Google I/O 2026: Gemini 3.5 Flash と「エージェンティック AI」の衝撃

Passive Assistant

Autonomous Agent
(エージェンティック AI)

Gemini 3.5 Flash: 次世代エージェントの心臓部

「速度」と「知能」のトレードオフを解消

Mixture of Experts (MoE) アーキテクチャを採用。アクティブパラメータを100億〜160億に抑え、フラグシップ等、約280トークンの超高速出力を実現。

推論時パラメータ数: 2508〜300B

100万トークンの巨大なコンテキストウィンドウ

巨大なソースコードやAPIドキュメントを一度に理解。外部ツールへのアクセスを最小限に抑えた長期的で複雑なタスク遂行が可能。

最大入力: 1,048,576 トークン (1M)
最大出力: 68,935 トークン

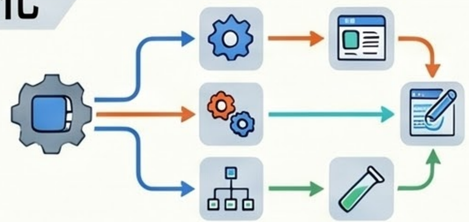
Dynamic Thinking（動的思考レベル）の導入

タスクの複雑度に応じて推論認定を4段階（MINIMAL〜HIGH）で推論可能。デフォルトは「MEDIUM」で、速度とコストのバランスを最適化。

Antigravity 2.0: ソフトウェア工場の自動化

スタンドアロンの「中央司令室」へ進化

単なるIDE拡張から AIエージェントの作業を監視、指揮するための役立したデスクトップアプリへと刷新。



動的サブエージェントによる並列処理

メインエージェントがタスクを遂行。「ロジック」「UI」「テスト」などの専門サブエージェントを生成し、コンテキストを汚染せずにバックグラウンドで並列処理。

12時間でOSコアを構築

93個のサブエージェントが連携、26億トークンの通信を経てOSコアを構成。数ヶ月かかる開発工程を劇的に短縮する「ソフトウェア工場」の実証に成功。

パフォーマンス: フラグシップを凌駕する数値

Gemini 3.1 Proを超えるベンチマークスコア

コーディングやマルチモーダル理解において前世代の最上位モデルを上回る、「置くて違いモデル」でなくても標準な基準解決が可能。

CharXiv
84.2%
チャートや科学の図表のマルチモーダル能力

MCP Atlas
83.6%
複雑ツールを押手機種なワークフロー処理

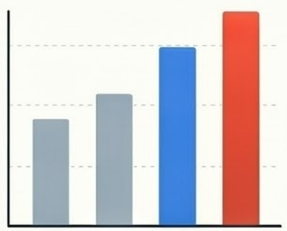
Terminal-Bench 2.1
76.2%
CLI環境でのエージェント実行助力

AI Analysis Index
55
同クラス平均 36 を大幅に上回る

コストの整: 進化の代償

API価格が前モデルの3倍に高騰

入力$1.50/1Mトークン、出力$0.00/1Mトークンに設定、「Flash＝安価」という従来のフラッドイメージが崩壊、上位のProモデルに実現する価格設定。



コストの雪だるま式増大と効率性の課題

- Gemini 3.5 Flash (知能スコア 27, コスト $172, 比率 1.0x)
- Gemini 3.5 Flash Preview (知能スコア 46, コスト $278, 比率 1.6x)
- Gemini 3.5 Flash (知能スコア 25, コスト $1,551, 比率 9.0x)

他社モデルの2倍近いトークンを出力する傾向 (Verbosity) があり、実質用コストが予想に増加するリスク。

Pelican SVGテストでの構造的入場

構造な装備を追加してトークンを大量頂戴した一方、ベタルと重複を驚く「根本的な構造」を描き損たるなど、推理の効率性に課題。