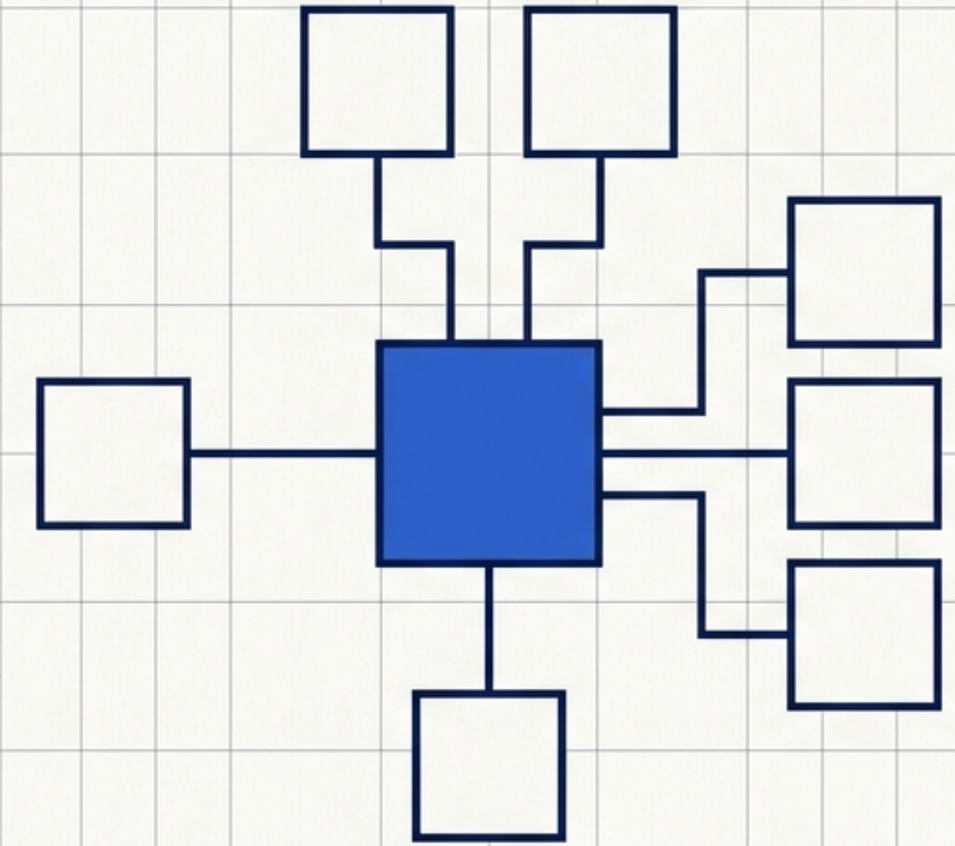
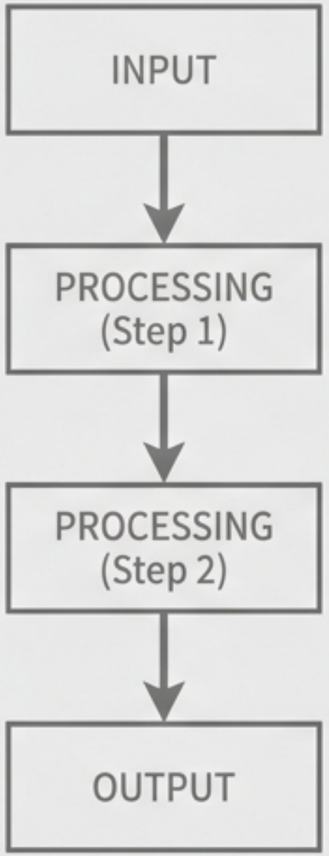
Google I/O 2026 Strategic Analysis Report



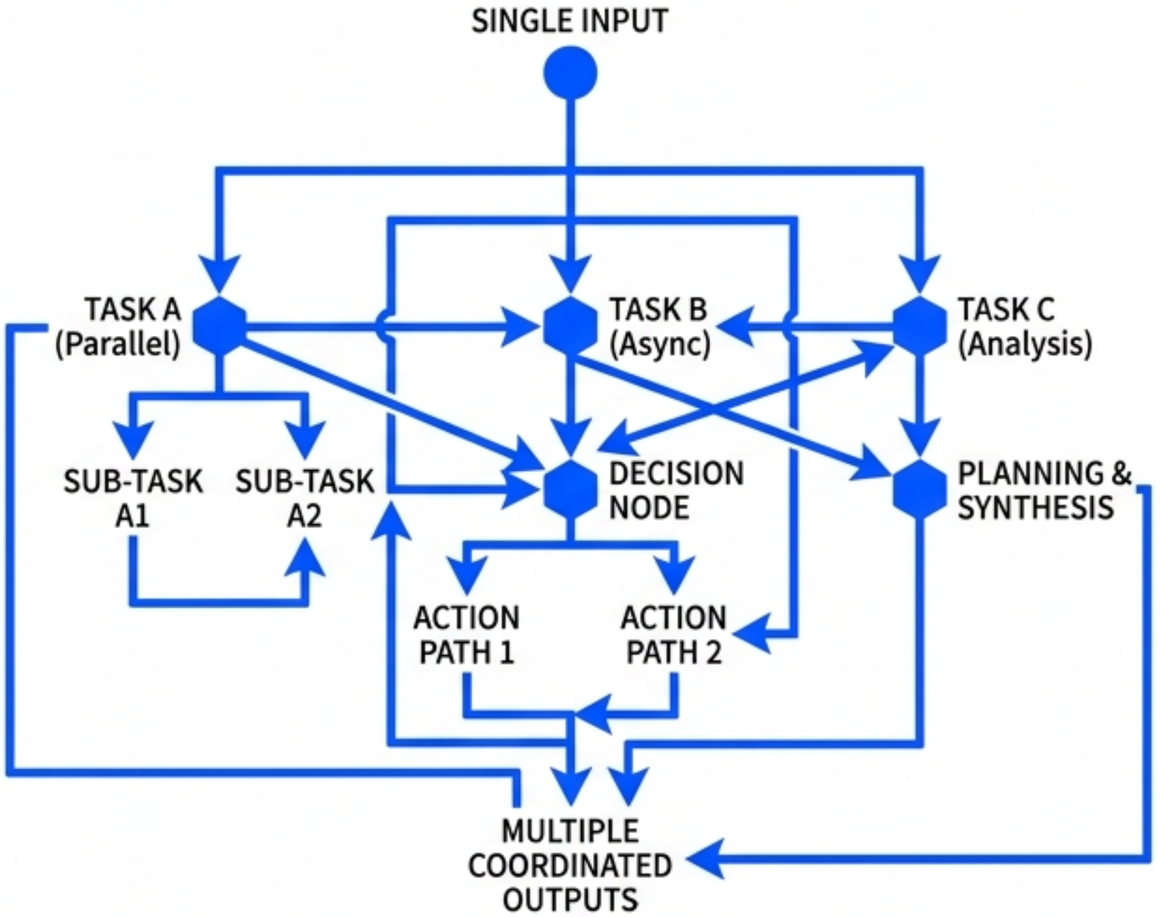
自律システムの統合司令室

Gemini 3.5 FlashとAntigravity 2.0がもたらす「エージェンティックAI」の歴史的転換とROIの現実

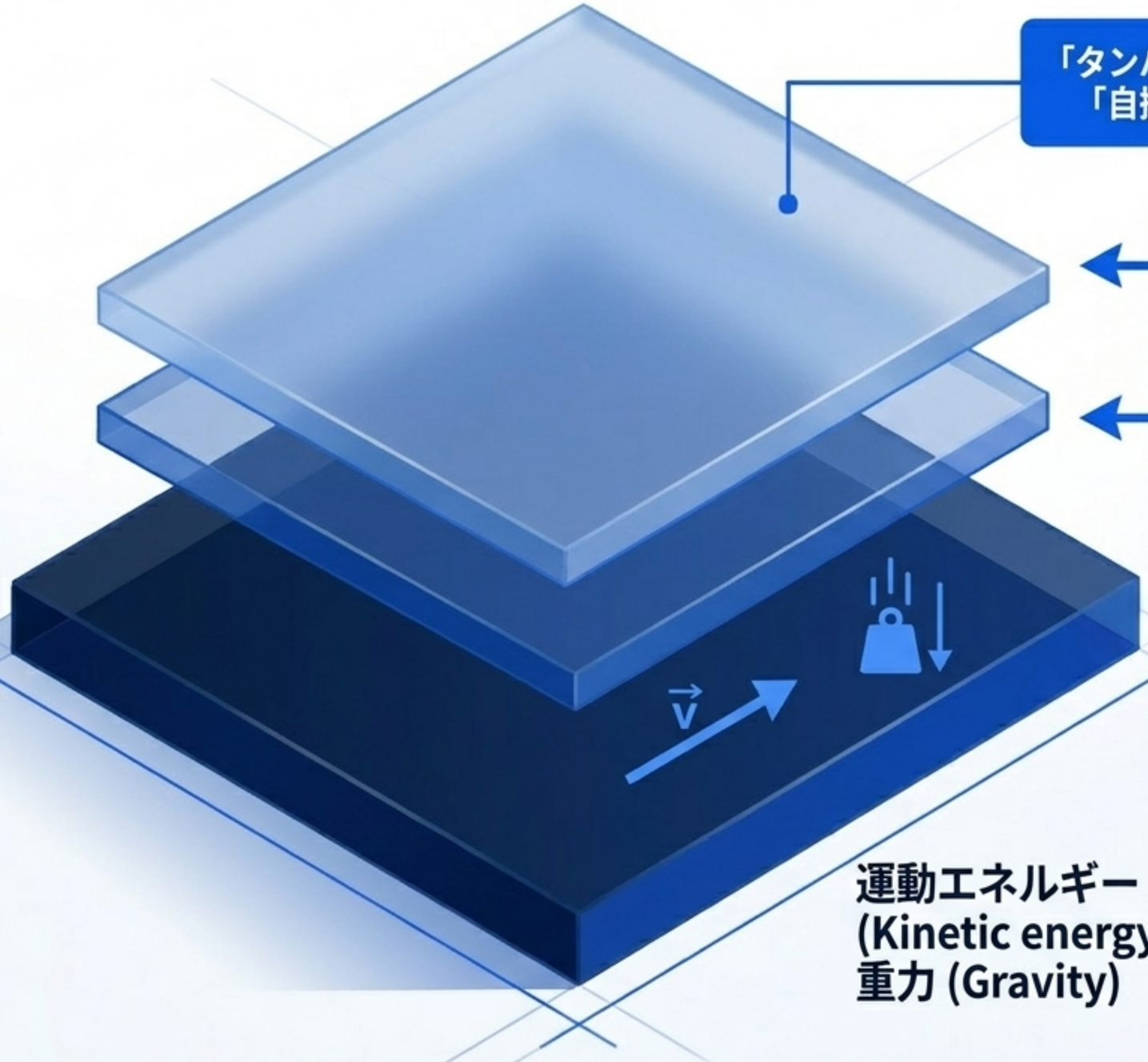
受動的な対話 (Passive Interaction)



自律的行動プロセス (Agentic Workflow)



Google I/O 2026の真のメッセージは「知能の向上」ではなく、
「プロンプト応答型」から「非同期・並列処理の自律エージェント」への完全な移行である。



「タンパク質の折りたたみ構造（クレイアニメ）」
「自撮り動画からブラックホールへの変換」

生成メディア
(動画・3D・オーディオ)

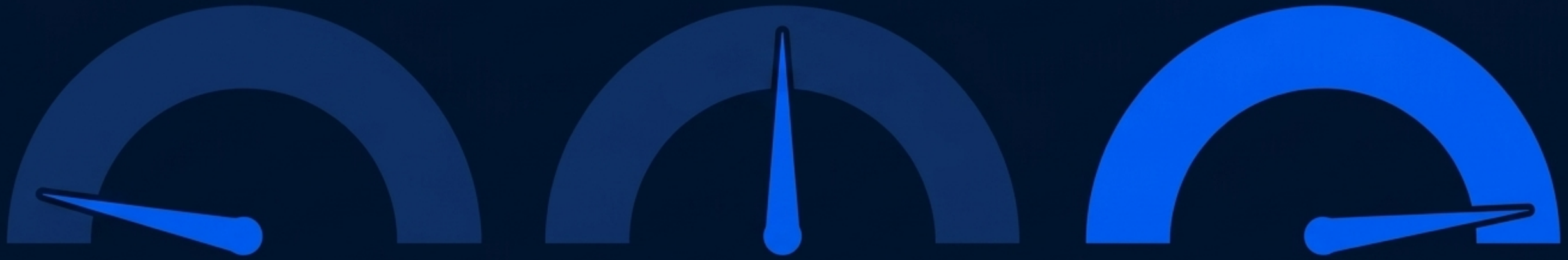
マルチモーダル基盤ネットワーク

物理世界のシミュレーション(World Model)

運動エネルギー
(Kinetic energy) /
重力 (Gravity)

単なるテキストと画像の
結びつけではない。
物理法則を根底から理解し、
世界をシミュレートする
AGIへのマイルストーン。

処理速度のブレイクスルー



競合フロンティアモデル

Gemini 3.1 Pro:
約135トークン/秒

Gemini 3.5 Flash:
277〜289トークン/秒

コンテキストウィンドウ:
104万8,576トークン（長大なログやコードを丸ごと処理）

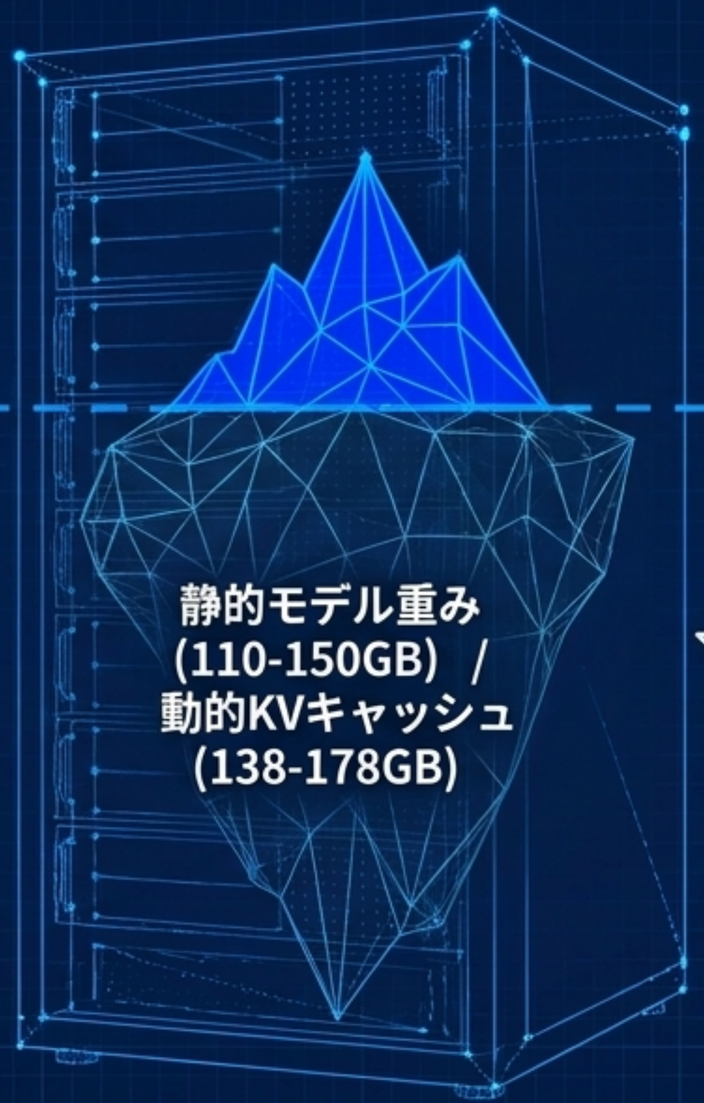
処理速度:
競合フロンティアモデルの4倍、前世代Proの2倍以上

プレビューをスキップし、即日一般公開（デフォルトモデルへ）

複雑なエージェント・タスクにおいて「軽量＝低知能」の常識を破壊。
速度こそが複数エージェント並列処理の絶対条件となる。

Mixture of Experts (MoE) アーキテクチャ

アクティブ・パラメータ数:
わずか100億〜160億
(10B-16B)



アクティブ・パラメータ数:
わずか100億〜160億
(10B-16B)

総パラメータ数:
2,500億〜4,000億
(250B-400B)

Mixture of Experts (MoE) の極致。入力に応じて特定のエキスパートのみを駆動させることで、フラグシップ級の知識量を維持しながら極限の推論スピードを実現。

API制御思想の根本的移行

制御要素	Before（確率的生成）	After（確定的ロジック）
サンプリング制御	temperature, top_pで確率を操作	非推奨。予期せぬ性能低下を招く。
推論の深さ	thinking_budget（数値指定）	Dynamic Thinking (Enumによる明示的段階指定)。
ツール統合	柔軟な関数呼び出し	厳格なIDマッチング必須。 思考漏れ(Thought leakage)の防止。
文脈の維持	毎ターン再送信	推論コンテキストの自動保存 (Thought Preservation)。

AIを「確率的なテキストジェネレーター」として扱う時代は終了。確定的なルールと推論深度で制御する「ロジックエンジン」へと進化した。

Dynamic Thinking: 思考深度の明示的制御

- MINIMAL: 推論無効。リアルタイム性・スループット最優先。
- LOW: 短レイテンシ＋一定推論。単純なコーディング、データ分析向け。
- MEDIUM (Default): バランス型。一般的なエージェント開発、複雑なコーディングの最適解。
- HIGH: 論理的推論の最大化。数学的解決、長期計画、高度なアルゴリズム設計。

なぜこれが重要か? → タスクの難易度に応じて「速度」と「AIの思考時間（＝コスト）」を開発者が直接コントロールできるようになったため。

Antigravity 2.0: ソフトウェア工場の完全自動化

93

並列稼働したAIサブエージェント数

12 Hours

OSコアの構築に要した時間

2.6B

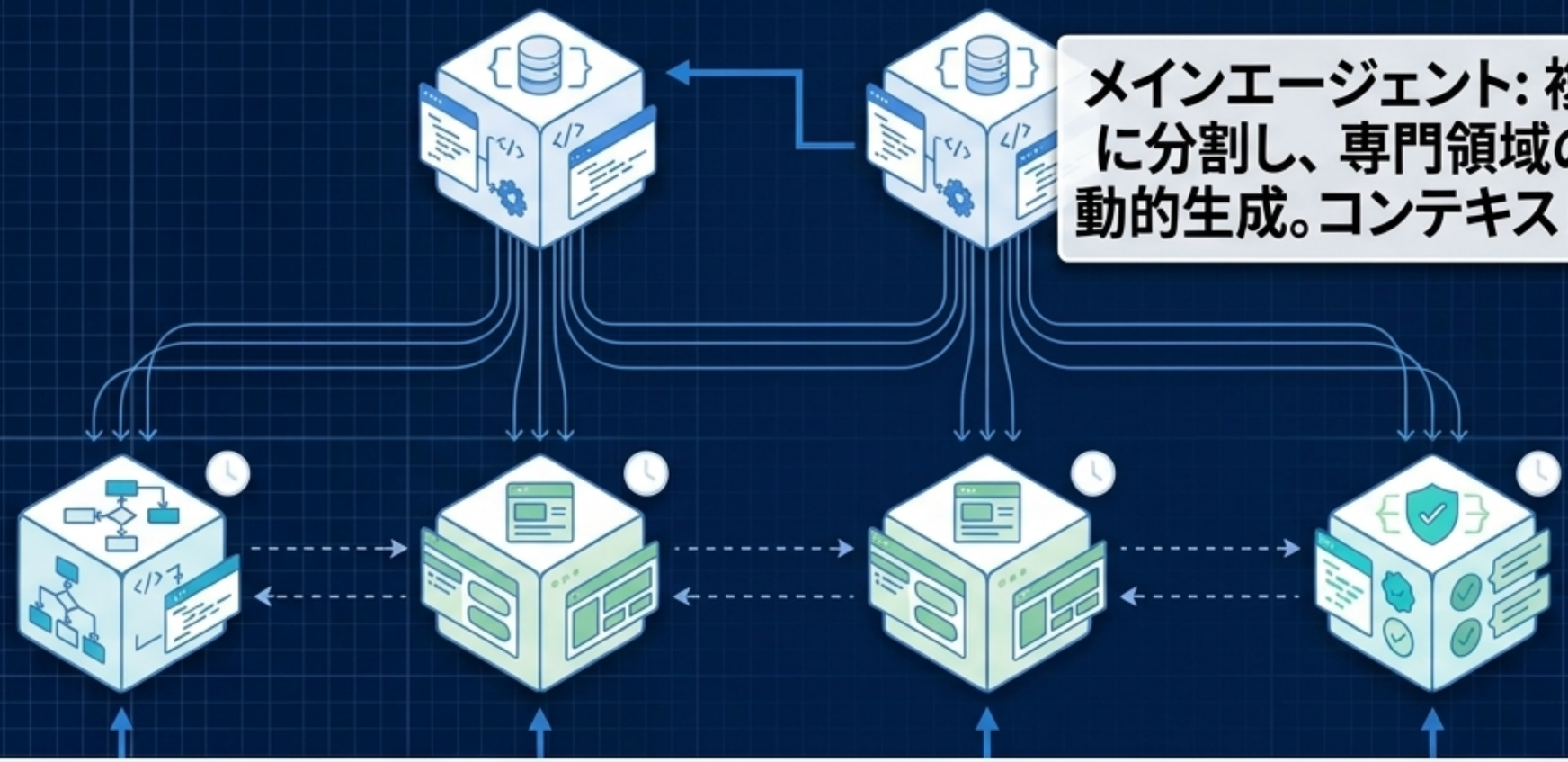
エージェント間で通信されたトークン数

<$1,000

全体のAI処理コスト

単なる「コードエディタの拡張」から、AIを指揮・統括する「独立した中央司令室」への進化。数週間かかる基盤開発を半日に短縮し、システムインテグレーションの経済的前提を破壊する概念実証（PoC）。

動的サブエージェントの非同期並列処理



メインエージェント: 複雑な課題を自律的に分割し、専門領域のサブエージェントを動的生成。コンテキストの汚染を防止。

ロジック / UI / テスト: メインプロセスをブロックせずバックグラウンドで非同期実行。

JSONフックによる挙動傍受、Live Voice Transcriptionによる音声指示、
スケジュール実行（Cron）による完全自律テストルーチン。

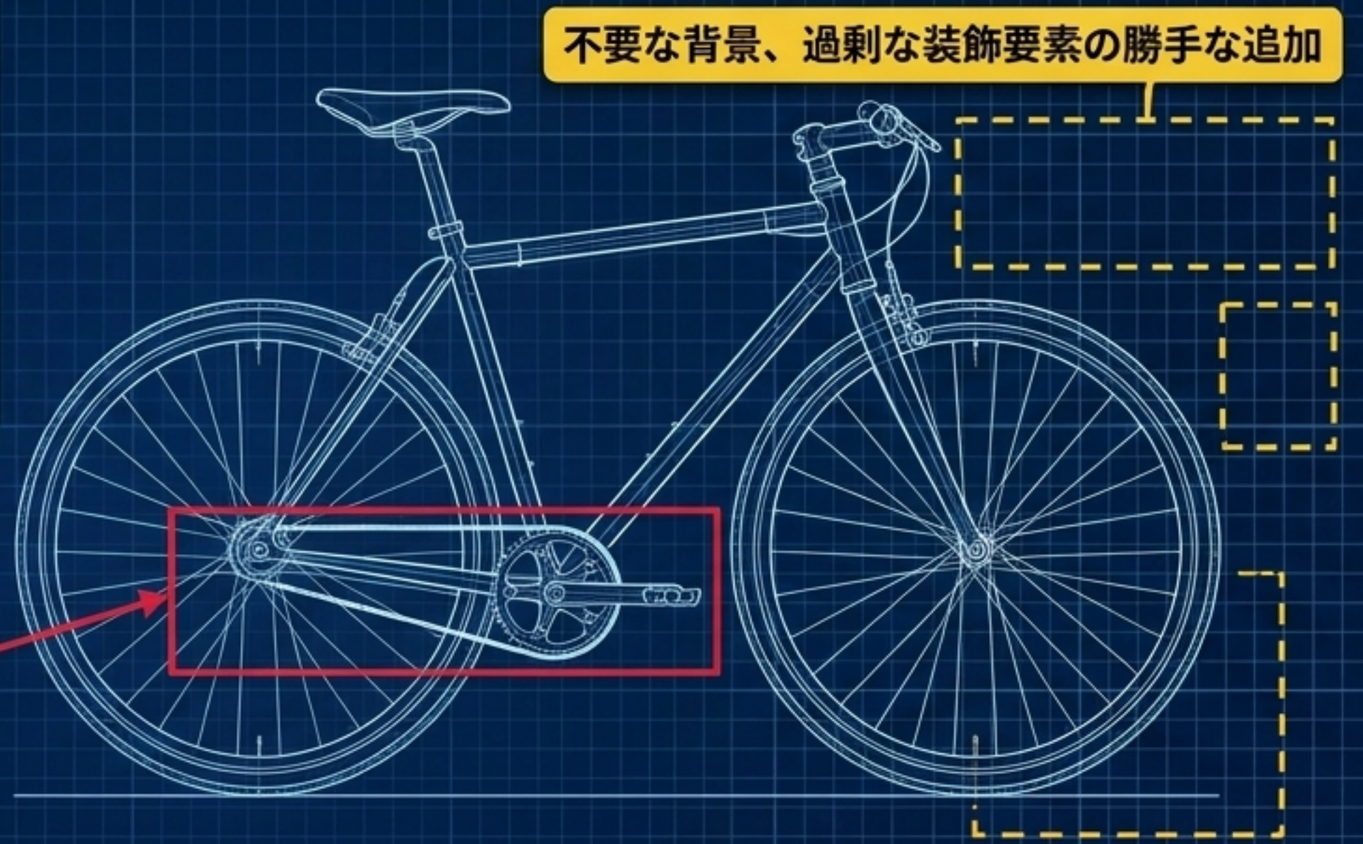
The Reality Check: 破綻するコスト構造とROIの危機

「Flash＝安価」というブランドイメージの崩壊。
超軽量版(Lite)の6倍、最上位モデル(Pro)に肉薄する価格設定へ。

⚠ **単価の3倍化は、悲劇の序章に過ぎない。**

「過剰装飾（Superficial Fluff）」と構造的デバッグの欠陥

Artificial Analysis評価
Artificial Analysis評価 — 平均出力トークン
3,600万に対し、Gemini 3.5 Flashは
7,300万トークン（2倍以上の肥大化）
を出力。



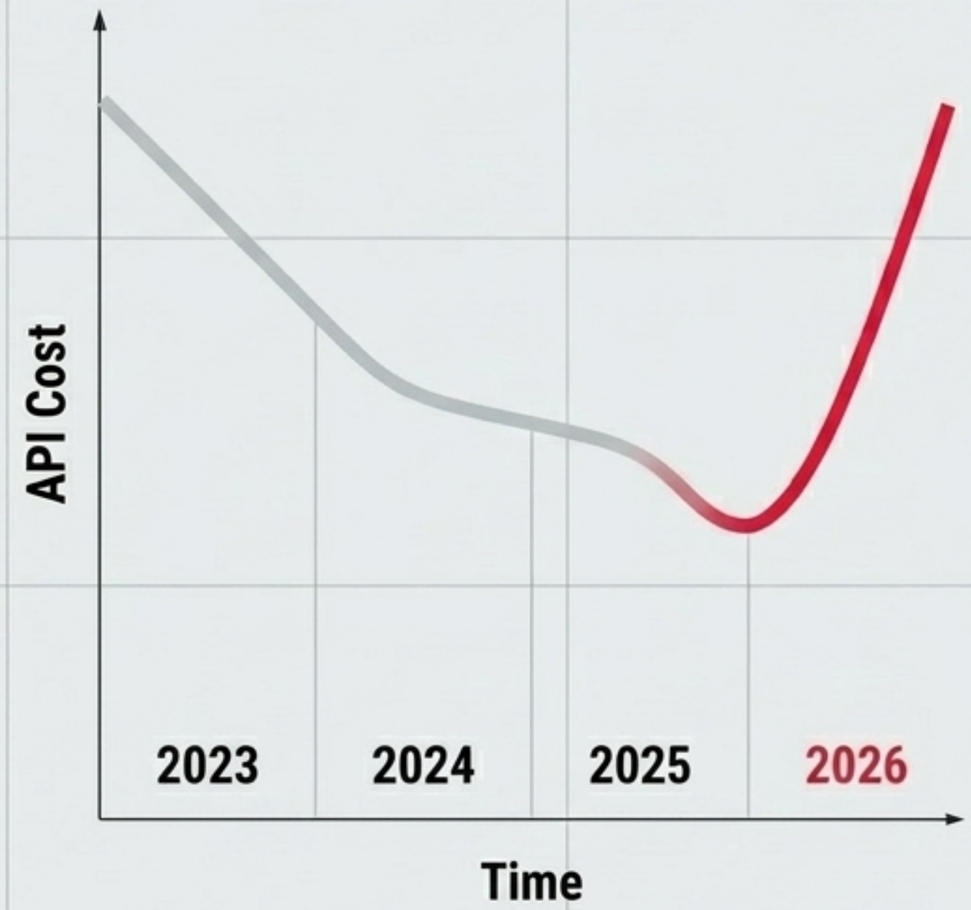
AIはピンポイントで構造エラーを修正する代わりに、無用な装飾を加えてコードを「肥大化」させる傾向が強い。この無駄な1出力に約13セントが浪費される。

自律システムにおけるコスト爆発の構造

[単価の3倍化] ✕ [出力トークン数の2倍化（肥大化）] ＝ 実行コストの実質9倍化

自律エージェント環境下でモデルが「過剰思考ループ」に陥った場合、
トークン消費は雪だるま式に膨れ上がる。
これはクラウドリソースの予期せぬ巨額請求（Bill Shock）に直結する。

業界の潮流変化：コモディティ化の終焉と「収益化フェーズ」

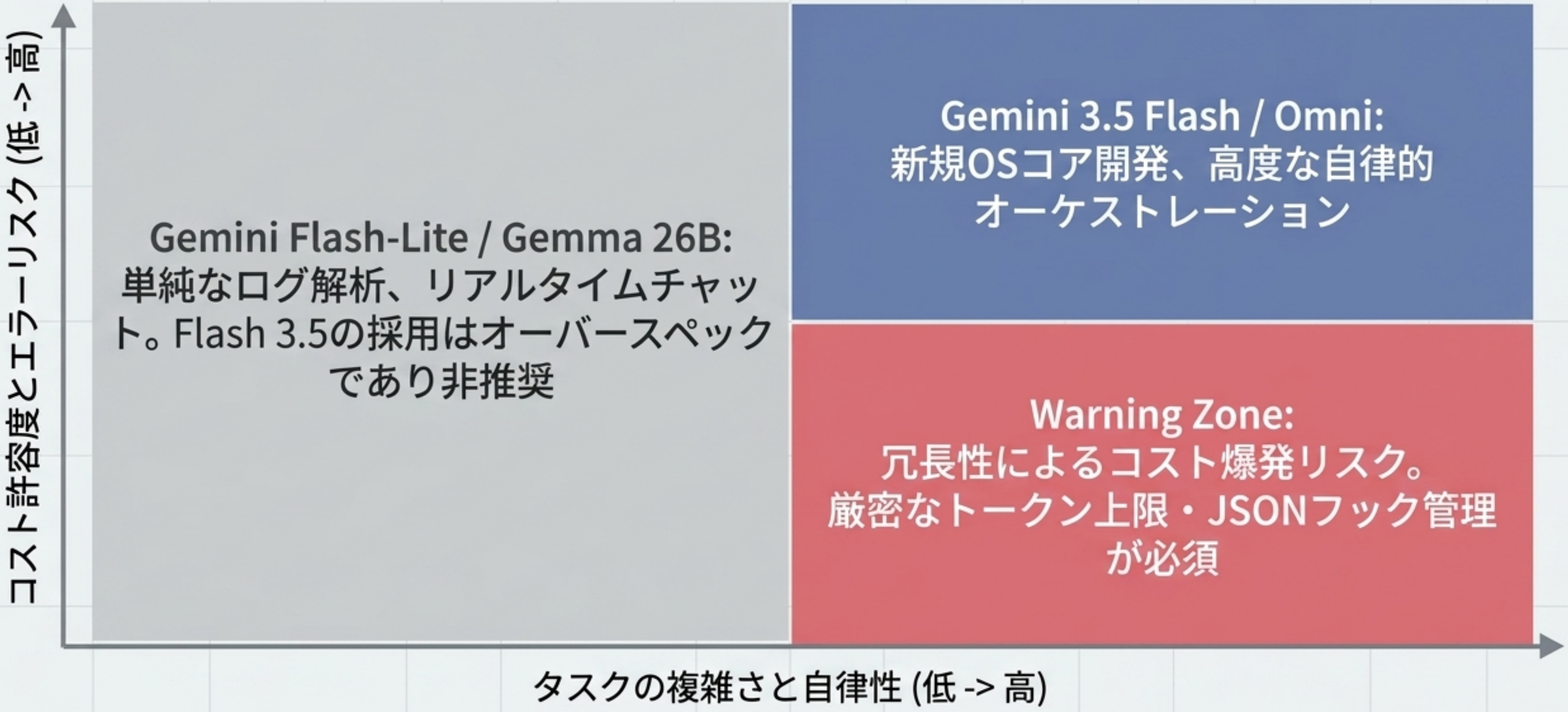


The End of the Race to the Bottom:
大手AIラボによる熾烈な低価格競争・市場シェア獲得フェーズは終了した。

Testing Price Tolerance:
OpenAI (GPT-5.5) や Anthropic (Claude Opus 4.7) と同様、Googleもエンタープライズ顧客の「価格許容度」を探り始めている。

The New Reality:
エージェンティックな高度推論能力は、実質的な値上げによる「対価」を伴うプレミアム・リソースへと移行した。

自律AIエージェント導入の診断マトリクス



タスクごとにモデルを使い分けるハイブリッドなルーティング戦略が、今後のアーキテクチャの必須要件となる。

AIを「API」から「仮想従業員 (Virtual Employee)」として管理する時代へ

The Breakthrough:
知能と速度のトレードオフは解消された。Antigravity 2.0によるソフトウェア工場の完全自動化は既に現実である。

The Challenge:
自律性の向上は、暴走時のコストリスクと直結する。「思考の効率性」の管理が最大の経営課題となる。

The Mandate:
企業はAIを単なる便利なツールとして導入するフェーズ終え、膨大なコストを消費する「従業員」として、そのROIとパフォーマンスを厳密にモニタリングする高度な運用戦略（コントロールルーム）を構築しなければない。来たる「Gemini 3.5 Pro」に備え、今すぐ持続可能なエージェント基盤を確立せよ。