

Google「Gemini 3.7 Flash」 深掘り調査報告：コーディングとエージェント能力を軸にした“高速ワークホース”戦略

調査基準日：2026年8月14日（日本時間）

エグゼクティブサマリー

Googleは2026年8月13日（米国時間）に「Gemini 3.7 Flash」を正式発表し、Gemini APIでも同日付で一般提供（GA）・本番利用可能とした。Google自身の表現では「coding and agents向けに最も賢いworkhorse model」であり、7月21日に公開したGemini 3.6 Flashからわずか約3週間での更新である。公式ブログには発表の時刻までは記されていないため、**Google自身による正確な公開時刻は未公表**である。一方、Reutersは8月13日17:02 UTCにローンチを報じており、これは日本時間では**8月14日02:02 JST**、米西海岸では8月13日10:02 PDT、米東海岸では13:02 EDTに相当する。したがって、「13日（米国時間）に公開」という理解は正しく、日本では14日未明から朝にかけてニュース化されたと見るのが妥当である。 ①

今回の更新の本質は、モデル規模の大型化をGoogleが公表したことではない。公式モデルカードが明示しているのは、**Gemini 3.6 Flashを基礎とし、コアの推論基盤にアルゴリズム上の改善を加えた**という点までである。パラメータ数、MoEか否か、expert数、層数、学習トークン量、3.6からの訓練データ差分、使用アクセラレータなどは3.7固有情報として公表されていない。したがって「新アーキテクチャ」「大規模化」「新しい学習データセット」と断定する根拠は現時点ではない。Googleが公開している技術的差分は、主として**推論アルゴリズム、thinking制御、コーディング・エージェント評価の向上**である。 ②

能力面では改善幅が大きい。Googleの同一評価表では、3.6 Flash比でFrontierCode 1.1 Mainが**34.4%→43.6%**、DeepSWE v1.1が**48.6%→65.3%**、Terminal-bench 2.1が**78.0%→85.8%**、AutomationBenchが**17.0%→30.4%**、GDP.pdfが**22.0%→34.0%**、OSWorld-2.0が**33.8%→47.9%**へ上昇した。特に、単発のコード補完よりも、デバッグ、issue解決、ターミナル操作、複数ツールをまたぐ業務自動化、UIを参照画像に合わせる作業といった**長い実行ループを持つ仕事**に重点を置いた更新と読める。 ③

ただし、3.7 Flashがあらゆるベンチマークで最強という意味ではない。Google自身の比較表でも、GPT-5.6 TerraはDeepSWE v1.1で69.6%、Terminal-bench 2.1で87.4%、OSWorld-2.0で50.2%と3.7 Flashを上回る。一方、3.7 FlashはFrontierCode、WebDev Arena、AutomationBench、GDP.pdfなどでGPT-5.6 TerraやClaude Sonnet 5を上回る。Artificial Analysis Intelligence Indexでは3.7 Flashが56、Claude Sonnet 5が55、GPT-5.6 Terraが57であり、**最高絶対性能というより、速度・価格・エージェント性能のPareto効率を狙ったモデル**という位置付けが最も適切である。 ④

価格戦略も重要である。2026年12月31日まで、Gemini 3.7 Flashは標準APIで**入力100万トークン当たり0.75ドル、出力3.75ドル**。2027年1月1日以降はそれぞれ**1.50ドル、7.50ドル**になる。期間限定価格ではClaude Sonnet 5の2ドル/10ドル、GPT-5.6 Terraの2ドル/12ドルよりかなり低い。Googleは同じ割引価格を3.6 Flashにも遡って適用しているため、「3.7が3.6の半額」という表現は**3.6の元の定価との比較**であり、現在の期間限定API料金同士では同額である点に注意が必要だ。 ⑤

エージェント面では、モデルそのものだけでなくGoogleの**Antigravity / Managed Agents**との結合が重要である。3.7 FlashはAntigravity agentの新しいデフォルトモデルになり、Interactions APIから1回呼び出す

と、Google側の隔離Linux環境で推論、コード実行、ファイル操作、Web閲覧などを行うことができる。もっとも、subagent、hooks、非同期タスク管理、agent harnessといったオーケストレーション機構自体はI/O 2026で既にAntigravity 2.0に導入されており、**3.7で新しいエージェントフレームワークが突然登場したわけではなく、その「推論エンジン」が3.7へ更新された**と理解するのが正確である。 ⁶

安全性についてGoogleはCBRN（化学・生物・放射線・核）とサイバー攻撃悪用への防御策を更新した。Frontier Safety評価では追跡対象のCritical Capability Levelには到達していないとする一方、**サイバー分野ではCCLのalert thresholdには達している**と評価している。さらにManaged Agentsはデフォルトで外向きネットワークアクセスが無制限であるため、Google自身がallowlist、最小権限credential、短命token、人間による出力確認を推奨している。性能向上と同時に、実運用上の最大リスクが「誤った文章」から「誤った行動」に移っている点は重要である。 ⁷

発表時刻・提供地域・製品ポジショニング

Google公式ブログの日付は**August 13, 2026**で、Gemini APIのchangelogとGoogle Cloudのモデルページも同日をGAリリース日としている。公式発表に「何時何分」という時計時刻は記載されていないため、正確なGoogle公開時刻は**unspecified（未公表）**とすべきである。Reutersは8月13日17:02 UTC時点ですでに「Google launched Gemini 3.7 Flash」と報じているため、少なくともその時点までには外部公開されていたことが確認できる。日本の窓の杜は8月14日06:45 JSTに「米Googleは8月13日（現地時間）発表」と報じている。 ⁸

時刻基準	確認できる時刻・日付	解釈
Google公式発表	2026-08-13	日付のみ。 正確な時刻は未公表。 ⁹
Reuters公開	2026-08-13 17:02 UTC	ローンチ済みであることを報道。 ¹⁰
米西海岸	2026-08-13 10:02 PDT相当	Reuters公開時刻を換算した参考値。
米東海岸	2026-08-13 13:02 EDT相当	同上。
日本	2026-08-14 02:02 JST相当	同上。日本では翌日扱い。
日本語主要報道	2026-08-14 06:45 JST	窓の杜掲載時刻。 ¹¹

提供チャンネルはかなり広い。開発者には**Gemini API、Google AI Studio、Android Studio、Google Antigravity**、企業には**Gemini Enterprise Agent Platform、Gemini Enterpriseアプリ**、個人にはGeminiアプリ内の**Gemini Spark**で展開されている。GoogleはSparkについてGoogle AI Pro/Ultra契約者を対象に160カ国以上で3.7 Flashへ切り替えるとしている。GitHubも同日からGitHub Copilotに段階的展開し、VS Code、Visual Studio、Copilot CLI、cloud agent、JetBrains、Xcode、Eclipseなどのモデルセクターから利用可能にすると発表した。Business/Enterpriseでは管理者によるpreview policyの有効化が必要である。 ¹²

Google Cloud側ではモデルのlaunch stageは**GA**で、提供ロケーションとして**global**に加えて**us**、**eu**のmulti-regionが記載され、ML processingは**us / eu**をサポートする。Provisioned ThroughputおよびStandard PayGoもglobal/us/euが対象である。したがって、「APIが米国限定で先行公開」というリリースではない。もっとも、Sparkなどコンシューマー製品の利用可否は対応国・アカウント・契約プランに依存する。 ¹³

製品ポジショニングについてGoogle Cloudは3.7 Flashを、深い推論を担うPro系と高スループット・低価格のFlash-Lite系をつなぐ**primary agentic workhorse**と表現している。つまりFlashの競争軸を単純な「小さくて速いモデル」から、**エージェントを大量実行できるだけの知能を、Pro級より低コストで提供するモデル**

へ移している。Reutersも、企業向けの自律システムで計画・ツール利用・多段ワークフローを低コスト化する製品としてGoogleが売り込んでいると整理している。 ¹⁴

技術仕様と公式機能

Gemini 3.7 FlashはGoogleによれば**native multimodal reasoning model**で、Gemini APIではテキスト、画像、動画、音声、PDFを入力でき、出力はテキストである。コンテキスト上限は**1,048,576トークン**、最大出力は**65,536トークン**。thinking levelは low、medium、high の3段階で、デフォルトは medium である。3.6 Flashで利用できた minimal は3.7ではサポートされず、明示するとvalidation errorになる。 ¹⁵

項目	Gemini 3.7 Flashの公式仕様
モデルID	gemini-3.7-flash ¹⁶
提供状態	Stable / GA、本番利用可能。 ¹⁷
基礎モデル	Gemini 3.6 Flash。 ²
パラメータ数・モデルサイズ	未公表 。モデルカードに記載なし。 ²
アーキテクチャ差分	「core reasoning foundationへのalgorithmic improvements」まで公表。層数、MoE構成、attention変更等は 未公表 。 ²
学習データの3.6からの差分	未公表 。3.6モデルカードへ参照を委譲。 ²
知識カットオフ	原則2026年3月。一部分野では2025年1月相当までの場合がある。 ¹⁸
入力	Text / Image / Video / Audio / PDF。 ¹⁶
出力	Text。画像・音声生成は非対応。 ¹⁶
コンテキスト	1,048,576 tokens。 ¹⁶
最大出力	65,536 tokens。 ¹⁶
Thinking	low / medium / high、default=medium。minimal 非対応。 ¹⁹
Function calling	対応。 ¹⁶
Code execution	対応。 ¹⁶
Google Search grounding	対応。 ¹⁶
Google Maps grounding	対応。 ¹⁶
File search / URL context	対応。 ¹⁶
Structured output	対応。 ¹⁶
Computer use	Preview 。 ¹⁶
Context caching	対応。implicit/explicit caching。 ²⁰

項目	Gemini 3.7 Flashの公式仕様
Live API	非対応。 ¹⁶
Fine-tuning / tuning	Google Cloud仕様では非対応。 ²¹
Batch / Flex / Priority	対応。 ¹⁶

モデルサイズと学習データについては、情報が欠けているのではなく、Googleが意図的に詳細を公開していない。 モデルカードのArchitecture、Training Dataset、Training Data Processing、Hardware、Softwareの各節は、いずれも「3.6 Flashを基礎とするので3.6のモデルカードを参照」としており、3.7固有のパラメータ数や学習規模を明らかにしていない。このため、Llama 4のように「17B active / 400B total」のような定量的なモデルサイズ比較はできない。 ²²

一方、品質・レイテンシーのトレードオフはAPI上で明示的に制御できる。Googleは **low** をインシデント対応、リアルタイムチャット、下書き、短時間分析などレイテンシー重視用途、**medium** を通常の複雑なコード・エージェント用途、**high** を難しい推論、数学、コーディング、多段tool use向けとしている。**high** ほどthinkingおよびfunction callにより多くのトークンを使うため、品質の代わりに時間とコストが増える。 ²³

絶対レイテンシーや公式tokens/secはGoogleから公表されていない。 第三者評価のArtificial Analysisでは、Google API上の3.7 Flash **high** を約**340 output tokens/sec**、平均Time per Taskを**1.7分**と測定している。同社の測定ではGPT-5.6 Terra **max** よりタスク時間が約40%短かった。ただし、これはArtificial Analysis固有のワークロード・推論レベル・API測定条件に基づく値であり、GoogleによるSLAや全ユーザーに保証されるスループットではない。 ²⁴

コーディングとエージェント能力

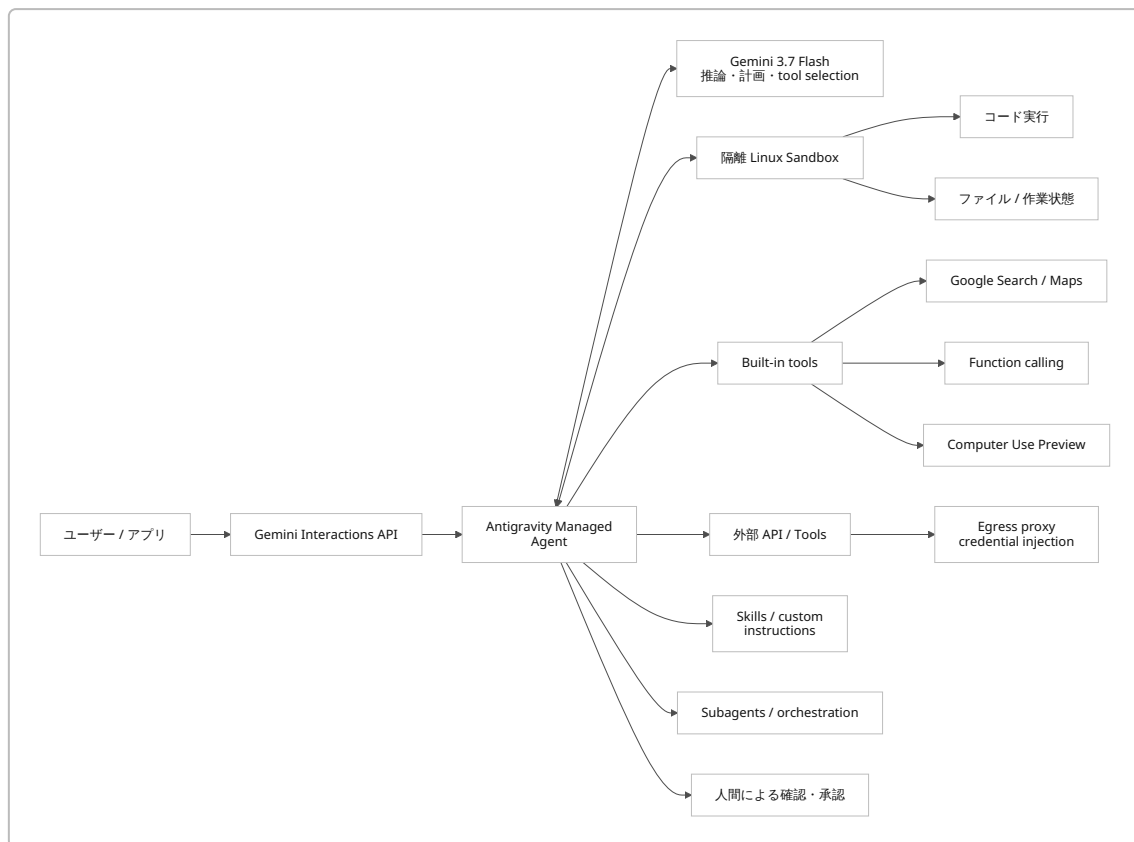
コーディング面でGoogleが最も強調しているのは、単なるコード生成率ではなく、**debugging、issue resolution、first-pass accuracy、production-ready code、長期のsoftware engineering**である。FrontierCode 1.1 Mainは3.6の34.4%から3.7の43.6%へ**+9.2ポイント**、DeepSWE v1.1は48.6%から65.3%へ**+16.7ポイント**改善した。WebDev Arena Eloも1538から1588へ上昇しており、スクリーンショット、画像、design systemを参照したUI生成でdesign adherenceと1:1 parityを重視した改良が入っている。 ²⁵

公式の具体例には、Three.jsによる3Dブラックホール生成、決済処理パイプラインのretry時race conditionを調査してlockを安全に書き直す例、テキスト指示からプレイ可能な3Dゲームを構築し画像資産をNano Bananaで生成する例、既存UIとdesign mockを比較・監査する例などがある。Googleはまた、3.7がroadblockに遭遇した際の適応、意図確認、instruction following、多段planning/tool callを改善し、失敗ループや手動介入を減らす方向で調整したとしている。 ²⁶

対応プログラミング言語の公式な網羅リストは公表されていない。 APIはテキストを生成する汎用モデルであり、「Python、Java、Go等を正式サポート」といった言語別保証表をGoogleは3.7 Flash仕様に載せていない。したがって、特定言語への対応可否をモデル仕様として断定するのは適切ではない。公式サンプルではJavaScript/Three.jsが使われ、Gemini SDKの呼び出し例はPythonとJavaScriptが提示されているが、これは**SDK例であって生成可能なプログラミング言語の限定を意味しない。** ²⁷

GitHubの初期テストでも、3.7 Flashは前世代よりweb/app development、agentic coding、code quality、codebase research、verificationが改善したと報告されている。Copilotでは主要IDEとCLI、cloud agentに段階展開されているため、Google自身のAntigravityだけでなく、既存の開発者ワークフローへ早期に組み込まれている点も重要である。 ²⁸

エージェントについては、**モデルとharnessを分離して考える必要がある**。3.7 Flash自体が提供するののは推論とtool callingの能力であり、実行環境・サブエージェント管理・Webブラウジング・ファイル操作を担うのは主にAntigravity/Managed Agents側である。3.7 Flashは今回、Gemini Managed AgentsおよびAntigravity SDKにおけるAntigravity agentの**新デフォルトモデル**になった。 29



この図はニューラルネット内部のアーキテクチャではなく、**Googleが公開しているサービスレベルのagent stack**を整理したものである。Managed Agentsは1回のAPI callでOSレベルに隔離されたLinux sandboxを用意し、推論、コード実行、ファイル管理、Web閲覧を行える。Antigravity 2.0のharnessには既にsubagents、hooks、非同期task managementなどが導入されており、3.7はそれらを動かすモデル側の品質を高める更新である。 30

外部フレームワークとの接続も広い。GoogleのAgent documentationはLangChain/LangGraph、LlamaIndex、CrewAI、Vercel AI SDK、Google ADK、Antigravity SDKを例示している。Antigravity SDKはGoogle製品と同じagent loopとcontext managementをプログラムから利用できる。したがって3.7 Flashは「Google独自エージェントだけに閉じたモデル」ではなく、既存のagent framework上で基盤モデルとして使える。 31

一方、**3.7固有の新しい永続メモリ機構は公式には発表されていない**。Managed Agentsのenvironment stateはVMが休止しても復元され、7日間利用がないenvironmentは削除されるが、これは実行環境のstate managementであって、3.7 Flashそのものに新たなlong-term memory architectureが追加されたことを意味しない。モデル固有のmemory容量・記憶保持期間・学習型memoryについては**unspecified**である。 31

Gemini Sparkについては3.7へ切り替わることでWorkspace tool use、複数skillをまたぐknowledge work、ファイル統合、メール下書き、status document更新などの精度・出力品質が改善するとGoogleは説明している。Sparkは24時間稼働型personal agentとして位置付けられており、今回の3.7は「チャットの回答品質」だけでなく**継続的に実務を進めるエージェントの経済性**を狙ったアップデートと評価できる。 32

API・SDK・料金・移行

Gemini APIではモデルIDを `gemini-3.7-flash` として利用する。Googleの最新ガイドはPythonの `google-genai` 系インターフェース、JavaScriptの `@google/genai`、RESTの `/v1beta/interactions` を例示しており、Interactions APIを3.7とagentic workflowの中心的な利用方法として案内している。Stable model IDも `gemini-3.7-flash` である。 ²⁷

代表的なPython呼び出しは次の形になる。 ³³

```
from google import genai

client = genai.Client()

interaction = client.interactions.create(
    model="gemini-3.7-flash",
    input="Analyze this codebase and identify the root cause of the failing tests.",
    generation_config={
        "thinking_level": "medium"
    },
)

print(interaction.output_text)
```

3.6 Flashからの移行では、基本的にはmodel IDを3.7へ置き換えることが中心になる。一方、3.5 Flash、Gemini 3 Flash Preview、Gemini 3.1 Proなど古いモデルから直接移行する場合、Googleは `temperature`、`top_p`、`top_k` などの **deprecated sampling parameters** と **prefilled model turns** を削除するよう明示している。3.6ではこれらは既に非対応だったため、3.6→3.7はAPI挙動上の変更が比較的小さい。 ³⁴

実務的なmigrationでは、単にIDだけを差し替えるよりも、`medium` を基準にして既存のcoding/tool-use regression suiteを回し、速度重視経路を `Low`、失敗コストの高い難問を `high` ヘルレーティングする設計が合理的である。これはGoogleが `Low` をlatency-sensitive、`medium` を一般的な複雑コード/agent、`high` を最大推論・tool-use向けとしていることから導かれる運用上の推奨である。 ²³

料金は次の通りである。**thinking tokensも出力料金に含まれる**ため、`high` を多用すると同じユーザー向け出力長でも実コストが増える可能性がある。 ³⁵

推論方式	2026年末までの入力 / 1M tokens	出力 / 1M tokens	2027年1月以降
Standard	\$0.75	\$3.75	\$1.50 / \$7.50 ³⁵
Batch	\$0.375	\$1.875	\$0.75 / \$3.75 ³⁵
Flex	\$0.375	\$1.875	\$0.75 / \$3.75 ³⁶
Standard context cache input	\$0.075	—	\$0.15 ³⁵
Cache storage	\$0.50 / 1M token-hour	—	\$1.00 / 1M token-hour ³⁵

Search groundingはGemini 3.xで共有される月5,000件の無料Search request枠の後、1,000件当たり14ドル、Maps groundingも無料枠後は1,000 search query当たり14ドルとされている。したがって、エージェントの総コストを評価する際は**モデルtoken料金だけでなく、tool call/grounding料金と反復回数**を見る必要がある。 ³⁵

APIのアクセスtierはFree、Paid、Enterpriseに分かれる。3.7 FlashのStandard推論はFree Tierでも利用可能だがrate limitがあり、Paidでは高いrate limit、context caching、Batch API、advanced modelsが利用できる。Enterpriseではさらに専用support、security/compliance、Provisioned Throughput、volume discount等が提供される。特にプライバシー上、Googleの価格ページは**Free Tierのコンテンツは製品改善に使用され得る一方、Paid Tierのコンテンツは製品改善には使用しないと明記**している。 ³⁵

Managed Agentsはモデルtokenとtool usageの従量課金で、1つのinteractionが複数のreasoning loopを起こし、Googleによれば典型的に**100K~3M tokens**を消費する場合がある。Preview期間中はenvironment compute自体は課金されない。これは、単純な「\$3.75/M output」という数字だけで長時間エージェントのコストを見積もると大きく外れる可能性があることを意味する。 ³¹

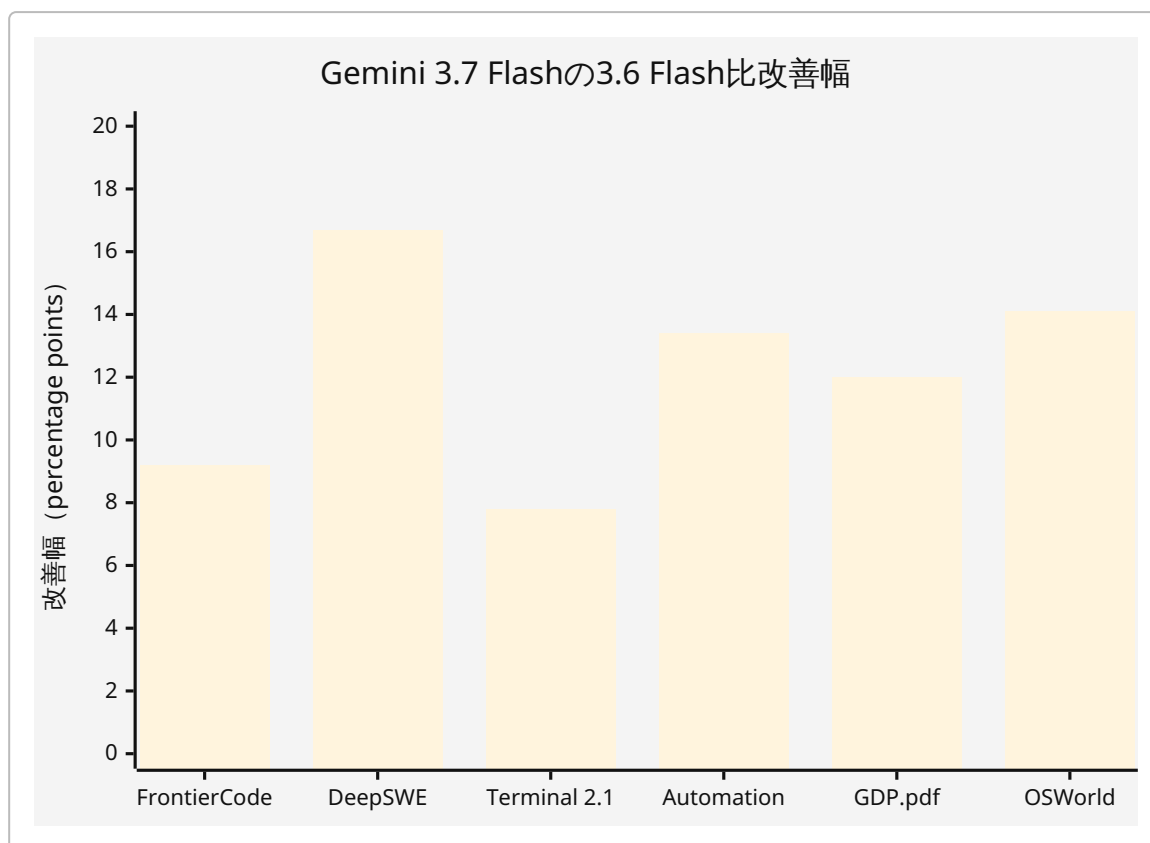
ベンチマークと競合比較

Google DeepMindのモデルカードは、Gemini 3.7 Flash、3.6 Flash、Claude Sonnet 5、GPT-5.6 Terra、Muse Spark 1.2を**同じ表・同じ評価方法**で比較している。以下はその中からコーディング、エージェント、業務処理に関係の深い項目を抜粋したものである。 ³⁷

Benchmark	3.7 Flash	3.6 Flash	Claude Sonnet 5	GPT-5.6 Terra	読み方
Artificial Analysis Intelligence Index	56	52	55	57	総合知能。GPTが僅差上位。 ³⁷
FrontierCode 1.1 Main	43.6%	34.4%	42.7%	41.3%	3.7が最高。 ³⁷
DeepSWE v1.1	65.3%	48.6%	53.8%	69.6%	長期SWEではGPTが上。 ³⁷
WebDev Arena	1588 Elo	1538	1541	1523	Web/UI開発は3.7が最高。 ³⁷
Terminal-bench 2.1	85.8%	78.0%	80.4%	87.4%	terminal agentではGPTが僅かに上。 ³⁷
Terminal-bench 3.0	14.9%	5.4%	14.6%	20.8%	難しい一般agent能力では依然差。 ³⁷
AutomationBench	30.4%	17.0%	10.7%	23.6%	業務自動化で3.7が最高。 ³⁷
GDPVal-AA v2	1525 Elo	1422	1598	1578	広範なknowledge workではClaude/GPT優位。 ³⁷
Harvey LAB-AA	90.7%	85.1%	90.1%	85.2%	複雑なlegal workflowで僅差首位。 ³⁷

Benchmark	3.7 Flash	3.6 Flash	Claude Sonnet 5	GPT-5.6 Terra	読み方
GDP.pdf	34.0%	22.0%	28.0%	24.7%	複雑PDF処理で首位。 37
OSWorld-2.0	47.9%	33.8%	—	50.2%	computer-useではGPTが 上。 37
Agent's Last Exam	26.3%	24.2%	33.3%	28.0%	desktop/OS agentでは Claudeが上。 37
GDM-MRCR v2 128K	97.0%	91.8%	81.5%	93.5%	long-context retrievalで 3.7が上。 37

3.6からの改善幅だけを見ると、コーディング・エージェント指標が広範囲で上がっている。以下のグラフは、同一モデルカードでpercentage表記されている代表評価の**絶対percentage-point差**であり、相対%増加ではない。 37



この結果から、3.7を「3.6の小規模ポイントアップ」と見るのは適切ではない。例えばDeepSWEは+16.7ポイント、AutomationBenchは+13.4ポイント、OSWorldは+14.1ポイントであり、3週間の更新としてかなり大きい。ただしGoogle自身が「developer feedbackとalgorithmic innovations」の成果としか説明していないため、改善が追加学習、post-training、推論制御、tool-use policy、data mixtureのどれに何%起因するかは**未公表**である。 38

第三者評価でも同じ方向性が確認できる。Artificial Analysisは3.7 Flash **high** をIntelligence Index **56**と評価し、3.6より4ポイント上昇、約340 output tokens/sec、平均1.7分/taskとしている。同社は3.7をIntelligence

vs. Time per TaskのPareto frontier上と評価しており、性能上昇の中心はTau3 Banking、Terminal-Bench、GDPval-AAなどagentic evaluationにあると分析している。³⁹

価格・コンテキストも含む競合位置を整理すると次のようになる。なお、GPT-4oとLlama 4はGoogleの3.7モデルカードの同一benchmark suiteには含まれないため、**同じbenchmarkの数字を無理に横並びにはしていない**。GPT-4oについてはOpenAI公式仕様、Llama 4についてはMeta公式モデルカードとの仕様比較に限定する。⁴⁰

モデル	入力 / 1M	出力 / 1M	Context	Max output	備考
Gemini 3.7 Flash	\$0.75*	\$3.75*	1.048M	65,536	*2026年末まで。2027年から\$1.50/\$7.50。 ³⁵
Gemini 3.6 Flash	\$0.75*	\$3.75*	約1M	64K級	3.7発売に合わせ同じ期間限定価格を適用。 ⁴¹
Claude Sonnet 5	\$2.00	\$10.00	1M	128K	Adaptive thinking、coding/agentic重視。 ⁴²
GPT-5.6 Terra	\$2.00	\$12.00	1.05M	128K	functions/web/file/computer use対応。 ⁴³
GPT-4o	\$2.50	\$10.00	128K	16,384	text/image input、text output。現在の3.7と世代差が大きい。 ⁴⁴
Llama 4 Maverick	provider 依存	provider 依存	1M	Metaカードでは統一API価格なし	17B active / 400B totalのMoE、weights提供型。 ⁴⁵
Llama 4 Scout	provider 依存	provider 依存	10M	同上	17B active / 109B total、16 experts。 ⁴⁵

Llama 4との比較では透明性の違いが際立つ。MetaはScoutを17B active/109B total、Maverickを17B active/400B totalの**Mixture-of-Experts、early-fusion native multimodal architecture**と公開し、学習トークン数もそれぞれ約40T、約22Tと明示している。一方Gemini 3.7 Flashでは同種のパラメータ情報が公表されていない。したがって、性能差を「Geminiの方が何倍大きい/小さい」とモデルサイズから説明することはできない。⁴⁶

GPT-4oはユーザー指定の比較対象として重要だが、2026年現在の直接的な競合基準としては古い。OpenAIの現行仕様ページではGPT-4oは128K context、16,384 max output、knowledge cutoff 2023-10-01であり、Googleの3.7評価はより新しいGPT-5.6 Terraを比較相手にしている。したがって「Gemini 3.7 FlashがGPT-4oをbenchmark Xで何%上回る」といった主張を、今回確認した一次資料から行うことはできない。⁴⁷

総じて3.7 Flashの競争力は、「**最高スコアだけを追うモデル**」ではなく、「**Claude Sonnet 5/GPT-5.6 Terra級の領域に食い込みつつ、期間限定価格では大幅に安く、しかも推論スループットが高い**」点にある。ただしDeepSWE、Terminal-bench、OSWorld、Agent's Last Exam、GDPVal-AAなどを見ると競合に明確に負けるタスクもあり、モデルrouterを使わず3.7だけですべてを処理することが常に最適とは限らない。⁴

セキュリティ・プライバシー・適用領域と限界

Googleは3.7 FlashについてCBRNおよびcyber offense悪用へのFrontier Safety safeguardsを更新した。モデルカードでは専門red teamを実施し、child safetyのlaunch thresholdを満たし、一般content safetyでも3.6と同等または改善したと説明している。またFrontier Safety Framework評価では、CBRN、cybersecurity、harmful manipulation、ML R&D/misalignmentについていずれも最終的なCritical Capability Levelには達していないとしている。 ⁴⁸

ただし詳細を見ると安全性は単純な「全面改善」ではない。3.6比の自動評価で、Multilingual Safetyは**-0.48pp（低いほど良いので改善）**、Image-to-Text Safetyは変化なしだった一方、Text-to-Text Safetyは**+1.17pp（低いほど良いので悪化）**、Unjustified Refusalsも**+0.84pp（低いほど良いので悪化）**、Toneは-0.47ppで同指標ではhigher-is-betterのため悪化方向である。Google自身は総合的には3.6と類似しており、手動レビューでは悪化フラグの大半がfalse positiveまたは深刻でないケースだったとしている。評価セットと方法が更新されているため、さらに古いモデルカードとの直接比較はできない。 ⁴⁹

Frontier Safetyで特に注目すべきなのは、**cybersecurityではCCLそのものには未達だが、そのalert thresholdには到達した**という記述である。CBRNでも一部のexpert red teamingで完全なharm journeyに沿ったactionable informationを引き出したためCCL alert threshold相当と評価したものの、平均スコアやexpert steering依存性からCCL自体には未達と判定した。Googleはこれを理由に予防的mitigationを継続している。8月14日時点で、詳細な「Gemini 3.7 Frontier Safety Framework Report」はモデルカード上では**“will be published shortly”**とされ、まだ後続公開予定である。 ⁵⁰

エージェントでは別種の安全問題がある。Managed AgentはOSレベルで隔離されたsandboxで動く一方、**outbound networkはデフォルトで unrestricted**である。Googleはdomain allowlistで通信先を制限できるようにし、外部APIについてtrusted toolのみを使うこと、最小権限service account/API key、short-lived token、定期的credential rotationを推奨する。Credentialはegress proxyのheader transformation経由で注入でき、sandbox内部に直接露出させずに済む。しかしGoogle自身が「agentはアクセス可能なcredentialを使える」と警告しているため、権限設計そのものが防御境界になる。 ⁵¹

Googleはさらに、生成コード、data transformation、configuration changeを本番に反映する前に**人間が検証すること**を明示的に推奨している。これは3.7の性能上昇でhuman-in-the-loopが不要になったことを意味しない。むしろAutomationBenchやOSWorldのスコアが大幅に改善し、「実際に操作できる」範囲が広がるほど、誤ったtool callのblast radiusも大きくなる。 ⁵²

Enterprise Agent Platformではonline prediction、batch inference、context cachingについて**data residency、CMEK、VPC Service Controls、AXT**がsecurity controlsとして記載される。データ所在地を厳密に管理する企業では、単に `global` endpointを使うのではなく、`us` / `eu` multi-regionおよび組織側のcompliance設計と組み合わせる必要がある。 ⁵³

Gemini Developer APIのデータ利用については、Free TierではコンテンツがGoogle製品改善に使用され得る一方、Paid Tierでは「Content not used to improve our products」と明記される。機密ソースコード、顧客情報、社内文書を扱う場合、Free APIとPaid/Enterpriseを同じプライバシー条件だと考えてはいけない。なお、これは**Developer API tierに関する規定**であり、Geminiアプリ/Spark全体のconsumer privacy policyが3.7で特別に変更されたという発表ではない。3.7固有のconsumer privacy変更は今回確認した公式資料では**未公表**である。 ³⁵

適用領域として最も合理的なのは、既存コードベースのdebugging・issue解消・migration、design mockからのUI実装と差分監査、Web/terminalを使う開発エージェント、PDFを含む企業文書分析、CRM/SaaSなど複数toolをまたぐ業務自動化、法務・金融・bioinformaticsの調査補助、Workspace中心のpersonal

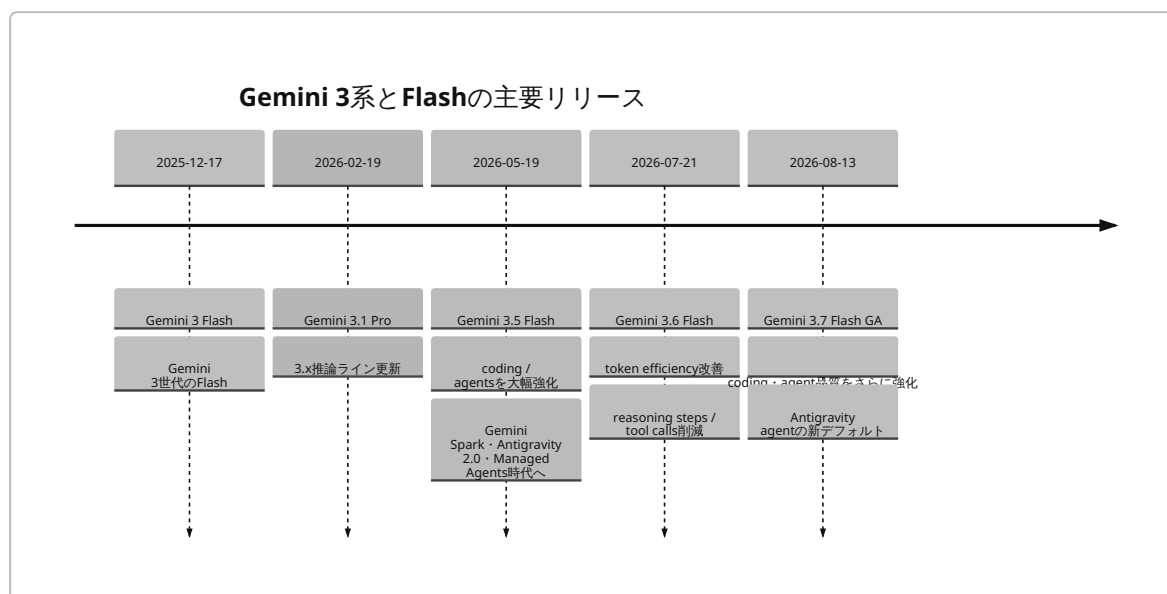
knowledge-work agentなどである。これらはGoogleが示したベンチマークと公式デモの双方で3.7の改善が確認される領域である。 54

一方、主要な限界は明確である。Google自身が**hallucination、jailbreak resistanceの継続課題、時折発生するslowness/timeout**を既知の限界として列挙している。知識カットオフは原則2026年3月で、一部分野では2025年1月までしか内在知識がない可能性がある。さらにLive API、native image generation、audio generation、tuningは非対応で、リアルタイム音声エージェントや画像生成用途には別モデルが必要である。 55

技術評価上の最大の不確実性は、**モデル内部の透明性不足**である。3.7固有のパラメータ数、architecture topology、訓練token数、data mixture、post-training recipe、compute量、絶対的なGoogle公式latency/tokens-per-second、プログラミング言語別成功率、persistent memory方式はいずれも公開資料からは分からない。したがって、今回確認できるのは「何を変えた結果性能が上がったか」のうち、結果と大まかな reasoning improvementまでで、「なぜその改善が生まれたか」を再現可能な技術詳細までは開示されていない。 56

リリース年表と主要ソース

Gemini 3.xの流れを見ると、Googleは2026年に入って特にFlash系の更新周期を急速に短縮している。Gemini 3 Flashが2025年12月17日、3.1 Proが2026年2月19日、3.5 Flashが5月19日、3.6 Flashが7月21日、3.7 Flashが8月13日である。3.5ではagent-firstな方向が明確になり、I/O 2026でSpark、Managed Agents、Antigravity 2.0が前面に出た。3.6はtoken efficiency・少ないreasoning/tool-call loop、3.7はその基盤上でcoding/agent qualityを大きく引き上げる、という進化経路になっている。 57



この約3カ月の流れから読み取れるのは、GoogleがFlashを単なる廉価モデルではなく、**大量のagent loopを高速・低コストで回す主力実行モデル**へ再定義している点である。3.5ではAntigravityとsubagent、3.6ではtoken/tool-call効率、3.7ではissue resolution、production coding、workflow automationの成功率を強化しており、モデル品質・agent harness・推論コストの三者を同時最適化する戦略が一貫している。 58

主要一次資料・公式資料

ソース	種別	主な用途
Google Blog — Introducing Gemini 3.7 Flash	Google公式・製品発表	発表日、狙い、coding/agent改善、Spark、価格、安全対策。 59
Google DeepMind — Gemini 3.7 Flash Model Card	Google DeepMind公式・技術/安全性一次資料	architecture開示範囲、training情報、benchmark、limits、safety。 60
Gemini API — Gemini 3.7 Flash model page	Google公式・開発者仕様	context、modalities、tools、thinking、API capabilities。 16
Gemini API — What's new in Gemini 3.7 Flash	Google公式・migration/API guide	API例、thinking levels、移行、Antigravity default化。 23
Gemini Developer API Pricing	Google公式・価格/データ利用条件	Standard/Batch/Flex料金、cache、Free/Paid/Enterprise。 35
Gemini API — Agents Overview	Google公式・Agent技術資料	Linux sandbox、external tools、credential、framework、human oversight。 51
Google Cloud — Gemini 3.7 Flash	Google Cloud公式・Enterprise仕様	GA、regions、Provisioned Throughput、security controls。 61
Google DeepMind — Model cards index	Google DeepMind公式	Gemini各世代の公開日・更新日の確認。 62
Google — Gemini 3.5: frontier intelligence with action	Google公式・過去リリース一次資料	3.5 Flash、Spark、agentic transition。 63
Google — Gemini 3.6 Flash / 3.5 Flash-Lite / Cyber	Google公式・過去リリース一次資料	3.6のtoken efficiencyとtool-call改善。 64
Google I/O 2026 announcements	Google公式・Agent platform資料	Antigravity 2.0、subagents、hooks、Managed Agents。 65
Artificial Analysis — Gemini 3.7 Flash	独立第三者ベンチマーク	約340 tokens/s、Time per Task、quality/cost Pareto分析。 39
Reuters — Google unveils Gemini 3.7 Flash	主要英語報道	公開時刻の外部アンカー、企業市場での位置付け。 10
窓の杜 — Google 「Gemini 3.7 Flash」を発表	主要日本語テック媒体	日本時間での報道、国内向け整理。 11
GitHub Changelog — Gemini 3.7 Flash in GitHub Copilot	GitHub公式・開発ツール一次資料	Copilot/IDE提供範囲、初期coding評価。 28
Anthropic — What's new in Claude Sonnet 5	Anthropic公式・競合一次資料	価格、1M context、128K output、thinking。 42
OpenAI — GPT-4o model	OpenAI公式・競合一次資料	GPT-4o context、料金、modalities。 44
OpenAI — Models / GPT-5.6 Terra	OpenAI公式・競合一次資料	GPT-5.6 Terraの価格、context、tools。 43

ソース	種別	主な用途
Meta — Llama 4 Model Card	Meta公式・競合モデルカード	MoE構成、parameter数、training tokens、context。 45

本調査で確認できた3.7固有の最も詳細な公式技術資料は**Google DeepMindのモデルカード**であり、そこでもparameter count、ニューラルarchitectureの具体的変更、3.7固有training corpus/computeは開示されていない。したがって現時点で最も堅い結論は、Gemini 3.7 Flashは「巨大化したFlash」ではなく、**3.6を土台に推論・software engineering・tool-use loopを集中的に改善し、Antigravity/Managed Agentsという実行基盤と低価格APIを組み合わせることで、コーディングと実務エージェントの“量産可能な主力モデル”を狙ったリリース**だということである。 66

-
- 1 3 8 9 12 25 32 38 48 54 59 **Gemini 3.7 Flash: our most intelligent workhorse model**
<https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/>
- 2 4 7 18 22 37 49 50 55 56 60 66 **Gemini 3.7 Flash - Model Card — Google DeepMind**
<https://deepmind.google/models/model-cards/gemini-3-7-flash/>
- 5 6 17 19 23 26 27 29 33 34 41 **What's new in Gemini 3.7 Flash | Gemini API | Google AI for Developers**
<https://ai.google.dev/gemini-api/docs/latest-model>
- 10 **Google unveils Gemini 3.7 Flash AI model for coding, agent workflows | Reuters**
<https://www.reuters.com/business/google-unveils-gemini-37-flash-ai-model-coding-agent-workflows-2026-08-13/>
- 11 **Google、「Gemini 3.7 Flash」を発表 ~わずか3週間での更新、性能アップで年末まで半額 - 窓の杜**
<https://forest.watch.impress.co.jp/docs/news/2132672.html>
- 13 14 20 21 53 61 **Gemini 3.7 Flash | Gemini Enterprise Agent Platform | Google Cloud Documentation**
<https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-7-flash>
- 15 16 **Gemini 3.7 Flash | Gemini API | Google AI for Developers**
<https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash>
- 24 39 **Gemini 3.7 Flash: On the Intelligence vs. Time per Task Pareto frontier**
<https://artificialanalysis.ai/articles/gemini-3-7-time-frontier>
- 28 **Gemini 3.7 Flash is now available in GitHub Copilot - GitHub Changelog**
<https://github.blog/changelog/2026-08-13-gemini-3-7-flash-is-now-available-in-github-copilot>
- 30 31 51 52 **Agents Overview | Gemini API | Google AI for Developers**
<https://ai.google.dev/gemini-api/docs/agents>
- 35 36 **Gemini Developer API pricing | Gemini API | Google AI for Developers**
<https://ai.google.dev/gemini-api/docs/pricing>
- 40 44 47 **GPT-4o Model | OpenAI API**
<https://developers.openai.com/api/docs/models/gpt-4o>
- 42 **What's new in Claude Sonnet 5 - Claude Platform Docs**
<https://docs.anthropic.com/en/docs/about-claude/models/whats-new-sonnet-5>
- 43 **Models | OpenAI API**
<https://developers.openai.com/api/docs/models>

45 46 llama-models/models/llama4/MODEL_CARD.md at main · meta-llama/llama-models · GitHub

https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md

57 62 Model cards — Google DeepMind

<https://deepmind.google/models/model-cards/>

58 63 Gemini 3.5: frontier intelligence with action

<https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/>

64 3.6 Flash, 3.5 Flash-Lite, and 3.5 Flash Cyber

<https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/>

65 100 things we announced at Google I/O 2026

<https://blog.google/innovation-and-ai/technology/ai/google-io-2026-all-our-announcements/>