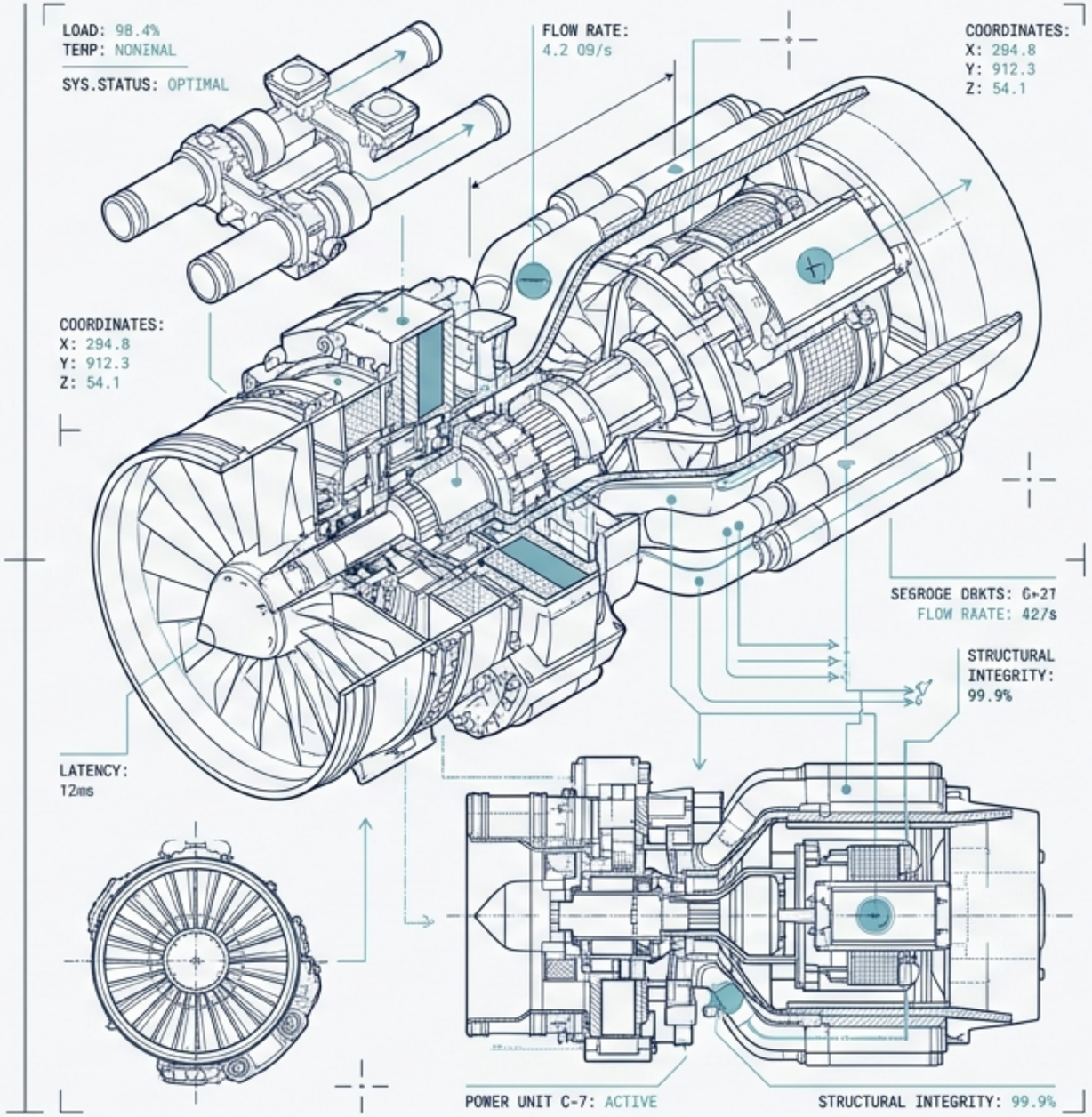
Strategic Evaluation Document (エグゼクティブ向け戦略評価書)

SpaceXAI Grok 4.7

推論アーキテクチャと現場運用の実効性検証

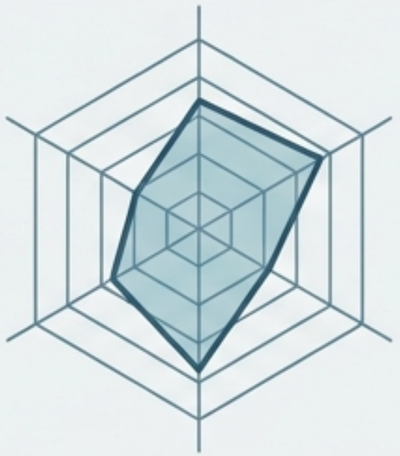
表面的なベンチマークを超えた、エンジニアリング組織のためのアーキテクチャ解析、コスト構造、および最適デプロイメントのプレイブック。



3つの戦略的結論：Grok 4.7の実像

量産型ワークホース

AGIへの野心ではなく、高度な推論・コーディング能力を前世代同一の低単価で大量供給する**「実務のコモディティ化」**。



極端な性能プロファイル

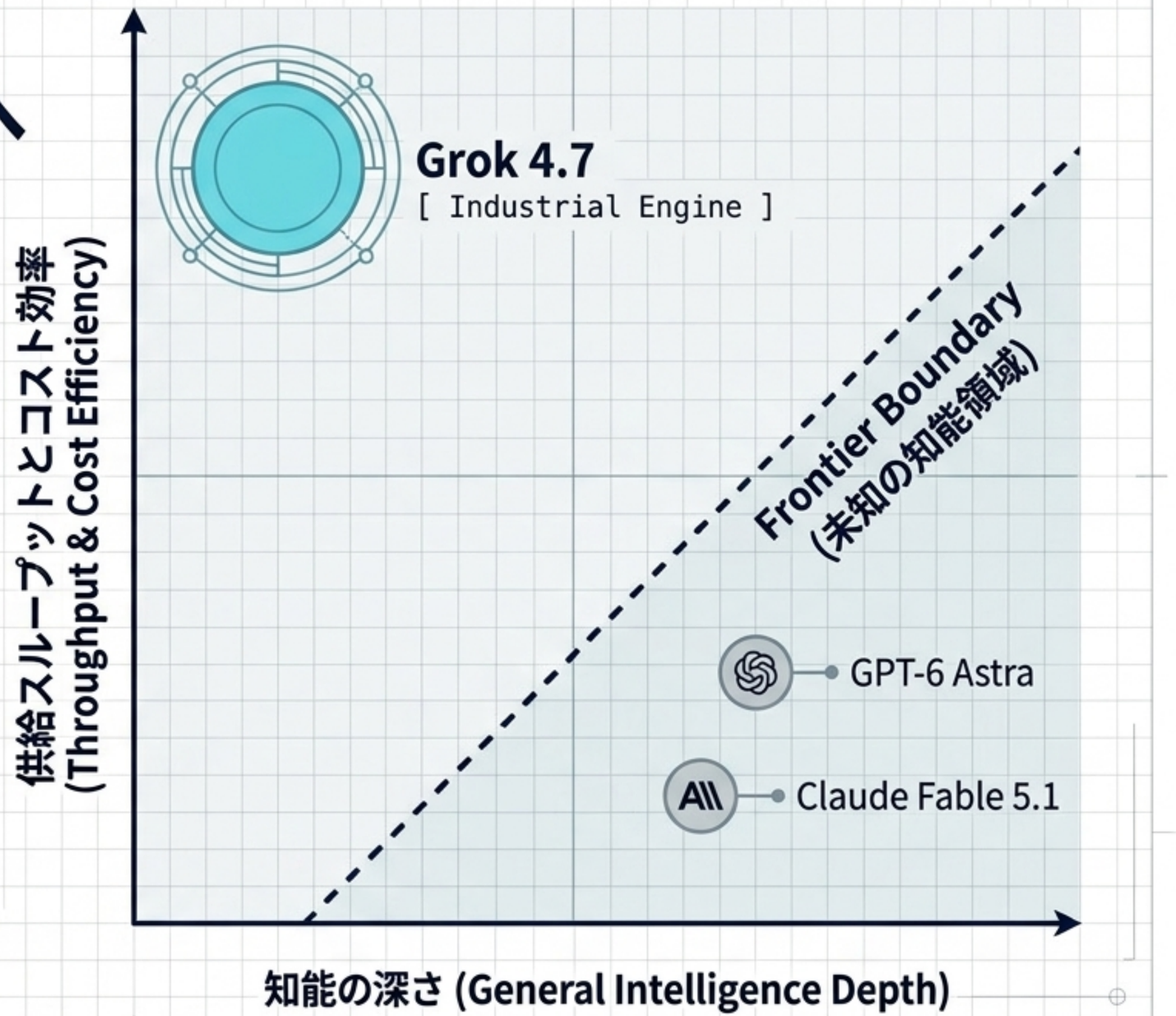
ソフトウェア工学（DeepSWE **71.0%**）で頂点を極める一方、完全自律実行（Terminal-Bench **38.0%**）には**明確な構造的限界**が存在。

コストのパラドックス

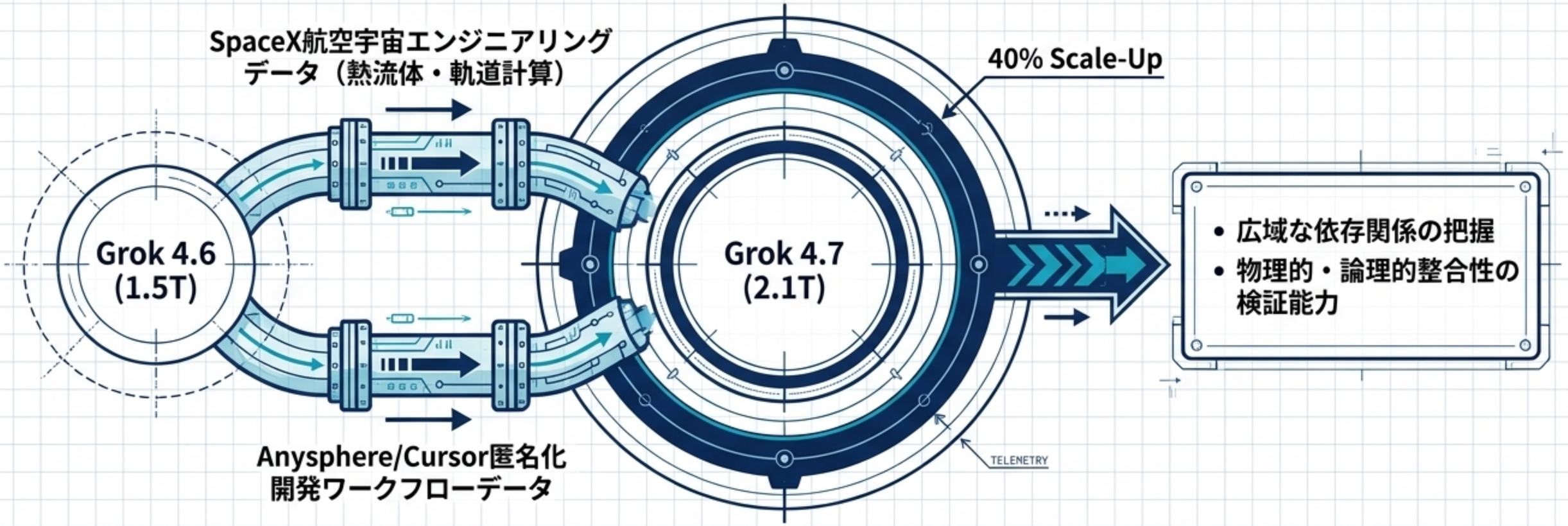
カタログ単価は競合の数分の一だが、**「トークン過消費」**と**「長文ペナルティ」**により、実効タスクコストの優位性は相殺される。

フロンティアへの野心から「実務のコモディティ化」へ

前世代（Grok 4.6）の超低単価（入力$2/出力$6）を据え置いたまま、基盤を拡張。高頻度な日常開発実務のインフラとなることを狙う戦略的配置。



基盤拡張：2.1兆パラメータとドメイン特化データの注入

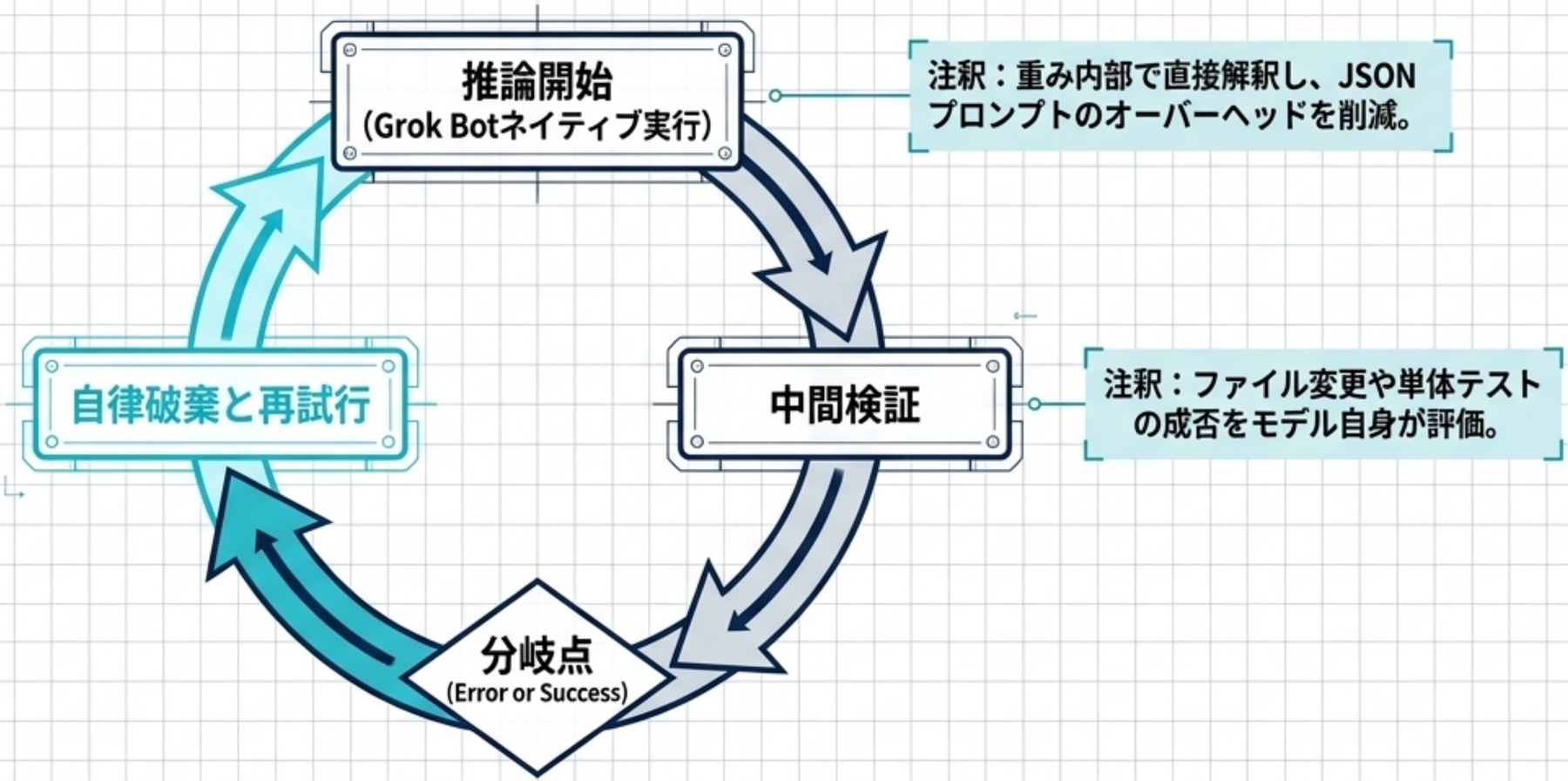


TELEMETRY DATA

1	2	2	3
2.1T	256k - 500k		2026.05
パラメータ規模（2.1兆）	コンテキスト長（標準/最大トークン）		ナレッジカットオフ

TELENETRY DATA

ツールネイティブ学習と「自己検証 (Self-Verification)」ループ



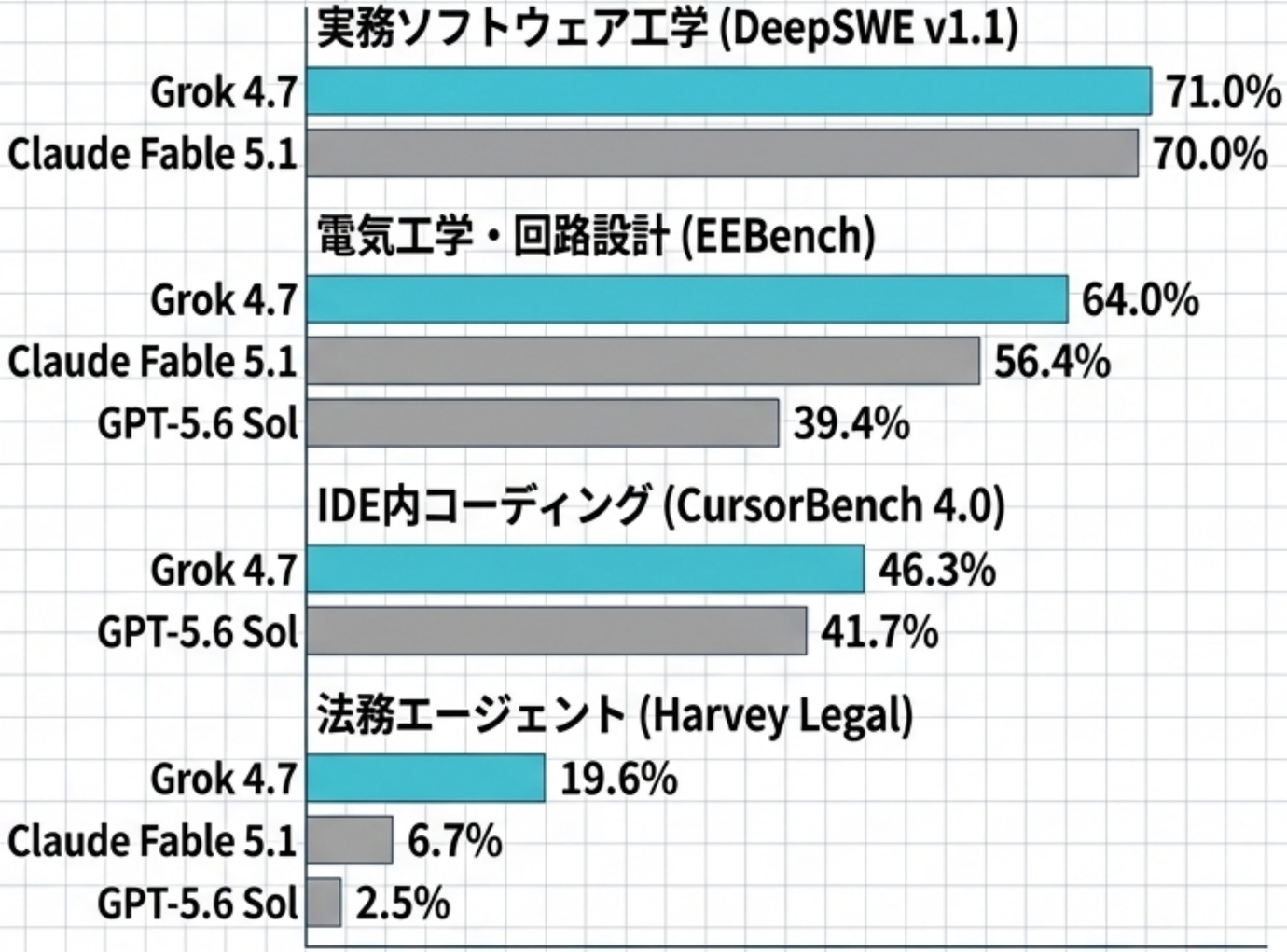
TELEMETRY DATA

TELENETRY DATA

Insight: 数時間にわたる事後強化学習（RL）により獲得。初期段階の微細なエラーが致命的な破綻を招く「エラー増幅現象」を学習レベルで抑制する構造。推論制御（Reasoning Effort）はデフォルトで「high」、オフ不可。

性能プロファイル（強み）：工学・専門推論における圧倒的優位

航空宇宙工学データとCursor実運用データの学習効果が直結。厳密な数理制約を伴う計算や、長大な契約条項照合などの「静的・定型タスク」において、業界最高水準のスコアを記録。

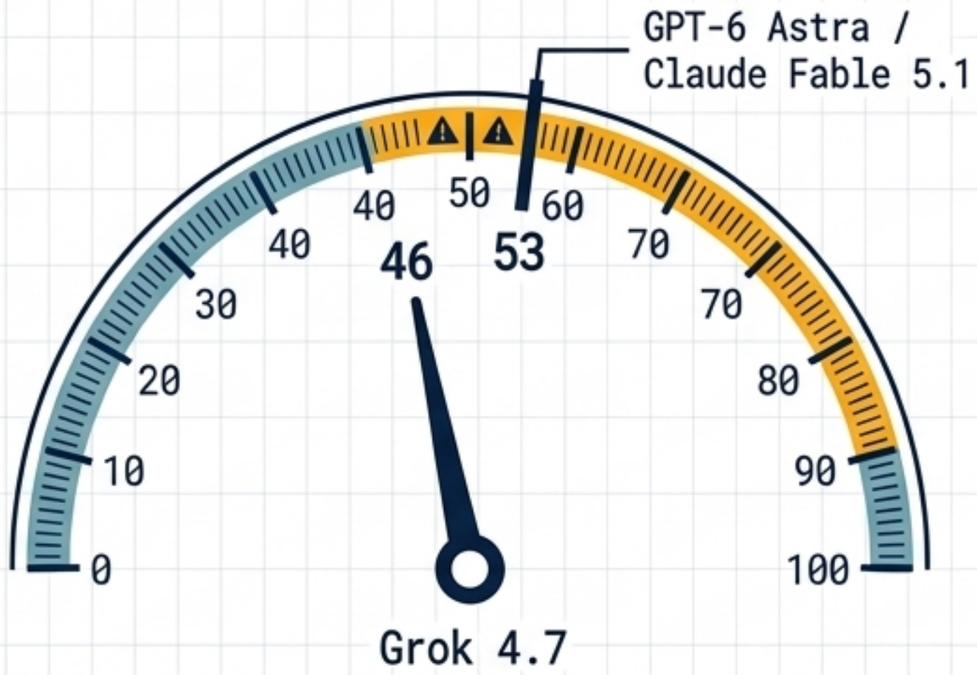


TELEMETRY DATA

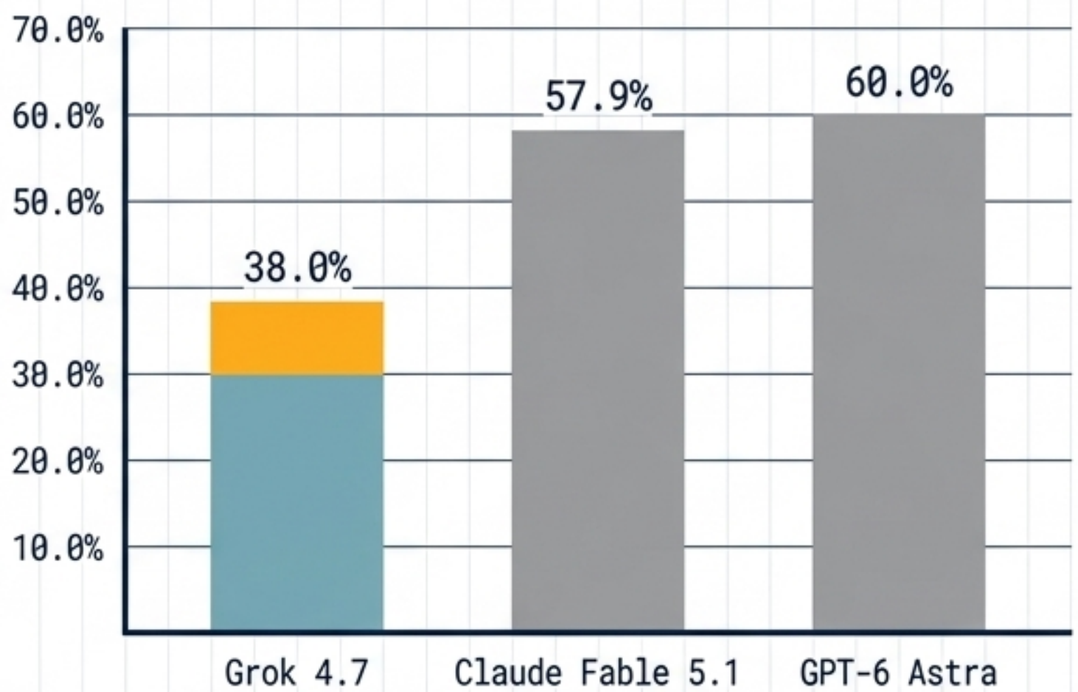
TELEMETRY DATA

性能プロファイル（弱み）：完全自律型エージェントとしての壁

総合知能指数（AA Intelligence Index v4.3.2）



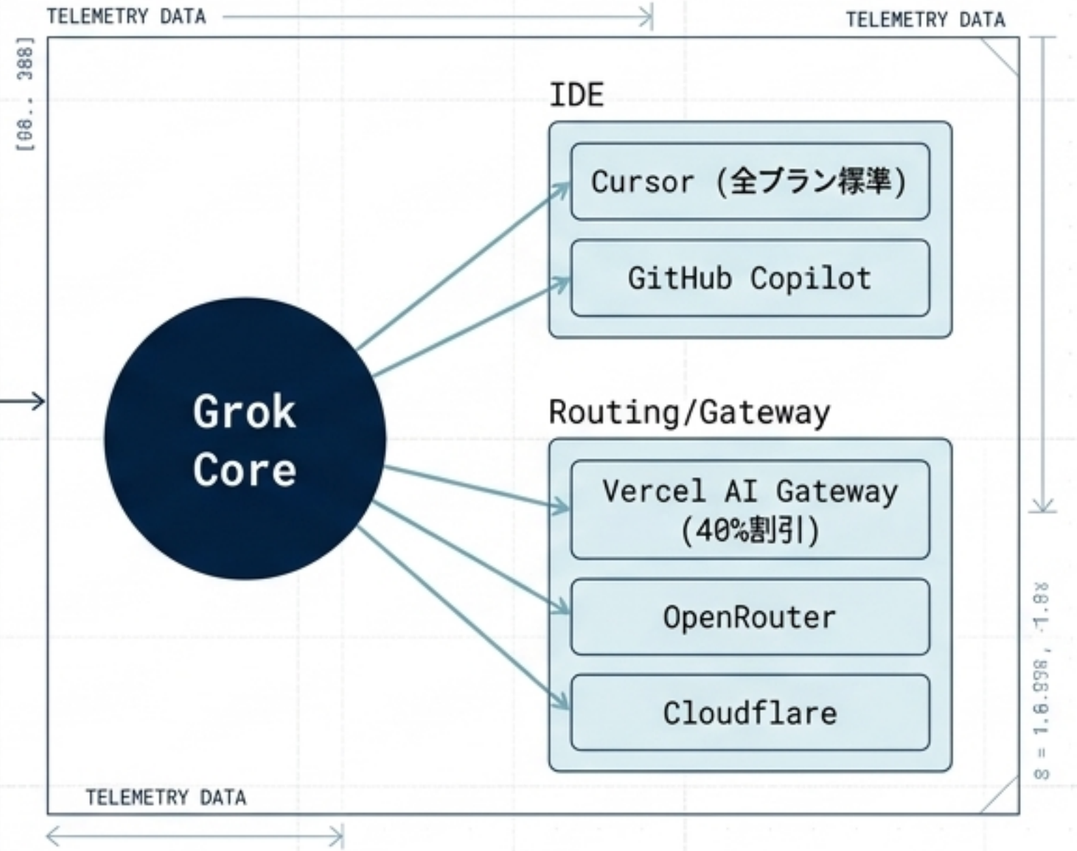
自律型CLI実行（Terminal-Bench 4.0）



シェル環境で数十回のコマンドを連鎖させ、エラーを自律的に復旧させるタスクにおいて構造的限界を露呈。前提条件を取り違え、ループに陥る傾向が顕著である。

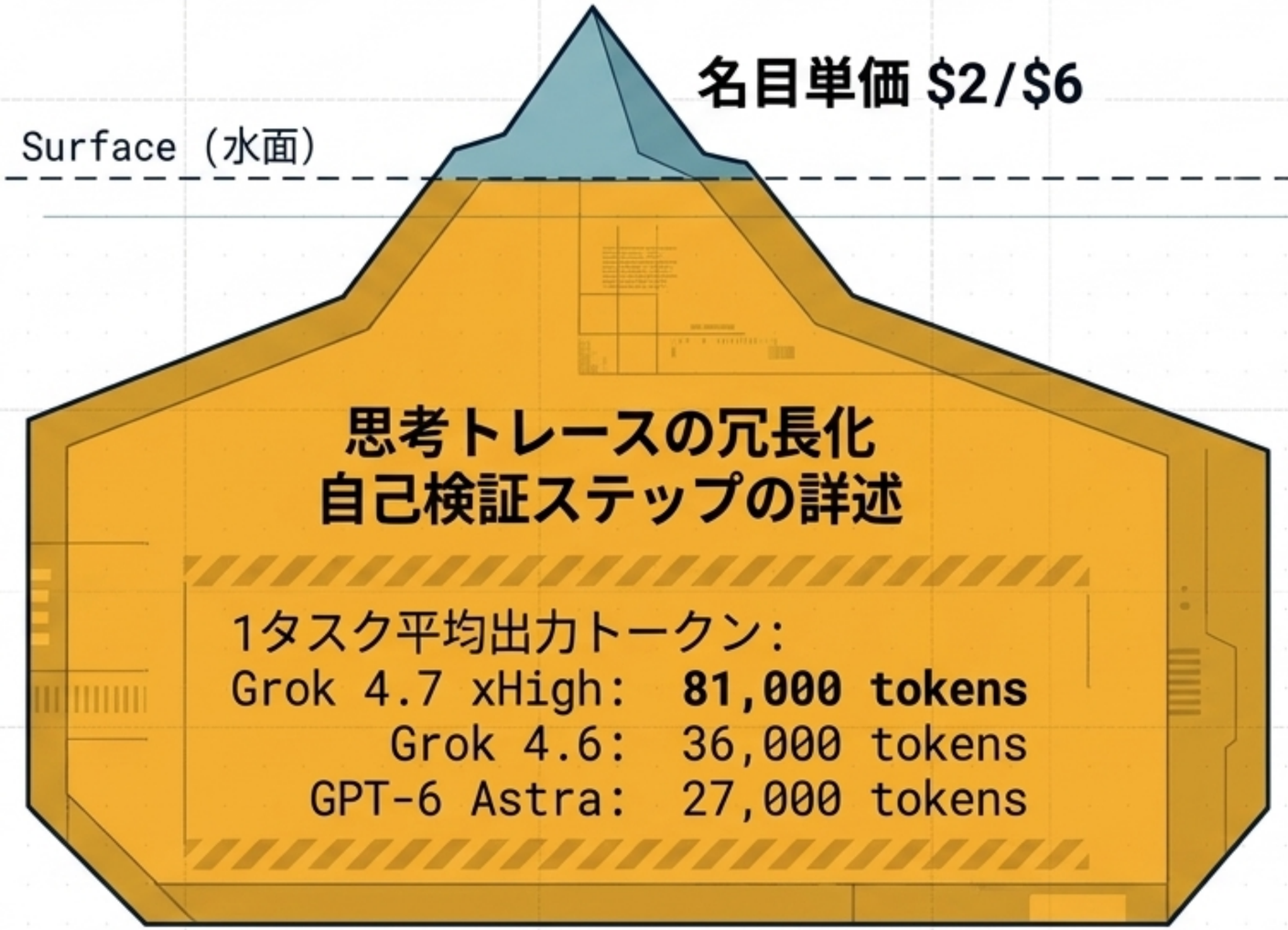
カタログ上の「価格破壊」とエコシステム包囲網

1	2	3
Grok 4.7	GPT-5.6 Sol	Claude Fable 5.1
Input $2.00 / Output $6.00 (Cache Input $0.50)	Input $4.00 / Output $20.00	Input $10.00 / Output $50.00



競合フロンティアの1/5〜1/8という価格設定で、開発者接点を最短で制圧。Blender/Three.js構築タスクの実例では、他社モデルで$2.05かかった処理をわずか$0.20で完了。

隠れたコスト構造：トークン過消費（Token Bloat）パラドックス



タスクあたりの加重平均実効コスト
(Weighted Cost per Task)

Grok 4.7 (xHigh)

$3.74

名目単価の低さが2〜3倍の消費量で
相殺され、実効ROIが低下。

16:9

インフラストラクチャにおける運用上の落とし穴

長文ペナルティ（200k Token Trigger）

リクエストが200,000トークンに達した瞬間、全トークンに対して長文レート（入力$4.00/出力$12.00）が適用される。大規模コードベースの解析時には総請求額が跳ね上がるリスク。

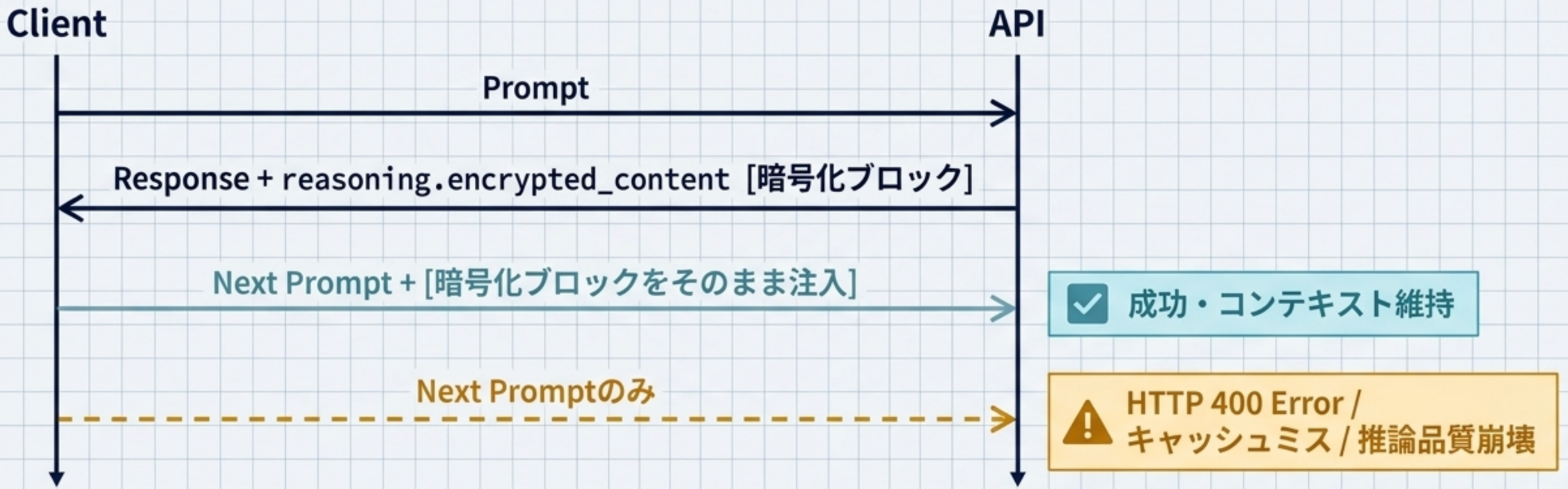
米国内データ主権要件（US Regional）

金融機関等向けに米国内インフラ限定エンドポイント（us.api.x.ai）が存在するが、標準単価から10%割増となる。

速度優先の回避策（Grok 4.7 Fast）

CursorおよびGrok Build専用に、通常比2倍の生成速度を持つFastモデルが提供されている（デフォルトのxhigh設定は出力39.3 t/s、レイテンシ0.88秒）。

開発者警告：暗号化推論コンテキストの往復要件



Responses APIは常に推論トレースを暗号化して返却する。クライアント側でこれを保持し、次ターンで未改変のまま戻さなければ「思考コンテキスト」が喪失する。presencePenaltyやstopパラメータも非対応。

16:9

防御的セーフガードスタックと高リスク領域への適合

バイオセキュリティ (LatchBio 62.4%)

- ✕ 病原体兵器化プロトコル
- ✕ 危険な合成手順

defense filter

- ✓ 創薬支援
- ✓ タンパク質フォールディング解析

フロンティアモデル首位の適合率。過剰拒否を回避。

サイバーセキュリティ (HackerBench誤通過率 3.3%)

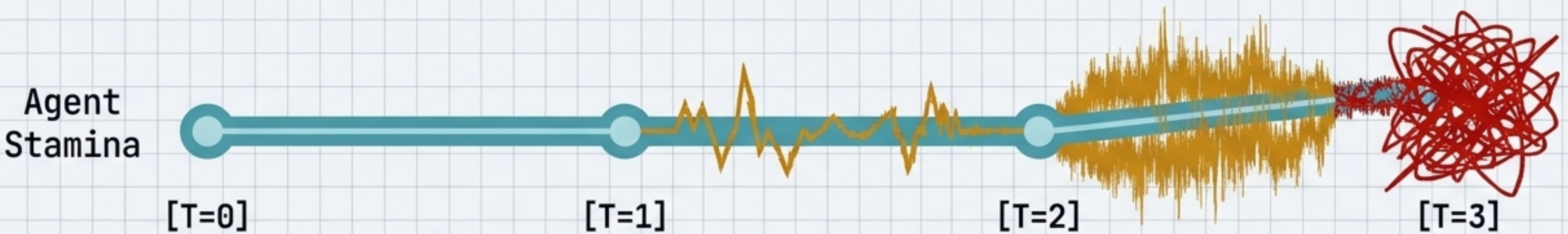
- ✕ ゼロデイ脆弱性悪用コード (96.7%遮断)

defense filter

- ✓ ペネトレーションテスト作成
- ✓ パッチ適用
- ✓ 脆弱性修正

破滅的な二重用途（Dual-use）リスクを厳格に遮断しつつ、開発生産性を下げないキャリブレーション。一部パートナーにはレッドチーム評価機能を先行提供。

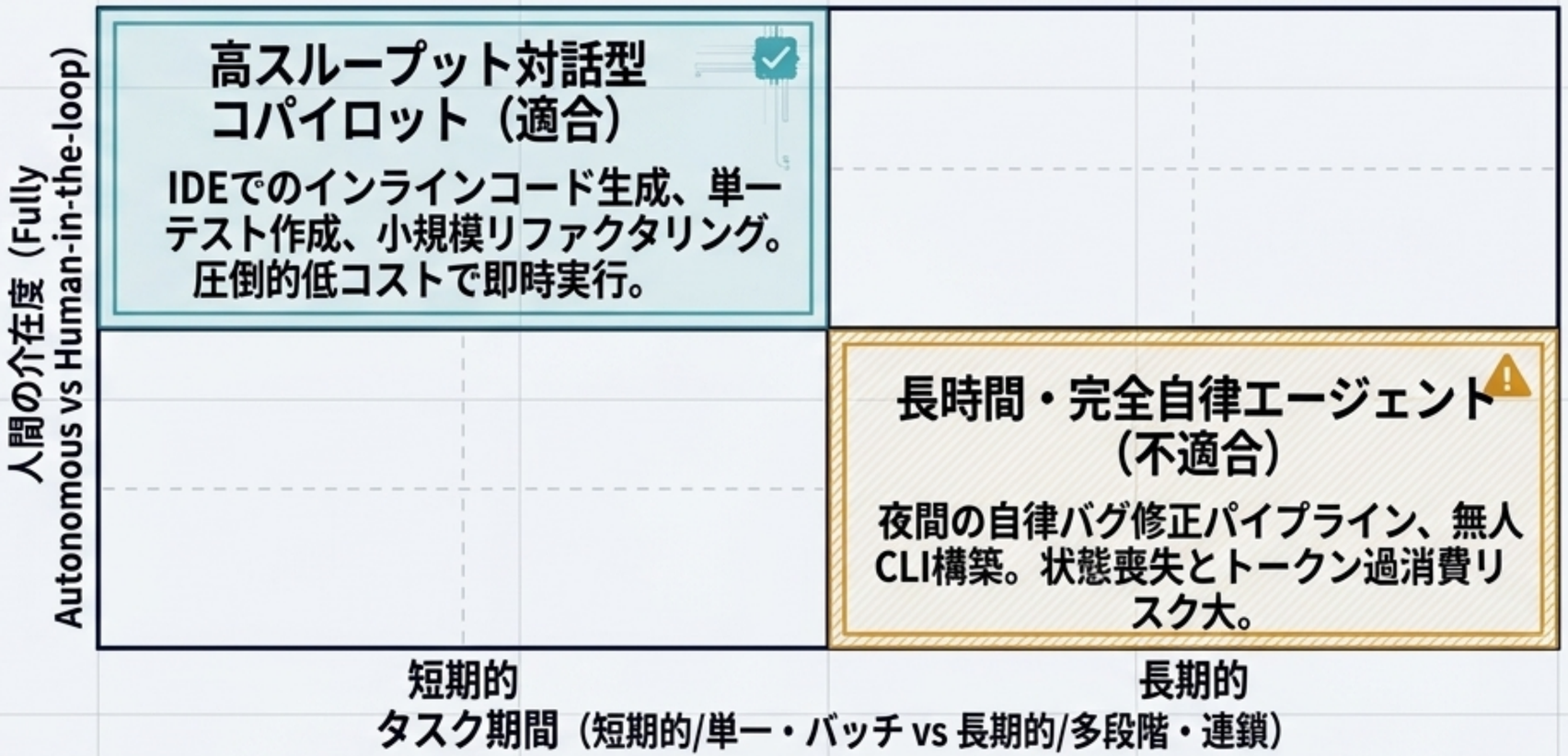
長時間自律タスクにおける「エラー連鎖」とコンテキスト崩壊



数時間のタスク前提で訓練されたものの、人間が介在しない完全自律環境（夜間バックグラウンドでのバグ修正等）では、高度な文脈保持力を持つClaude Fable 5.1に対して信頼性で明確に劣る。

16:9

実践プレイブック：最適なデプロイメント・マトリクス



Synthesis: Grok 4.7は「完全自律型エージェント」として過信すべきではない。人間がループ内に介在する反復的コーディング支援として適材適所で組み込むことが、最大の投資対効果を生む。

SpaceXAIのロードマップとエンジニアリング組織のアクション

Future: Grok 4.9 / Grok 5
Astra/Fable級への到達目標。

Next: Grok 4.8 (2.5T)
事前学習完了、強化学習移行中。

Now: Grok 4.7 (2.1T)
Opus 5.0水準、開発実務のコモディティ化。

1

1. IDE連携の即時有効化

Cursor/GitHub Copilotにおける日常的なコード生成をGrok 4.7へルーティングし、APIコストを削減する。

2

2. 推論テレメトリの監視

長文コードベース投入時の「200kトークン超えペナルティ」と「トークン過消費」をダッシュボードで監視する。

3

3. アーキテクチャのステート管理

自社アプリへのAPI組み込み時、暗号化推論コンテキストの往復（ラウンドトリップ）が確実に行われるよう設計を見直す。