Gemini 3 Flash:「知能」と「速度」のトレードオフを破壊する

Pro級の推論能力を、Flashの速度とコストで。
AI開発の新時代が始まる。

これまでのAI開発における「妥協」という名の壁

開発者は常に選択を迫られてきました。高度な思考能力を持つが、遅く、高価な「Pro」モデルか。あるいは、高速で安価だが、複雑なタスクには力力不足な「Flash」モデルか。この二者択一が、イノベーションの足枷となっていました。

Proモデル

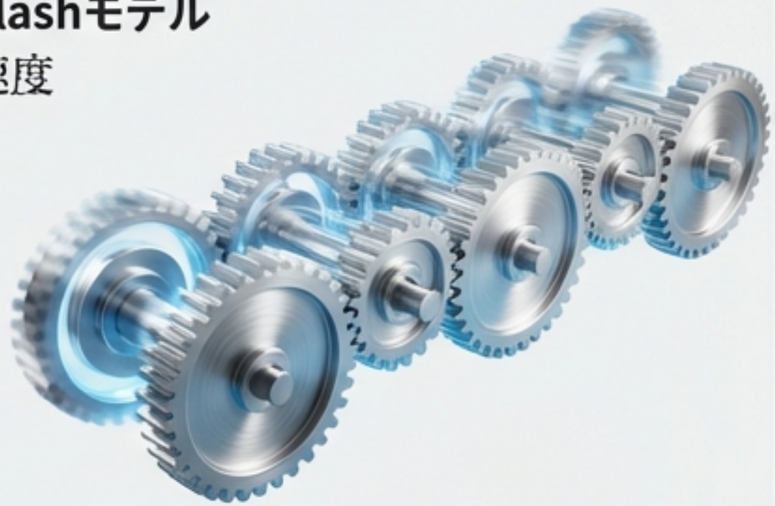
知能

- 高度な推論
- 複雑なタスク
- 高コスト
- 高レイテンシ

Flashモデル

Flashモデル

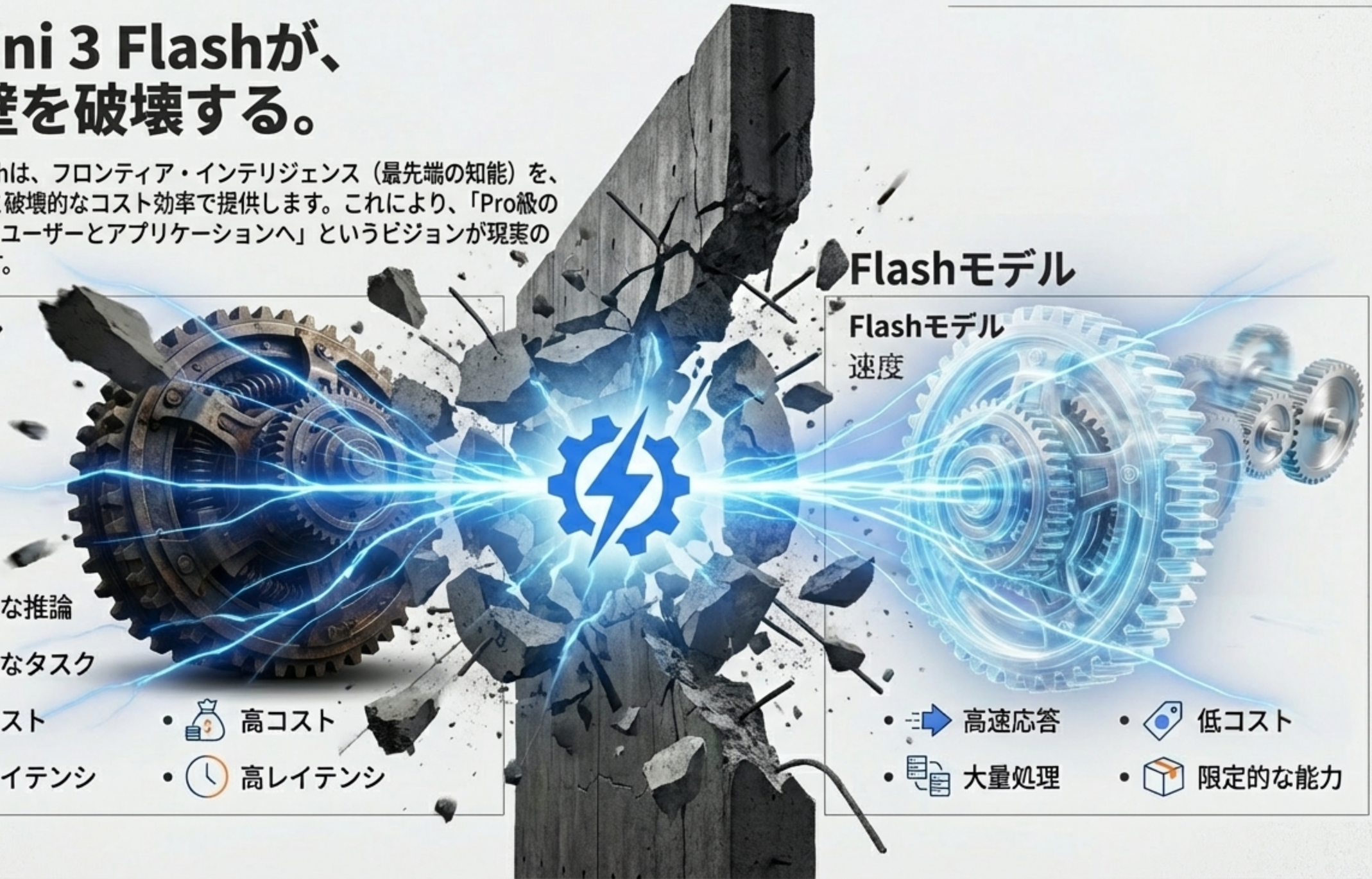
速度



- 高速応答
- 低コスト
- 大量処理
- 限定的な能力

Gemini 3 Flashが、その壁を破壊する。

Gemini 3 Flashは、フロンティア・インテリジェンス（最先端の知能）を、驚異的な速度と破壊的なコスト効率で提供します。これにより、「Pro級のAIを、すべてのユーザーとアプリケーションへ」というビジョンが現実のものとなります。



Proモデル

知能

- 高度な推論
- 複雑なタスク
- 高コスト
- 高レイテンシ

Flashモデル

Flashモデル

速度

- 高速応答
- 低コスト
- 大量処理
- 限定的な能力

高コスト

高レイテンシ

性能を支える3つの柱

1. Pro級の知能

博士課程レベルの推論能力と、Proモデルを超えるコーディング性能。

2. 前例のない速度

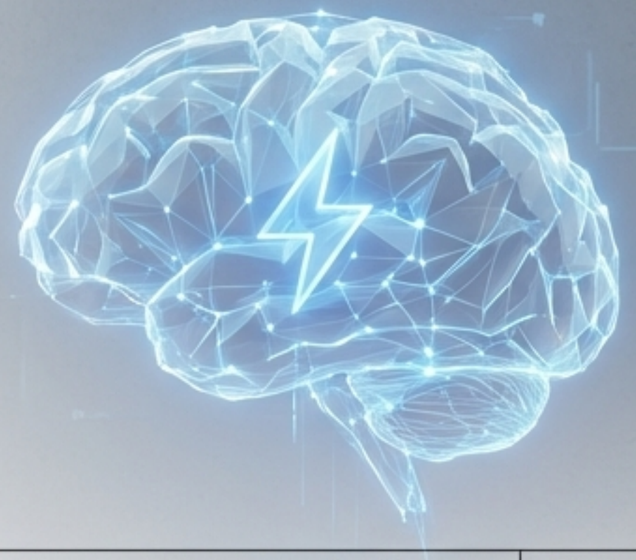
Gemini 2.5 Proの3倍高速。リアルタイム性が求められるタスクに最適。

3. 破壊的なコスト

知能あたりのコストパフォーマンスで市場をリード。大規模な実装を可能に。

Pillar 1: ベンチマークが証明する「博士課程レベル」の知能

Gemini 3 Flashは、速度のために知能を犠牲にしません。複数の主要なベンチマークで、既存の大型モデルに匹敵、あるいはそれを凌駕するスコアを記録しています。



GPQA Diamond（大学院レベルの科学的質問）

90.4%

「博士課程レベルの推論能力」

SWE-bench Verified（エージェント型コーディング能力）

78%

Gemini 3 Proをも上回るスコア

Artificial Analysis（総合性能）

71

Opus 4.5 (70) を超え、最上位モデルに肉薄

新しい制御次元：「思考レベル（Thinking Levels）」

開発者は API を通じて、モデルの「思考の深さ」を動的に調整可能に。
これにより、単一のモデルで応答速度と回答精度をタスクに応じて最適化し、
コストパフォーマンスを最大化できます。

Minimal (最小思考)

Use Case: チャットの応答、単純なデータ抽出
Benefit: 超低レイテンシ、コスト最小化

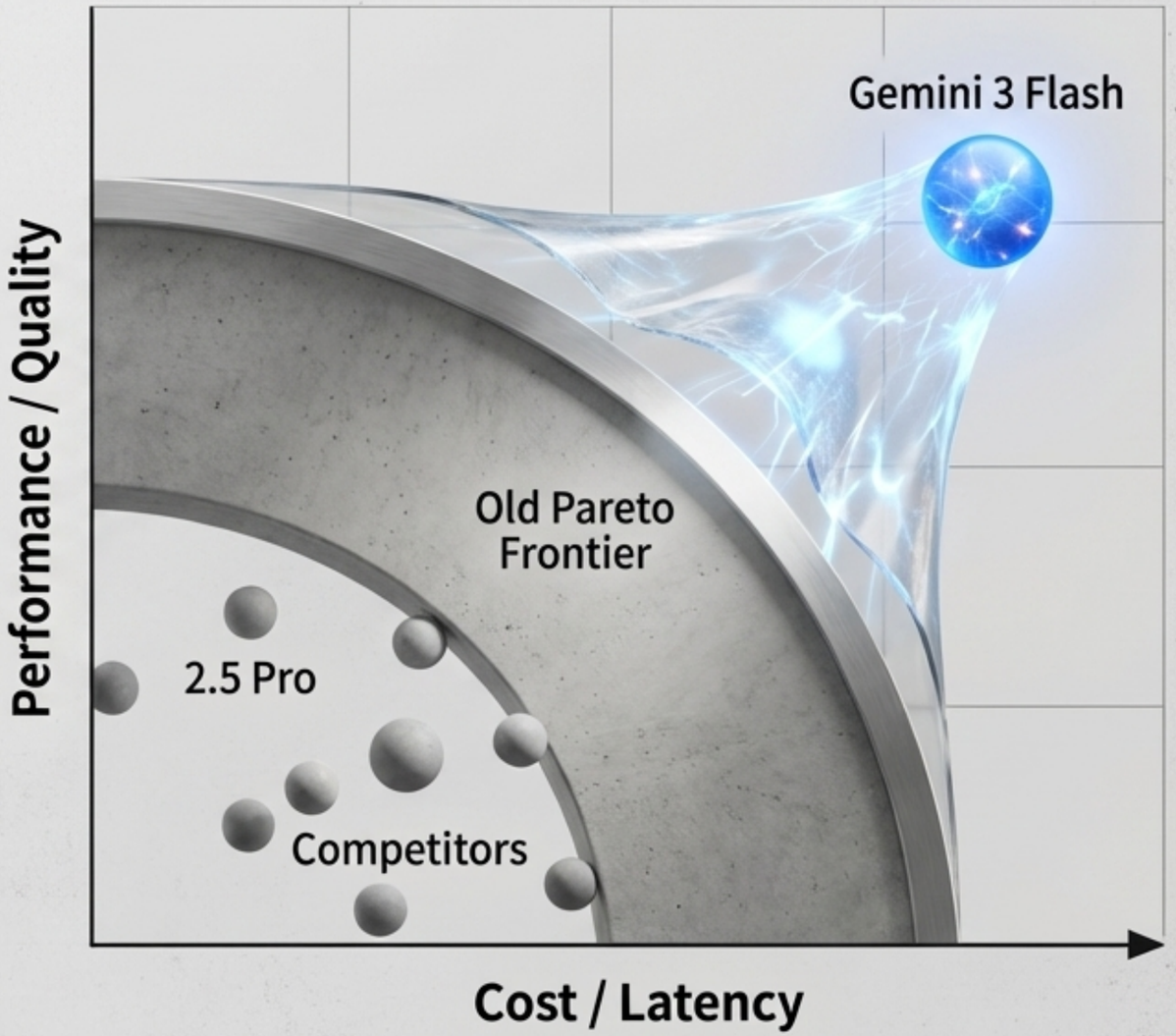
High (高度思考)

Use Case: 複雑な数学問題、法務分析
Benefit: Proに匹敵する精度

Pillars 2 & 3: パレートフロンティアを押し上げる速度とコスト

Gemini 3 Flashは、品質、速度、コストのバランスを新たな次元へと引き上げます。平均してGemini 2.5 Proより30%少ないトークンで、3倍高速に、より高品質な結果を提供します。

料金: $0.50 / 100万入力トークン



開発者にとっての変化：
ツールから「自律型パートナー」へ

Gemini 3 Flashの登場により、AIはもはや命令を待つだけのツールではありません。開発者の意図を汲み取り、自ら計画し、実行し、修正する「自律型エージェント」として、開発プロセスに深く統合されます。

「自律型エージェント」

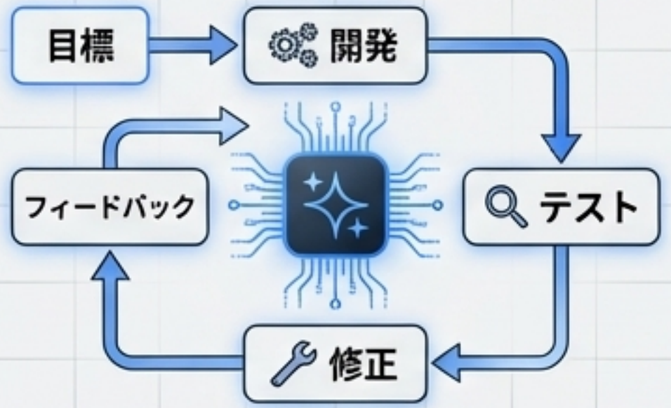
「バイブコーディング」

「リアルタイム・コア・ループ」

ユースケース①：エージェント主導のコーディングとUIプロトタイピング

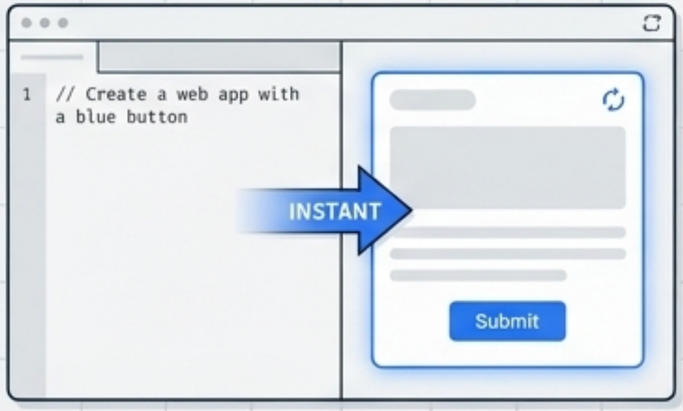
SWE-bench 78%の能力と低レイテンシを活かし、AIが開発の「コア・ループ」を回します。
タイピングを妨げない速度で、リアルタイムにバグを修正し、UIを生成します。

Google Antigravity



目標を投げるだけで、エージェントが自律的に開発・テスト・修正を繰り返す

Google AI Studio (Build Mode)



自然言語で指示するだけで、プレビュー画面のWebアプリが秒単位で更新される

Figma Integration

Figma上で『ボタンを追加して』と指示すると、動作するプロトタイプが即座に生成される

ユースケース②：高度な空間推論と100万トークンのデータ処理

動画、画像、そして数百万語のドキュメントを人間のように理解。
静的な情報が、インタラクティブな洞察に変わります。

リアルタイム動画解析

ゲーム内でのリアルタイムAIアシスタンス

100万トークンの契約書分析

数千ページの契約書群から、矛盾する条項をPro級の精度
かつFlashのコストで瞬時に抽出

Google製品全体へのインパクト：「知能の民主化」

Gemini 3 Flashの効率性により、これまで一部の機能に限定されていた最先端のAIが、Googleの主要製品で「全ユーザーへデフォルト解放」されます。これは「補助ツール」から「代理人（エージェント）」への質的な転換を意味します。

「質問に答えるツール」から
「意図を汲んで仕事を終わらせるエージェント」へ

あなたが毎日使うツールは、こう変わる

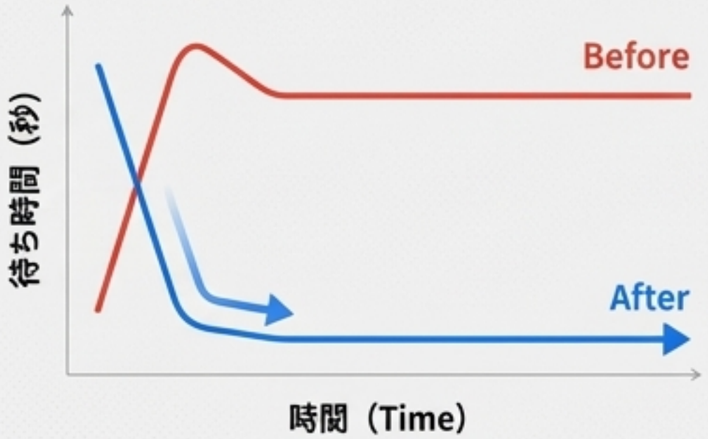
Google Search (AI Mode)

Before

AIによる回答には数秒の待ち時間があった

After

従来の検索と変わらない速度で、動的なUI（表や比較リスト）をその場で生成



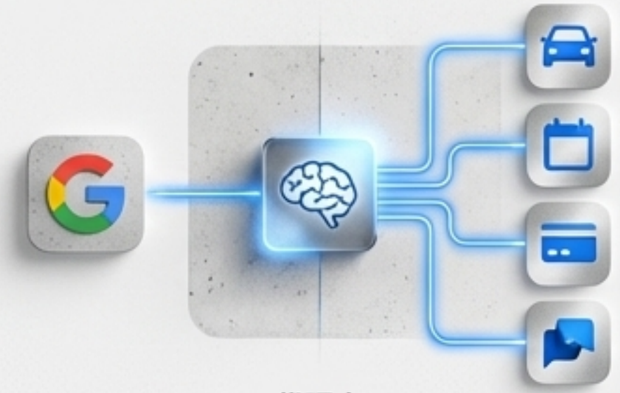
Google Workspace

Before

要約や下書きの「補助」

After

「メールを元にレンタカーを予約して」といった外部サービスを跨ぐタスクを実行する「代理人」



NotebookLM

Before

資料の「要約ツール」

After

数百の論文から矛盾点を指摘し、議論をしながら図解を即時生成する「リサーチ・パートナー」

状況別モデル選択：Gemini 3 Flash vs. ChatGPT 5.2

どちらのモデルも強力ですが、得意な領域が異なります。「何を達成したいか」に応じて最適なツールを選択することが重要です。

比較項目	Gemini 3 Flash	ChatGPT 5.2
得意なこと	**大量処理、高速応答、高コストパフォーマンス**	専門的な知恵、洗練された文章作成
コンテキスト窓	**100万トークン**	256k トークン
APIコスト	**圧倒的に低い**	数倍〜10倍以上
レイテンシ	**リアルタイム**	精度重視のため変動
エコシステム	**Google製品とネイティブ統合**	幅広いサードパーティ連携

ジョブに最適なツールを選ぶ

「とにかく速く、安く、
大量に、正確に処理したい」

- 自律エージェント
- 動画・コード解析
- 大規模システム
- リアルタイム対話

Gemini 3 Flash

「最高の知恵を借り、成果物
の質を極限まで高めたい」

- 経営判断の壁打ち
- 企画立案
- 完璧な一通のメール
- 創作活動

ChatGPT 5.2

今すぐ、未来の開発を始めよう

Gemini 3 Flashは、本日よりプレビュー版として主要なプラットフォームで利用可能です。
あなたのアイデアを、新しい時代のAIで実現してください。

開発者向け

- Google AI Studio
- Google Antigravity
- Vertex AI
- Gemini API / CLI
- Android Studio

すべてのユーザー

- Gemini App (Default Model)
- Google Search (AI Mode)

エンタープライズ向け

- Gemini Enterprise
- Vertex AI