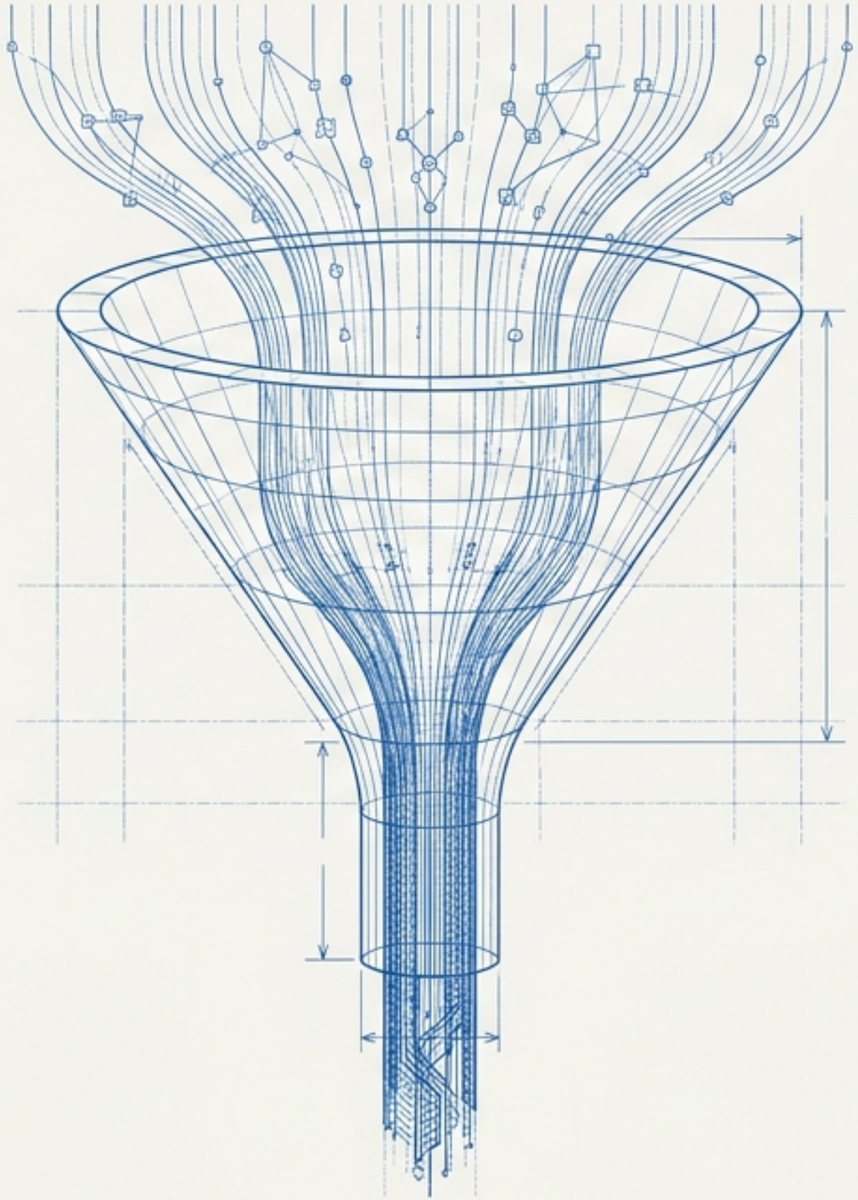
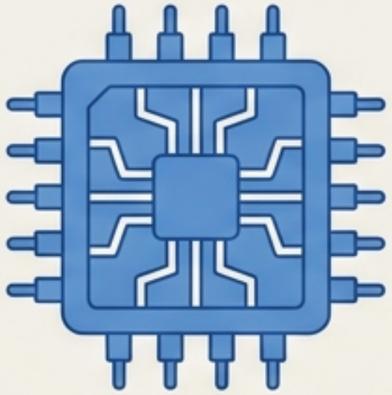
DeepSeek-V4-Pro-0813の全貌

アーキテクチャ刷新・エージェント性能・コスト構造の包括的分析

リリース: 2026年8月13日（General Availability）
対象読者: AI/IT戦略担当者、CTO、AIエージェント開発者





1．極限の計算効率化

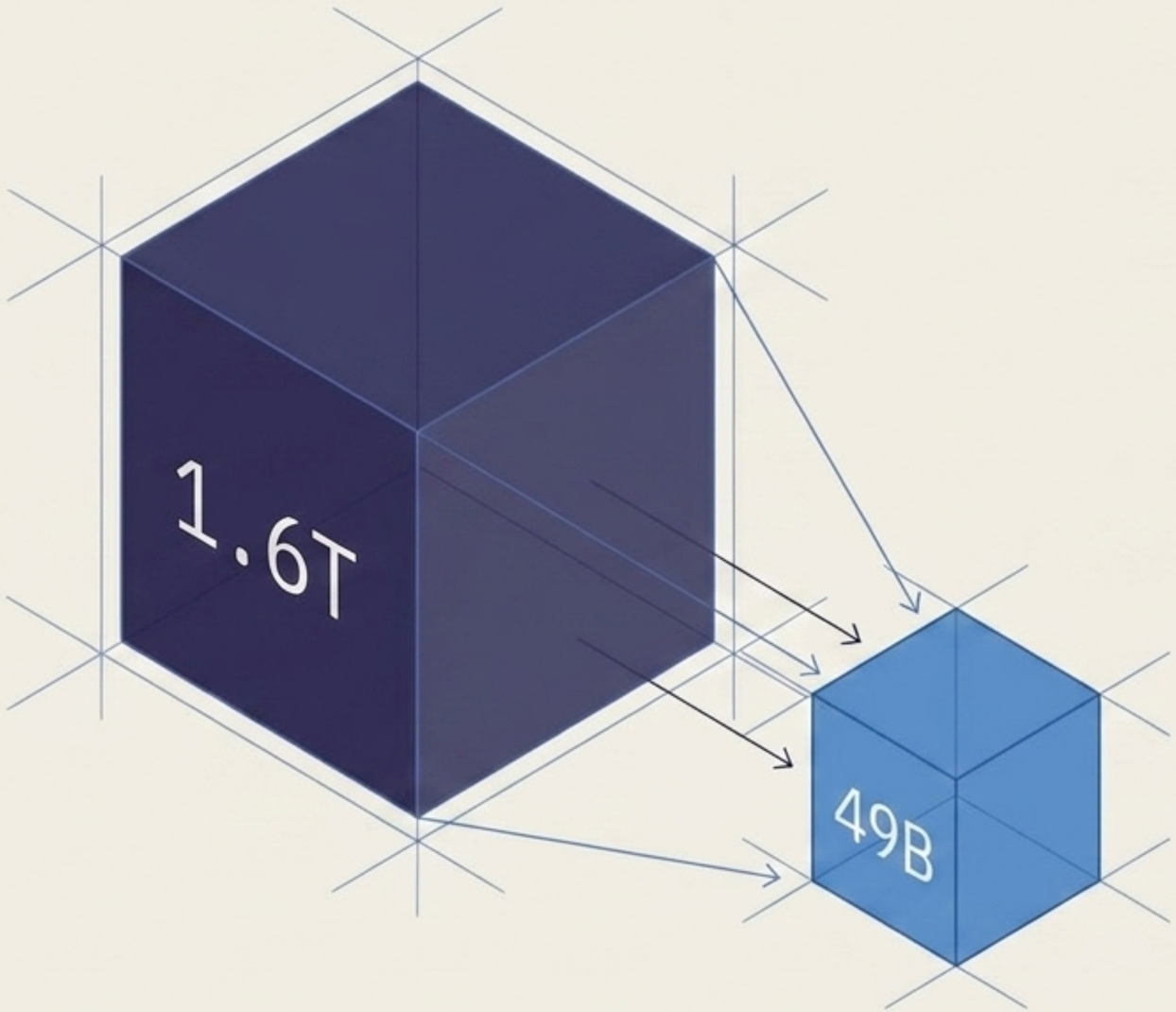
1.6兆パラメータから490億の有効パラメータへの絞り込み（MoE）。ハイブリッド・　アテンションによるコンテキスト処理の極限圧縮。

2．自律タスク性能の飛躍

継続的なポス・トレーニングによる進化。Claude Fable 5に肉薄するターミナル操作とコーディング能力の実現。

3．価格構造の破壊

既存フロンティアモデルの「1/50以下」のコスト。強力なキャッシュ割引とオフピーク料金が、自律型AIエージェントの常時稼働を初めて経済的に可能に。



モデル仕様

総パラメータ数: 1.6兆（1.6T）

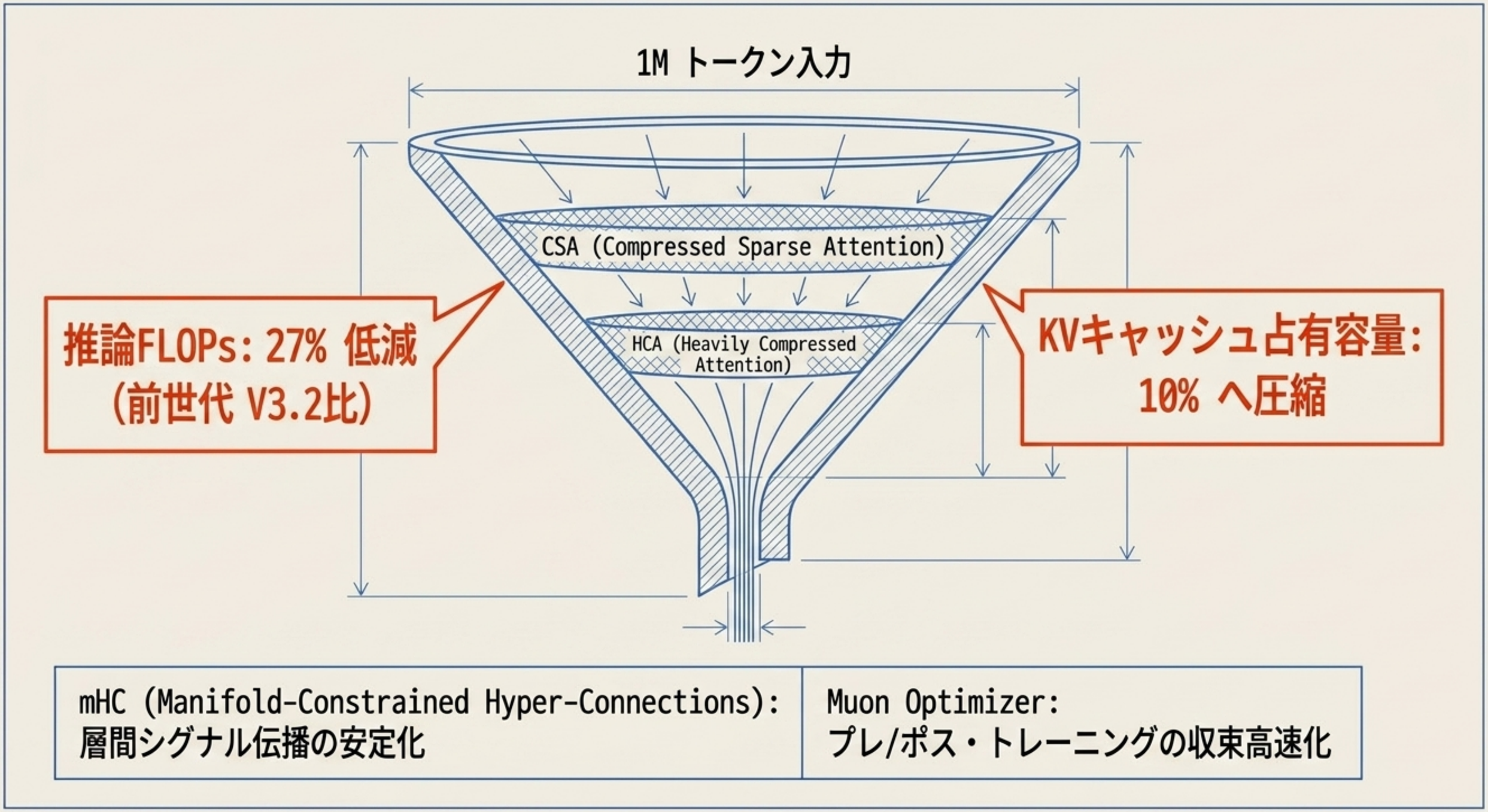
アクティブパラメータ数: 490億（49B）

コンテキスト長: 1,048,576 トークン（1M）

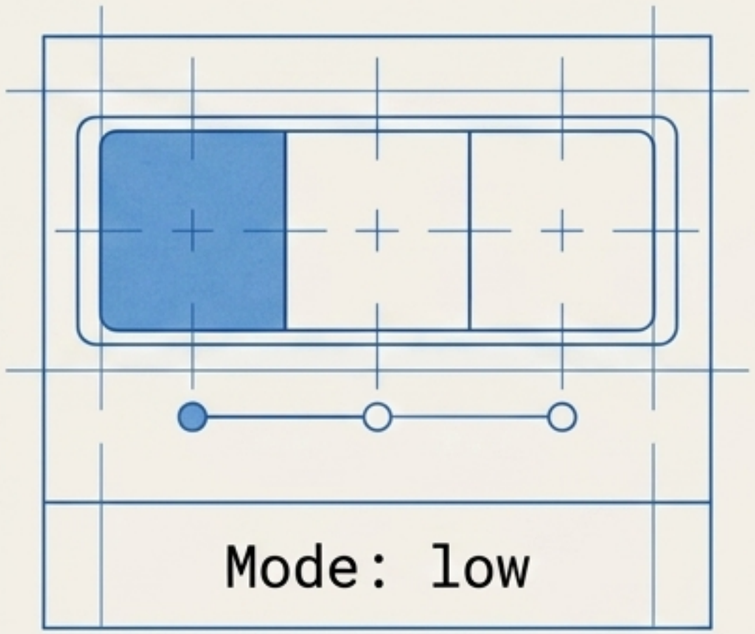
最大出力トークン数: 384,000 トークン

モデル構造: MoE + ハイブリッド・アテンション

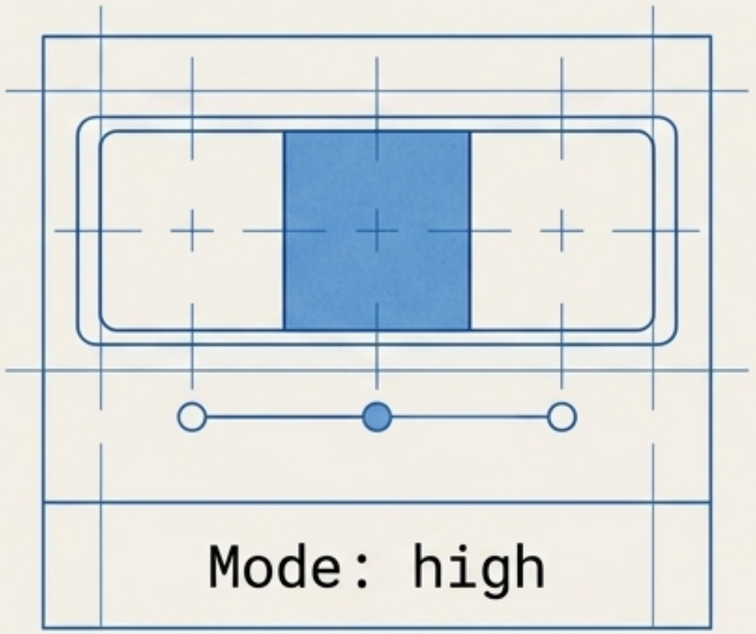
量子化精度: FP4（MoEエキスパート層）+ FP8 混合精度


1M トークン入力
CSA (Compressed Sparse Attention)
HCA (Heavily Compressed Attention)
推論FLOPs: 27% 低減
(前世代 V3.2比)
KVキャッシュ占有容量:
10% へ圧縮
mHC (Manifold-Constrained Hyper-Connections):
層間シグナル伝播の安定化
Muon Optimizer:
プレ/ポス・トレーニングの収束高速化

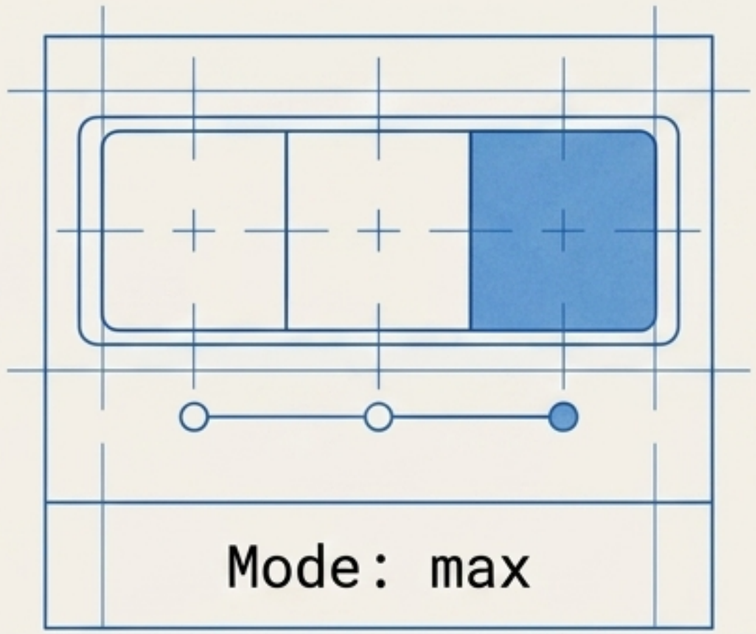
Thinking Effort Control: 動的な推論リソース配分



低遅延優先。定常的な日常会話、単純な指示追従。



デフォルト設定。通常のエージェントタスク、開発処理（出力前に論理プロセスを挟挟む）。

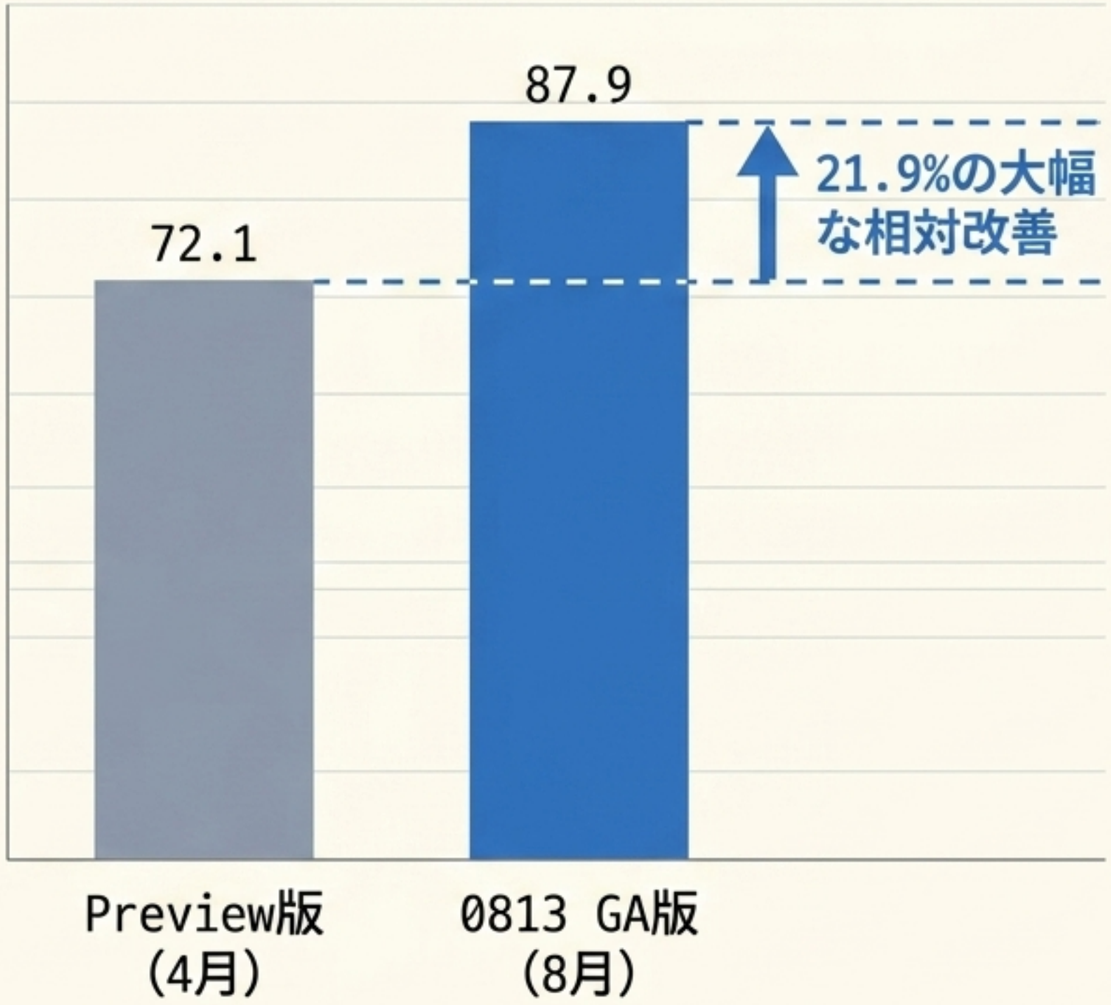


能力の最大化。最高難易度の課題解決、広範な探索を要する研究タスク。

Trade-off Note: 高度な思考モード（high/max）では内部推論処理による初期遅延が発生。TTFT（ファーストトークン到達時間）は25.67秒、生成スピードは約83 tokens/sec（BenchLM測定）。

エージェント性能の飛躍: Terminal-Bench 2.1

ターミナル操作およびシステム開発タスクにおける劇的なスコア上昇



$$\text{相対改善率} = \frac{87.9 - 72.1}{72.1} \times 100 \approx 21.91\%$$

Performance Insight: 実システム操作の精度が劇的に向上。一部で「15.8%向上」と誤認されるが、正しくは「15.8ポイント上昇」による約21.9%の相対改善である。

主要エージェント・ベンチマーク比較

評価対象領域	DS 0813 (GA)	Fable 5	DS Preview(4月)
Terminal-Bench 2.1 (ターミナル操作)	87.9	88.0	72.1
CyberGym (サイバーセキュリティ)	83.3%	-	52.7%
DeepSWE (AIソフトウェアエンジニアリング)	62.7%	-	12.8%
Toolathlon-Verified (複合ツール利用)	74.1%	-	-
NL2Repo (自然言語からのリポジトリ構築)	61.5%	-	-

Important Context: 87.9点は最小ハーネスまたはClinePassのhigh/max推論強度での最高評価値。BenchLMのKnowledge領域　(MMLU-Pro 87.5%等) でも高評価を獲得。

コスト構造の破壊：フロンティアモデルとの単価比較

Claude Fable 5 と DeepSeek V4-Pro 0813 の100万トークン単価比較

Output Token Cost

$$\text{出力単価比率} = \frac{\$50.00}{\$0.87} \approx 57.47\text{倍}$$

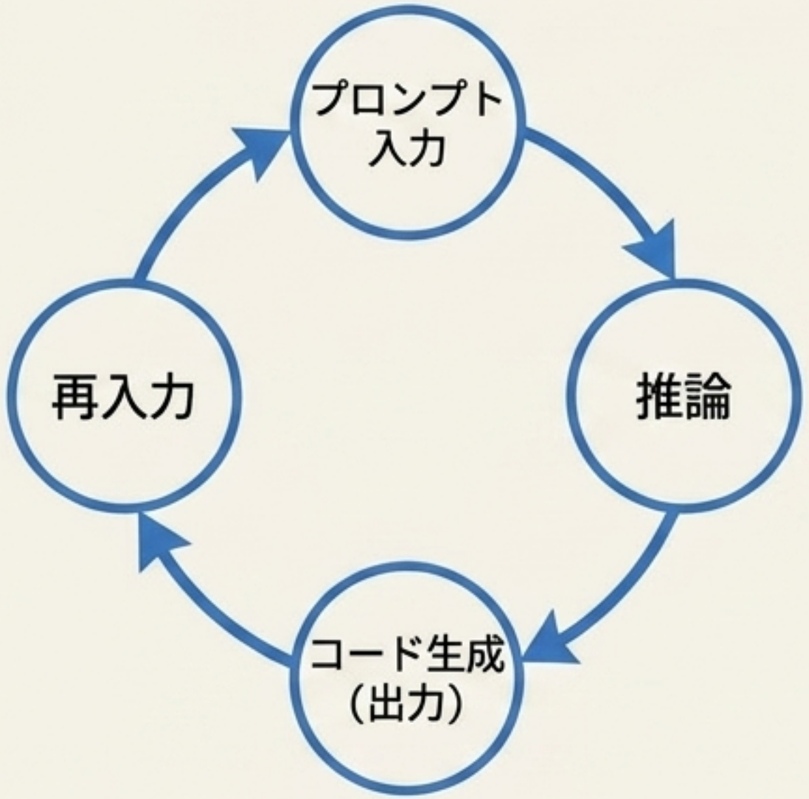
Input Token Cost (Cache Miss)

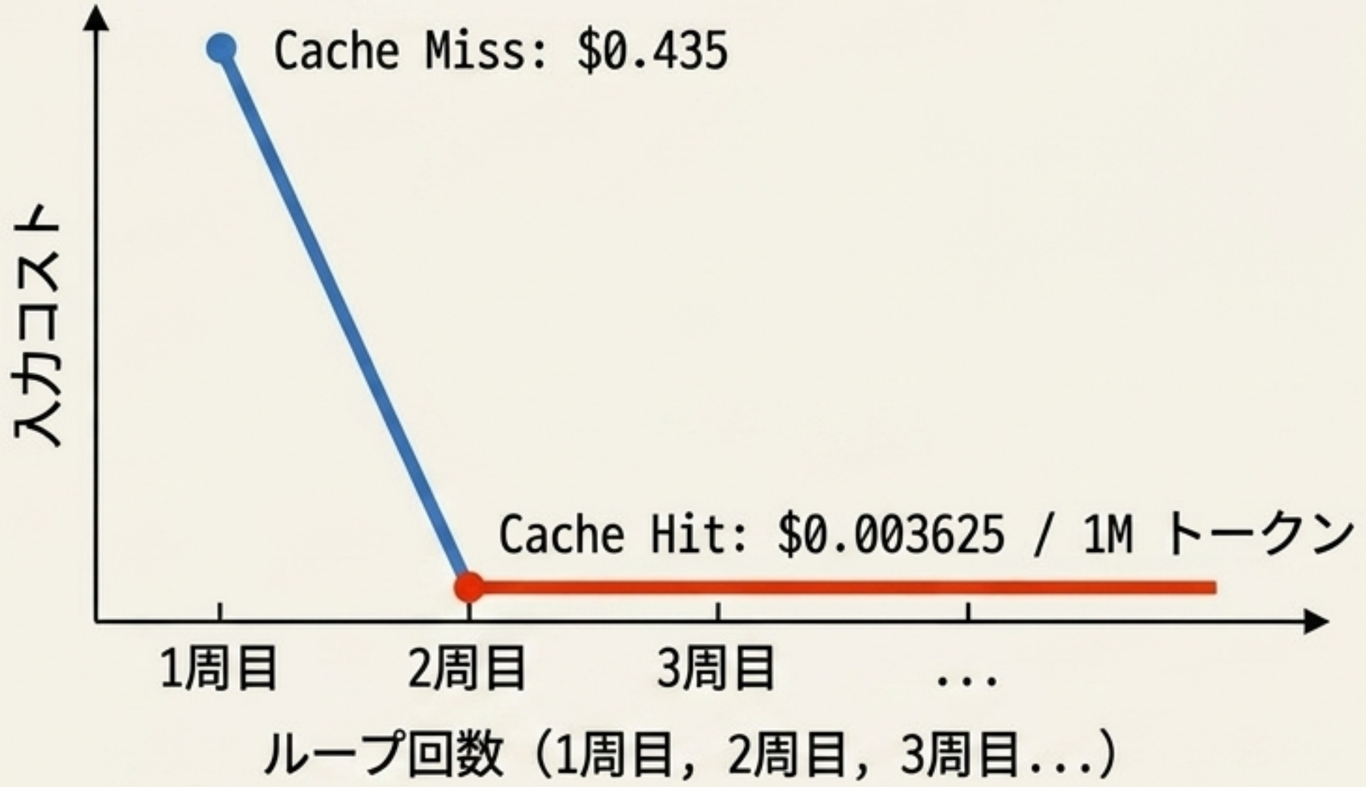
$$\text{入力単価比率} = \frac{\$10.00}{\$0.435} \approx 22.98\text{倍}$$

Key Insight: 単に「57倍安い」のではなく、入出力で削減比率が異なる緻密な価格構造。

コンテキストキャッシュの魔法

自律型エージェント運用における経済的ブレイクスルーの鍵

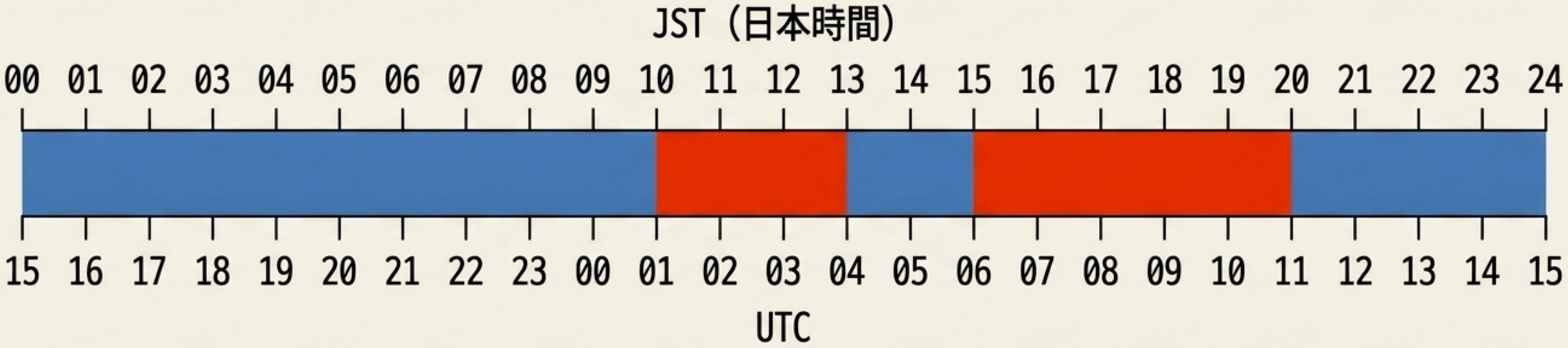




Operational Impact: 長文のコードベース（数十万トークン）を保持したまま、何度もプロンプトを投げて試行錯誤やツール呼び出しを繰り返すエージェントシステムにおいて、システム全体の総合費用を極限まで抑え込む最大の武器となる。

ダイナミックプライシング戦略: ピーク vs オフピーク

トラフィック平準化を活用したバッチ処理のスケジュール最適化



- ピーク料金: 10:00–13:00 & 15:00–19:00 JST（日中の主要業務時間）。1M出力 $3.96
- オフピーク料金: 上記以外の全時間帯。ピーク時の半額（1M出力 $1.98）

Actionable Strategy: 企業ユーザーや大規模バッチ処理を行う開発者は、夜間（JST）にタスク実行をスケジューリングすることで、インフラコストを劇的に削減可能。

エコシステム統合と開発運用上の留意点

```
~~base_url = 'https://api.openai.com/v1'~~
base_url = 'https://api.deepseek.com'
```

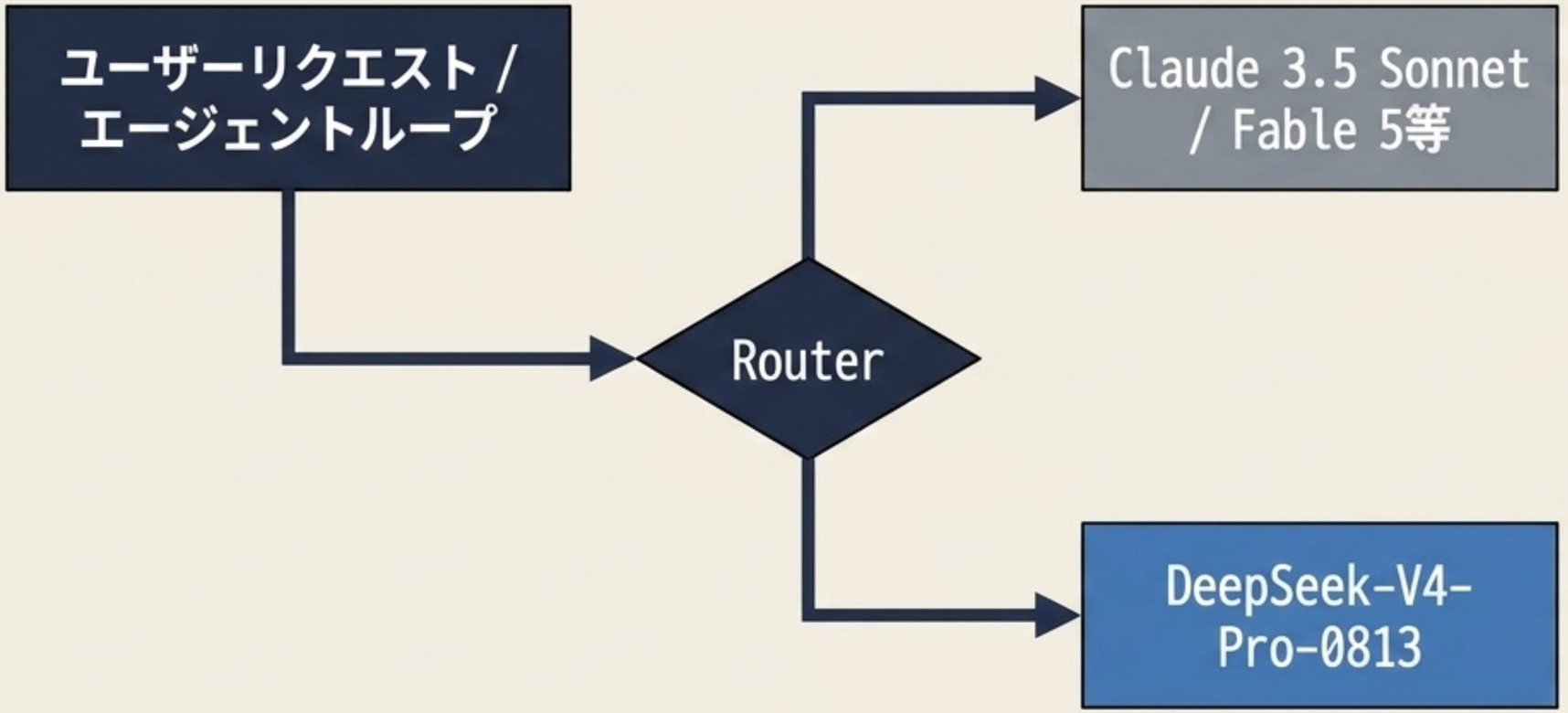
- OpenAI互換 & Anthropic互換APIの標準提供
- FIM (Fill-In-the-Middle)機能対応（非思考モード時)
- OpenAI Responses APIでのネイティブ思考制御(`reasoning_effort`指定)

⚠️ Caution (Hugging Face Deployment)

MITライセンスで重み公開中だが、HF上のデータは「4月プレビュー仕様」の記述が残存。ローカル展開時（vLLM等）はAPI提供の「0813 GA版」と微小な挙動差分が生じる可能性があり、個別検証が必須。

波及効果 1:「DeepClaude」パラダイム

推論プロセスと出力プロセスを分割最適化するハイブリッド・ルーティング戦略

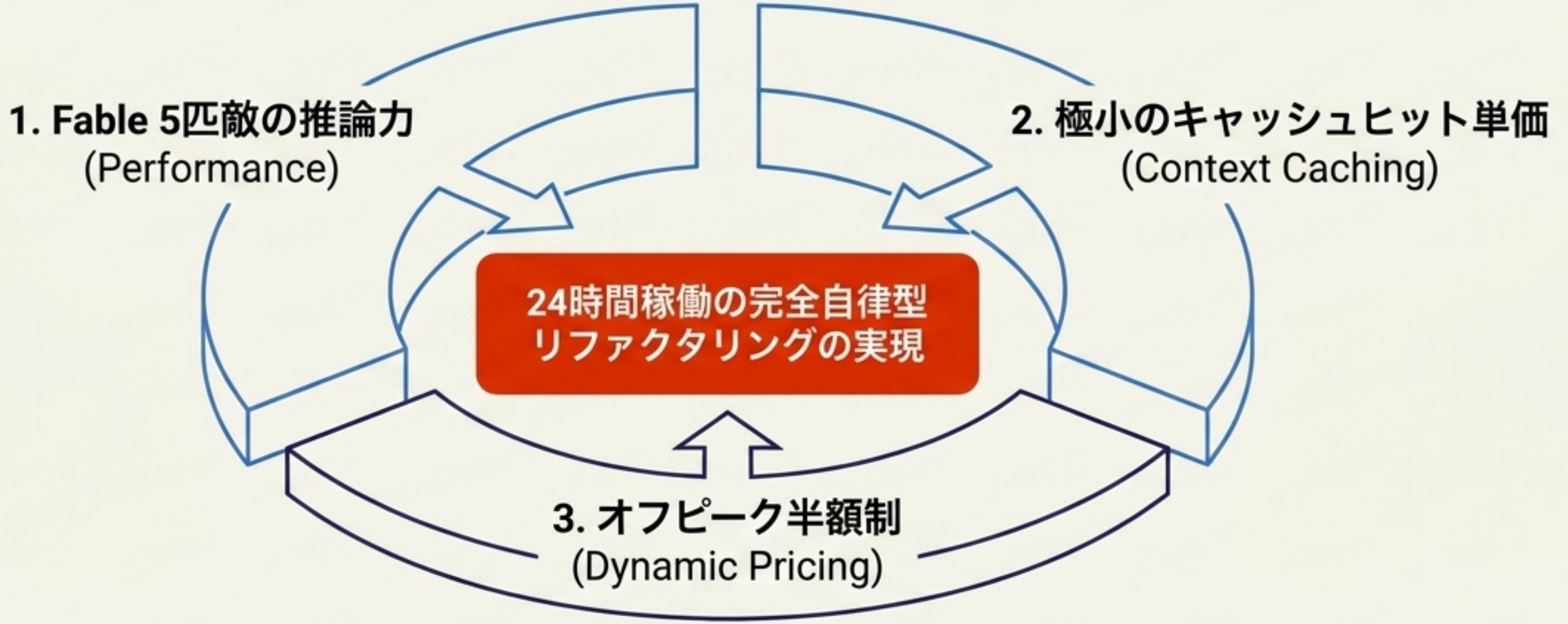


高額なフロントエンド。最終的な整形・高度な出力フォーマット・ユーザー対話を担当。

安価なバックエンド・ワーキングメモリ。大量のコンテキスト読み込み・論理推論・ドラフト作成を実行。

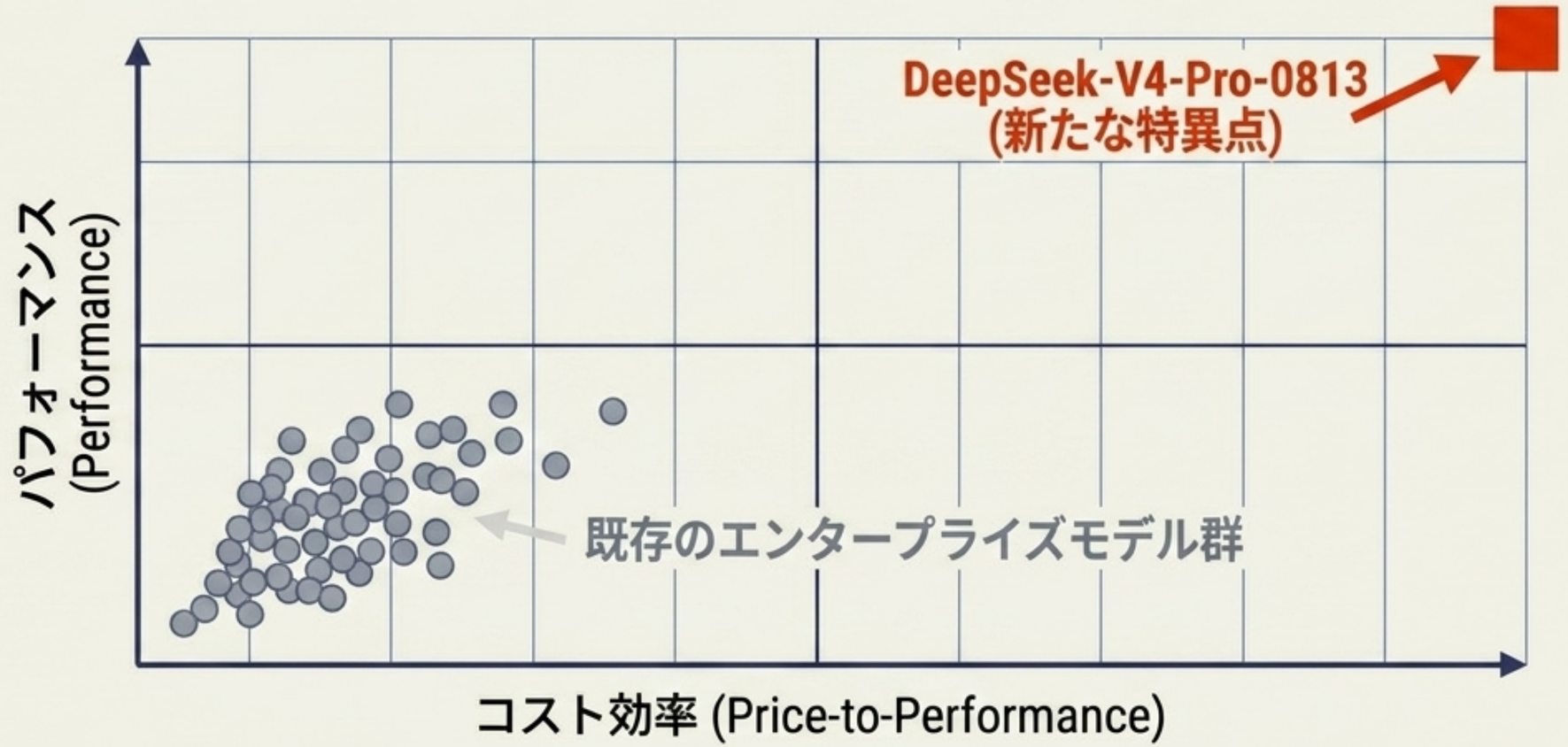
Strategic Impact: 価格破壊により、単一モデルへの依存から脱却。大企業やSaaSベンダーは、適材適所のハイブリッド構成を標準的なアーキテクチャとして採用し始めている。

波及効果 2: Agentic Workflowsの経済的実行可能性



Synthesis Insight: これまで1タスクあたりのAPI費用が高額すぎて実験室レベルに留まっていたAgentic Workflowsが、この3要素の組み合わせによって、**初めて「常時稼働（全自動コード改修や常時モニタリング）」に耐えうる経済性（Economic** Feasibility）を獲得した。

結論: The New Frontier Standard



単なる「安いモデル」ではない。極限まで研ぎ澄まされた計算アーキテクチャが、自律型AIの経済的障壁を完全に破壊した。エンタープライズAIインフラの「新たな標準」はすでにここにある。

Next Action: エージェントループにおけるキャッシュヒット率の検証、およびオフピークバッチのPoCを開始せよ。