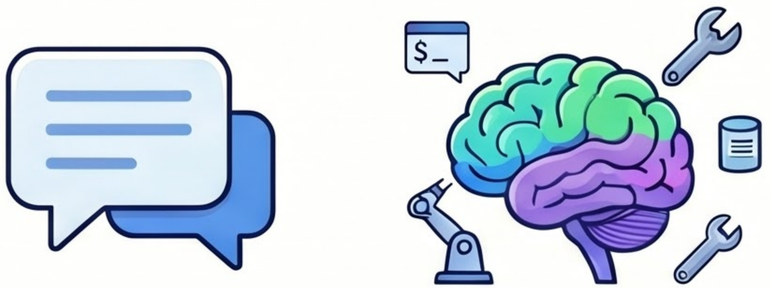
Gemini 3.5 Flash：高速性と実行力を兼ね備えた「エージェント時代」の主力モデル

コンセプト：チャットから「エージェント」へ



Frontier Intelligence with Action（実行を伴う最先端知能）
単なる対話ではなく、サブエージェント展開、問題解決、長期的なツール利用など、自律的な「エージェント」としての実行能力を重視した設計です。

「高速・廉価」から「高速・高性能」モデルへのシフト
従来のFlashモデルのイメージを刷新し、速度を収益性やUKに直結させるエージェント処理向けのハイエンドな位置づけに変更されました。

Google Search AI Mode　Google Antigravity　Android Studio　Gemini Enterprise Agent Platform

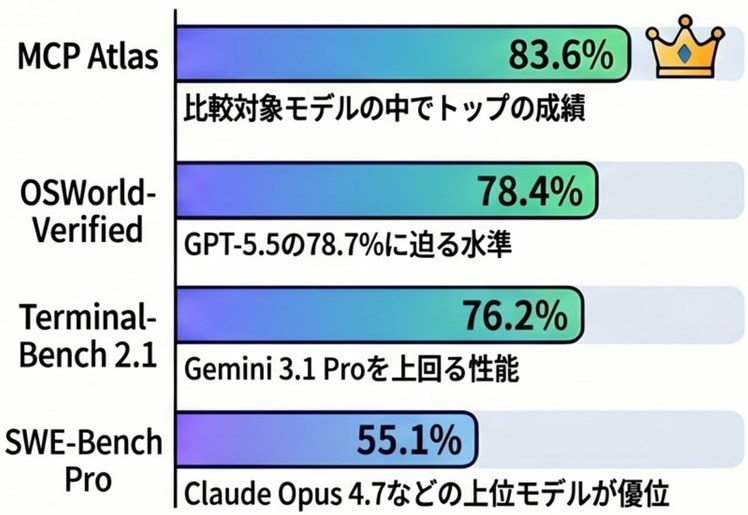
広範なGoogleエコシステムへの統合
これらの各プラットフォームで即座に利用可能です。

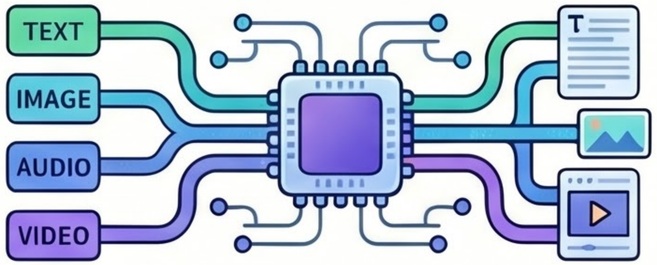
技術仕様と圧倒的なスピード

277.0トークン/秒

驚異的な出力速度 Artificial Analysisの調査で147モデル中2位を記録。他のフロンティアモデルの約4倍の速度で処理が可能です。

主要なベンチマークスコアによる他モデルとの比較





1Mトークンのコンテキストとマルチモーダル対応
テキスト、画像、音声、動画をネイティブに処理し、膨大なドキュメントや長期にわたるワークフローの中間推論を保持できます。

高度なエージェント・コーディング機能
反復的なコーディングや長期的なツール利用、サブエージェントの展開が安定して実行可能なGA（一般提供）版として提供されています。

導入における懸念点とコストの現実

Flash系としては異例の「価格上昇」
有料API価格は入力$1.50/1M、出力$9.00/1M。旧Flash Previewの3倍、Flash-Liteの6倍と、Proモデルに近い価格帯へ。

「出力の冗長性」による総コストの増大
他モデルに比べ出力が冗長な傾向があり、評価実行コストが1,551.60ドルに達するなど、単価以上の費用が発生するリスクがあります。

安全性・倫理面の課題
自由度の高い自律エージェントの提供に伴う誤用リスクに加え、モデルカード上の安全性評価で前モデルより一部数値が悪化しています。