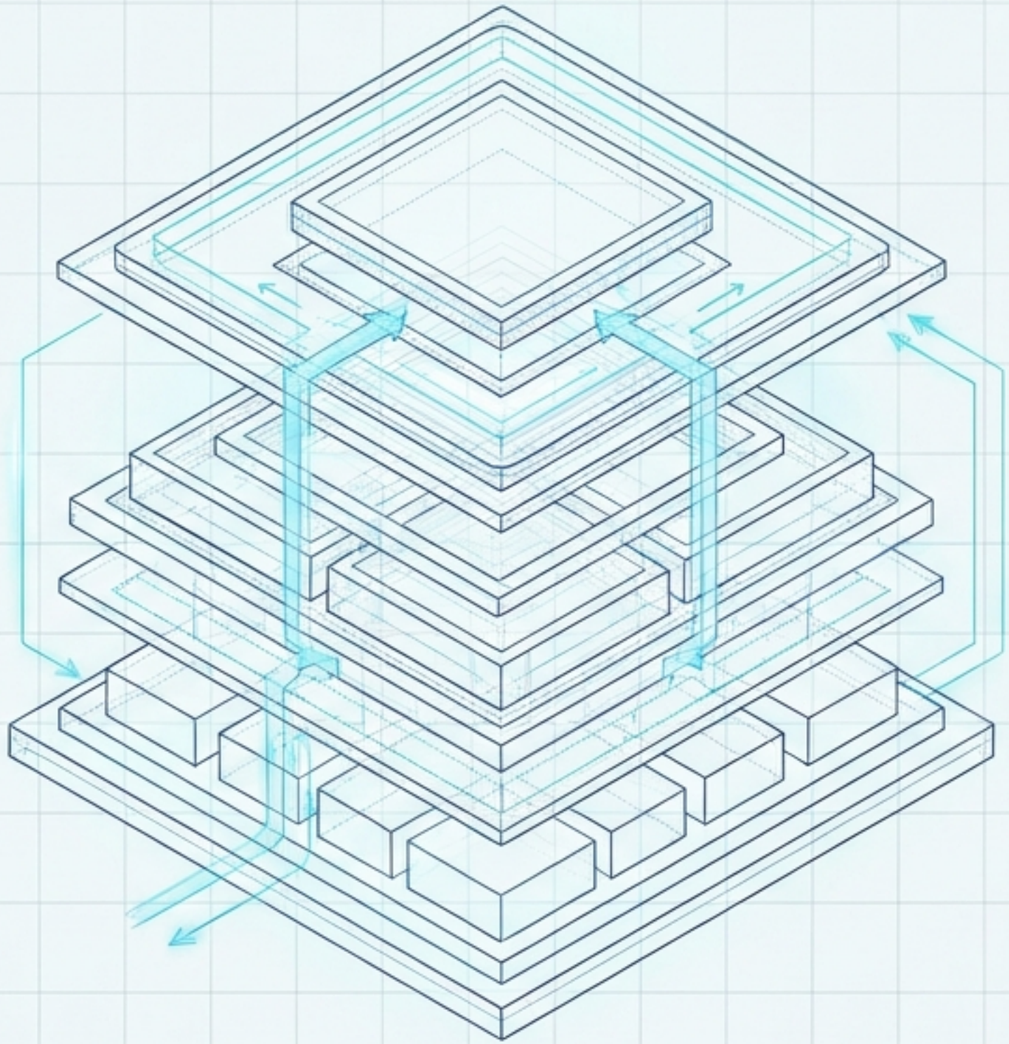
Qwen4-Max-Preview 徹底調査：次世代モデルの解剖と導入戦略

15日間限定公開・アーキテクチャ・経済性・実測評価の全貌



DOCUMENT ID: QWEN4-PREVIEW-EVAL

TARGET: Enterprise IT & AI Engineering

STATUS: Restricted Preview (350 Hours Only)

Gemini Notebook

ミッションの俯瞰と制約事項



制約 1: 公式ベンチマーク非公開

GPQAやSWE-bench等のスコア開示なし。サードパーティの実測が唯一の指針。

制約 2: ウェイト（重み）固定

期間中モデルの更新なし。終了後にアクセス不可。

制約 3: 公開ストレステスト

正式版（GA）に向けた、開発者の実務ワークフロートレース収集が主目的。

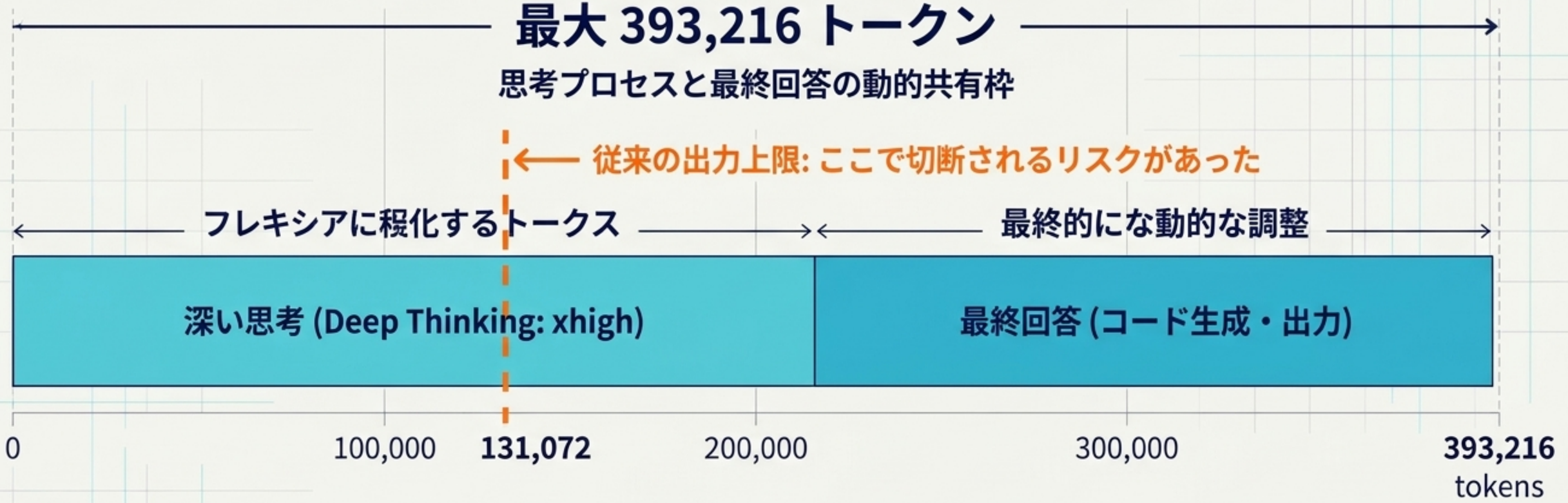
Gemini Notebook

仕様の進化：世代間比較マトリクス

Qwen4-Max-Preview (プレビュー版)	Qwen3.8-Flash-Next (先行実験版)	Qwen3.8-Max (現行正式版)
入力: 1,000,000 トークン	入力: 1,000,000 トークン	入力: 1,000,000 トークン
出力: 393,216 トークン（動的共有枠）	出力: 思考 262,144 + 回答 131,072	出力: 131,072 トークン
思考モード: xhigh（既定でハイブリッド）	思考モード: xhigh（既定）	思考モード: API版は Hybrid
API: Token Plan限定	API: セルフホスト対応	API: 従量課金対応

Gemini Notebook

出力革命：共有バジェットの視覚化



複数ファイルにまたがるコード改修や、数千行のフロントエンド実装を『単一パス』で完遂するための専用設計。

Gemini Notebook

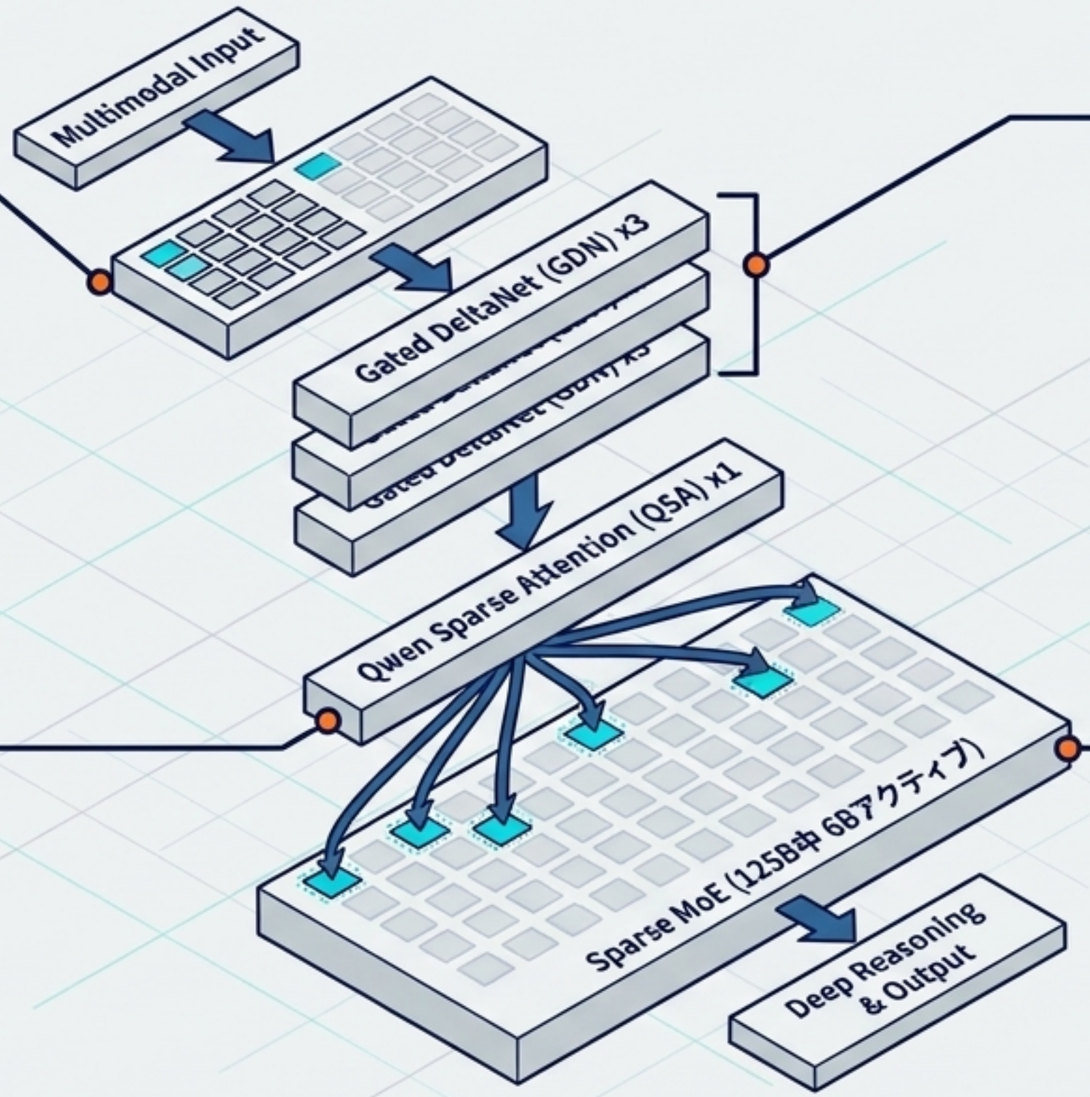
Qwen4 アーキテクチャの解剖図

N-gram Memory Lookup (51B)

演算FLOPsゼロ。
非同期ルックアップによる知識容量の拡張。

Gated DeltaNet (GDN) x3

線形アテンション。
KVキャッシュの肥大化を防ぎ、文脈を線形圧縮。



Qwen Sparse Attention (QSA) x1

MQAインデクサ。
上位ブロックのみを選択的に検索し、従来比8.6倍のスループットを実現。

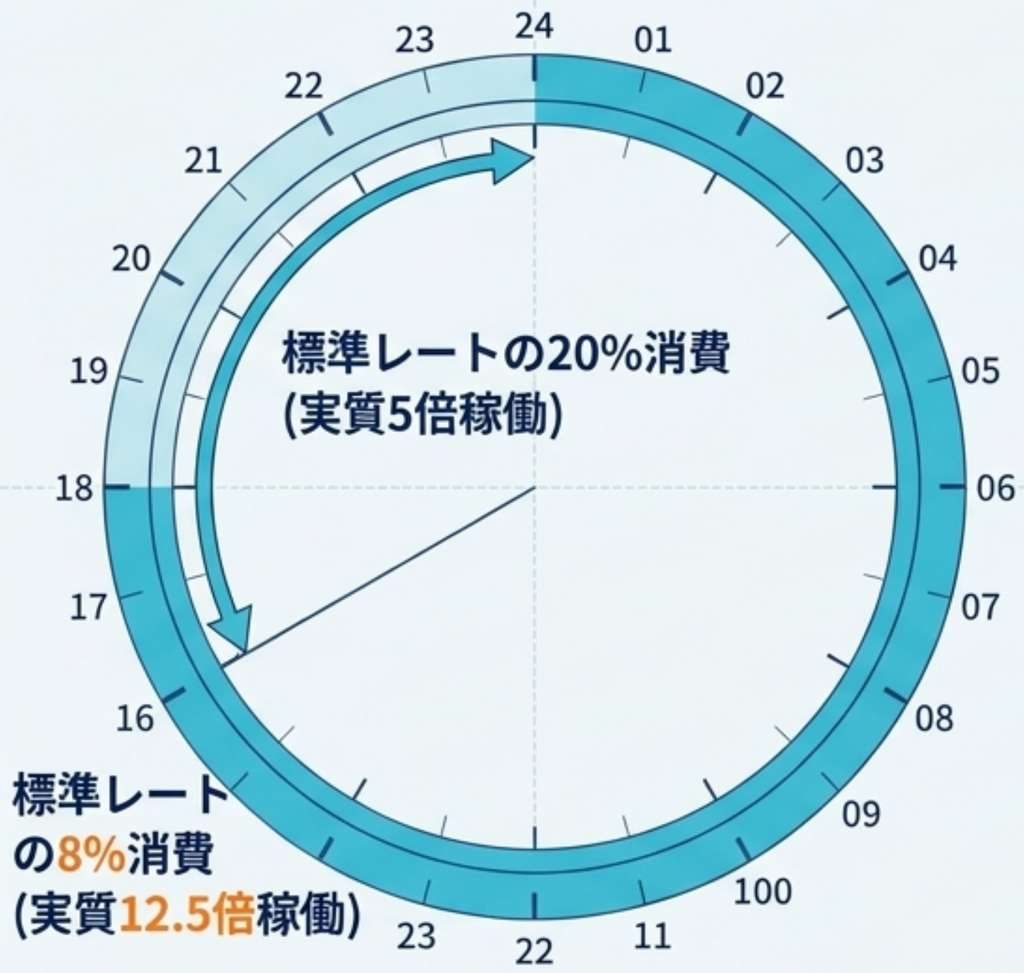
Sparse MoE (125B中 6Bアクティブ)

驚異の高希薄設計。
アクティブ率わずか約4.8%で推論効率を極限まで高める。

※先行公開版 qwen4_exp (Qwen3.8-Flash-Next) の技術仕様に基づく推測

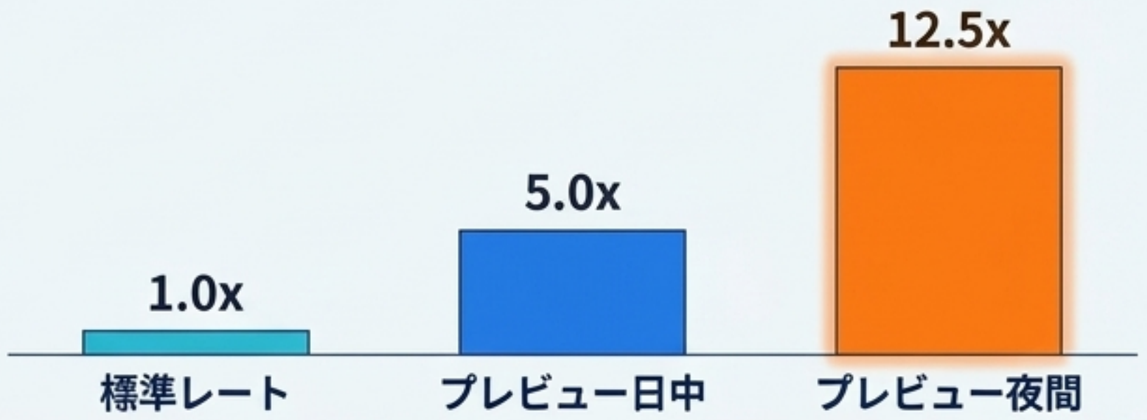
Gemini Notebook

経済性の真実：Token Planと夜間優位性



圧倒的な夜間優位性

アップグレードの算段：10,000クレジットあたりの実質単価
Lite / Essential
初月割引からスタート推奨
⚠ **追加パックの罠**
$7.50 / 2万cr と割高
Proプランへの日割りアップグレード
枠不足時は上位プランへの移行が最も経済的



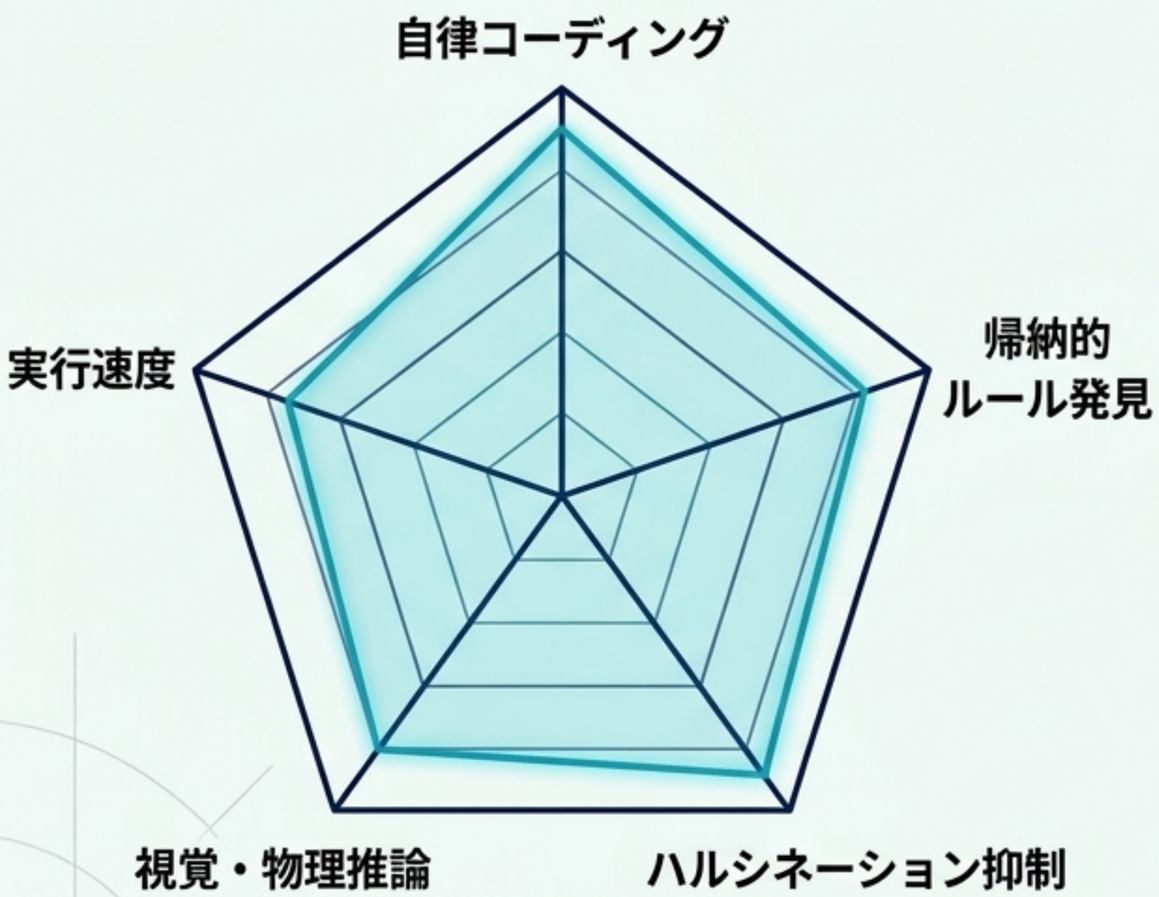
Gemini Notebook

法人導入の境界線：セキュリティ・ゲートウェイ



Gemini Notebook

戦場での実力：実測パフォーマンス



Fact 1: 自律コーディング

既存Pythonコードのエラーを自律検知し、57件の単体テストを生成・通過させてリファクタリング完遂。

Fact 2: ルールの帰納的発見

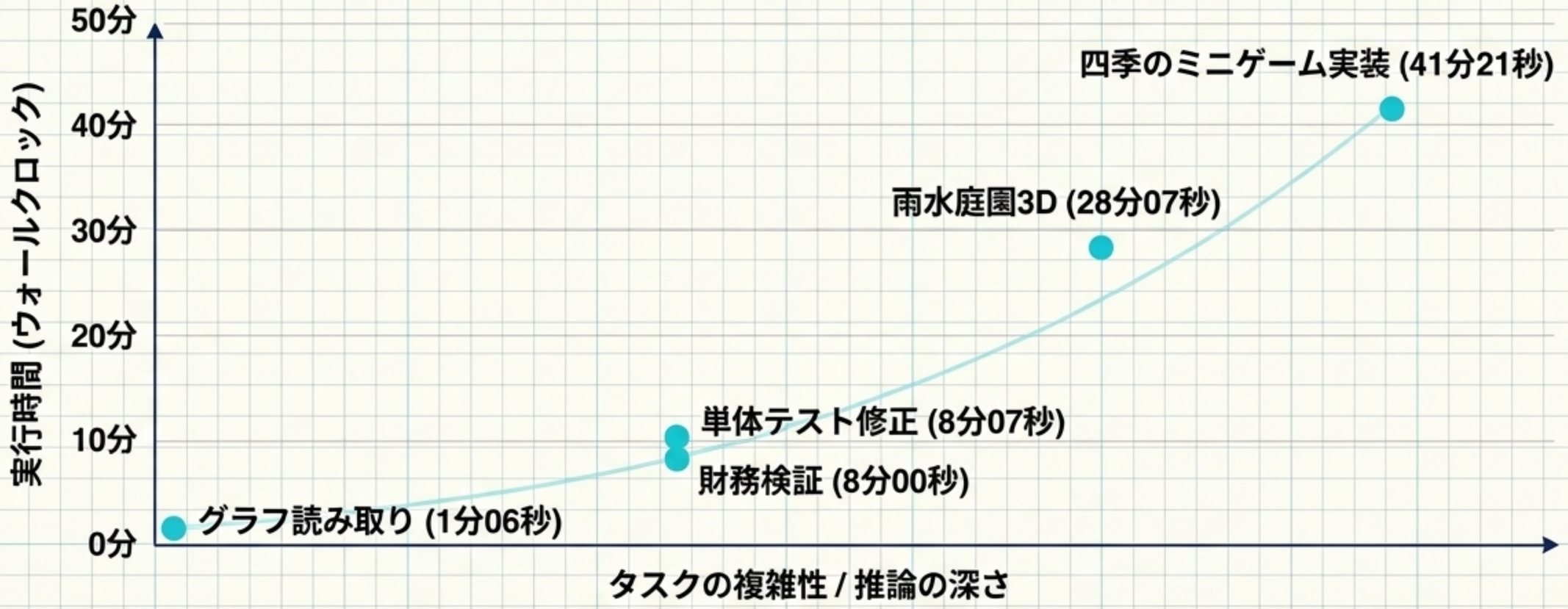
未明文化された価格ルールを数百の請求データから1セント単位で特定し、他モデルを凌駕。

Fact 3: ハルシネーション抑制

Apple等3社の財務横断検証で、時価総額の引っかけ質問や文脈混同トラップを完全回避。

Gemini Notebook

The Cognitive Trade-off: 推論力 vs 実行速度



⚠ 高品質な出力を得るには『待つ』設計が必要。サンドボックス検証では、DeepSeek比で2倍の思考時間を要し、時間切れ（タイムアウト）が4倍発生。

Gemini Notebook

マルチモーダル推論：視覚・物理・法務の統合

事例1: 法務推論

視覚入力（画像） → 視覚検知: シートベルトが脇の下を通るズレ → 外部知識: 9カ国の法律比較 → 結論: オーストラリアにおける罰金リスク（運転手対象）を特定

事例2: 物理推論

視覚入力（衛星画像） → 自律コード実行: ピクセル距離から縮尺（約1km/px）算出 → 外部知識: 断熱減率(上昇5.5°C低下 / 下降9.8°C上昇）の適用 → 結論: 乾燥側の約4°C昇温を導出

Gemini Notebook

The Developer's Minefield: 5つの落とし穴



1. max_tokens 切断バグ

症状: 思考と回答の合算値になったため、従来通り小さく設定すると回答前に切断される。

処方箋: 未指定または39万を設定。



2. マルチターン分離必須

症状: reasoning_contentを独立フィールドで返送しないと処理が失敗する。

処方箋: 対応SDK/クライアントを使用。



3. パラメータ排他制約

症状: 思考強度と特定関数強制(tool_choice)は併用不可。

処方箋: 関数強制時は思考を false に。



4. 検査エラーの無限ループ

症状: HTTP 400 発生時に再送ループに陥るツールあり。

処方箋: 400エラー時の即時停止実装。



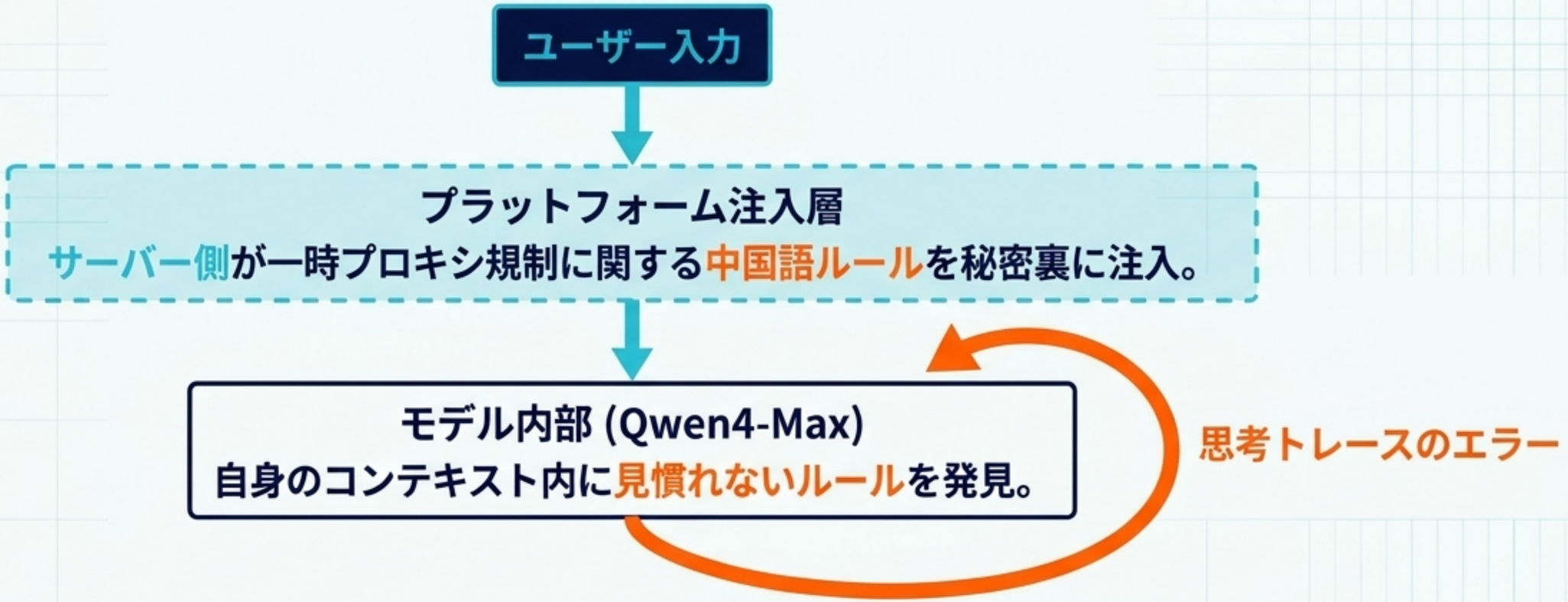
5. 動的レート制限 (429)

症状: 深い推論による応答遅延。

処方箋: stream: true を使用し、タイムアウトを180秒以上に設定。

Gemini Notebook

システムプロンプト漏洩とアーキテクチャの癖



モデル自身が「プロンプトインジェクションの疑いあり」と誤認し、自己防衛するループが発生。

インサイト:
クライアント側の不具合ではなく、サーバー側のルール注入による仕様。
思考プロセスの透明化がもたらした副産物。

Gemini Notebook

最適モデル選定の意思決定マトリクス

R&D / プレビュー検証

処理速度・低レイテンシ

深い推論・自律性

左上: 速度重視・プレビュー

DeepSeek-V4.1-Flash

右上: 推論重視・プレビュー

Qwen4-Max-Preview
テスト駆動リファクタリング、
財務突合に最適

左下: 速度重視・本番

Qwen3.8-Max /
Qwen3.8-Flash-Next

右下: 推論重視・本番

Claude Opus 5.5 /
Gemini 4 Carbon

プロダクション本番稼働

Gemini Notebook

Action Plan for GA: 正式版に向けた3つの準備

1. Test Now (今すぐ検証)

日本時間23時以降の「8%消費枠」を活用し、自社で最も難易度の高いタスク（既存コードのテスト駆動リファクタリング等）を投入して次世代の到達点を測る。

2. Avoid Production (本番環境へのハードコード禁止)

qwen4-max-previewは退役し、利用規約上もAPIバックエンド化は禁止。プロダクション依存を避ける。

3. Update Client Architecture（クライアント構造の準備）

正式版移行に備え、39万トークンのバジェット管理と、reasoning_contentの分離保持（マルチターン対応）ができるクライアント実装を整える。

The true test begins after General Availability.

Gemini Notebook