificialanalysis.ai/articles/gemini-3-5-flash-everything-you-need-to-know
https://artificialanalysis.ai/articles/gemini-3-5-flash-everything-you-need-to-know

49 https://x.com/simonw
https://x.com/simonw

50 https://www.reddit.com/r/Bard/comments/1thwjdz/35_flash_is_10x_faster_and_costs_a_third_of_gpt/
https://www.reddit.com/r/Bard/comments/1thwjdz/35_flash_is_10x_faster_and_costs_a_third_of_gpt/

51 https://www.reddit.com/r/singularity/comments/1thtxs8/behold_gemini_35_flash/
https://www.reddit.com/r/singularity/comments/1thtxs8/behold_gemini_35_flash/

52 https://news.ycombinator.com/item?id=48196771
https://news.ycombinator.com/item?id=48196771

53 https://deepmind.google/models/gemini/flash/
https://deepmind.google/models/gemini/flash/

60 https://docs.cloud.google.com/gemini-enterprise-agent-platform/resources/locations?hl=ja
https://docs.cloud.google.com/gemini-enterprise-agent-platform/resources/locations?hl=ja

62 https://blog.google/intl/ja-jp/company-news/technology/gemini-3-5/
https://blog.google/intl/ja-jp/company-news/technology/gemini-3-5/