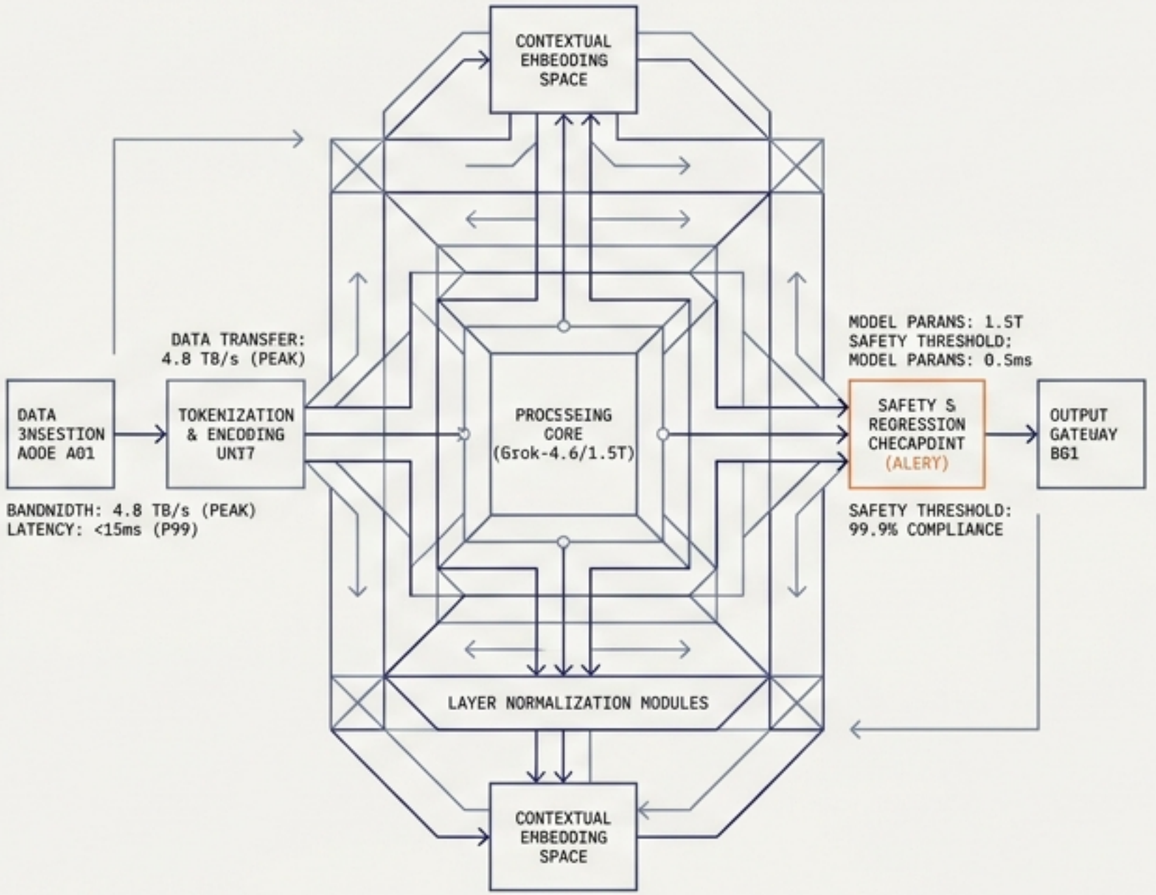
Grok 4.6 実装に向けた技術解剖と戦略ブリーフィング

TARGET: Enterprise Architects & AI Strategy Leads

FOCUS: Cost Economics, Safety Regressions, Contractual Risks

MODEL ID: grok-4.6 (1.5T-scale family)



最終判定：Grok 4.6の真価は「知能の向上」ではなく「自律タスクの経済性」にある

61

AA Intelligence Index

フロンティア級の推論能力。GPT-5.6 Solと同等の知能水準に到達。

$6 vs $30

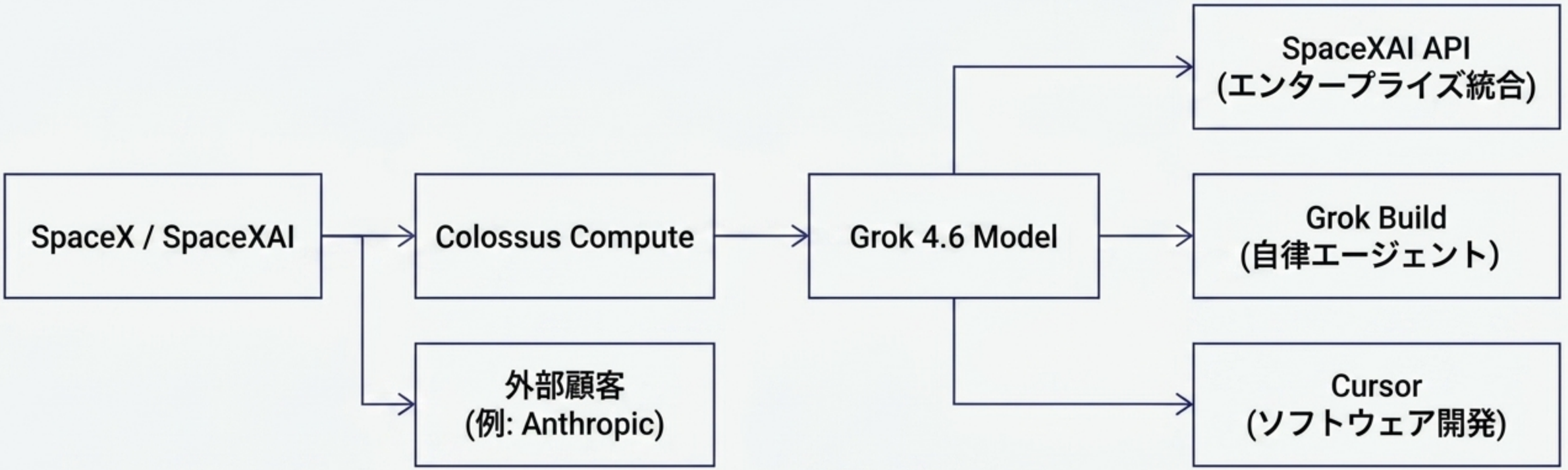
長時間の自律エージェント処理における圧倒的なコスト優位性とステップ効率。

⚠ 3つの企業導入トラップ

1. 20万トークン境界での劇的な価格倍増
2. 幻覚率（Hallucination）の悪化（0.98% → 1.7%）
3. 契約上の「ベンチマーク実施禁止」条項

SpaceXAIの垂直統合エコシステム

インフラから開発環境までのシームレスな統合



二面戦略：計算インフラ市場で競合を顧客としつつ、Cursor統合により開発者への配布摩擦をゼロ化。

Grok 4.6 アーキテクチャと学習パイプライン

ASSEMBLY LINE

Phase 1: データ収集
公開データ・社内データ・匿名化Cursorデータ

↓

Phase 2: 基礎学習
1.5Tスケール / 潜在的なMoEアーキテクチャ

↓

Phase 3: SFTとフィルタリング
軌跡の再生成とモデルベースの問題除外

↓

Phase 4: エージェント特化の強化学習
知識労働・コーディング・CAD環境での徹底的なAgentic RL

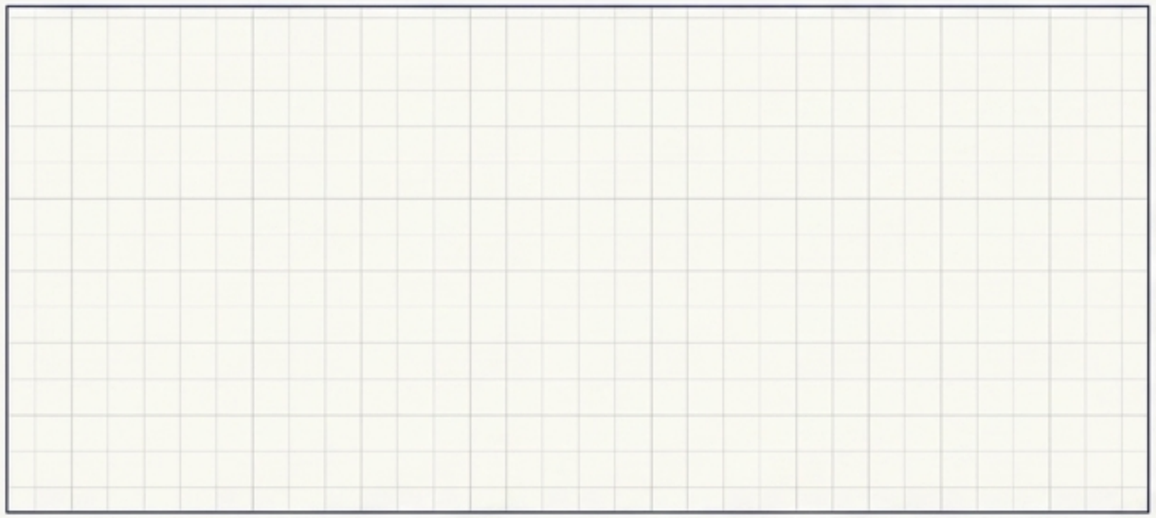
TECHNICAL SPECS

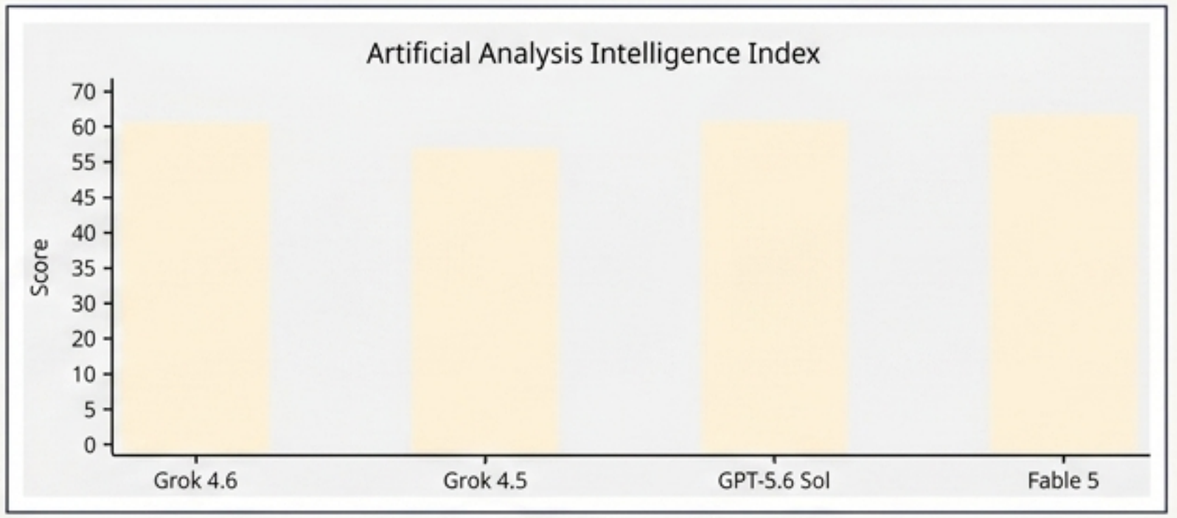
CONTEXT:
500k tokens

KNOWLEDGE CUTOFF:
2026-02-01

LATENCY:
約31秒 (Time to First Token)

フロンティア・ランドスケープ：現行モデルとの診断マトリクス





[Model]	[Target Peer]	[Context]	[Base Price (In/Out)]	[Verdict]
Grok 4.6	GPT-5.6 Sol	500k	$2 / $6	低単価の長期エージェント向け
GPT-5.6 Sol	Grok 4.6	—	$5 / $30	2026年の最上位推論基準
Gemini 3.1 Pro	—	1M	$2 / $12	超大容量コンテキスト処理
Llama 3.1 405B	—	128k	Self-host	オンプレミス・データ主権

※GPT-4oや初代Llama 3等の「2024年世代」との直接比較は評価軸として不適切である。

性能の二面性：総合スコアが隠す「勝敗」の境界線

優位領域：知識労働・エージェント

APEX-Agents	Grok 4.6: 57.5% Sol: 56.7%
OfficeQA Pro	Grok 4.6: 63.2% Sol: 60.2%

複数ステップを要するリサーチや文書分析など、長時間のオーケストレーションで凌駕。

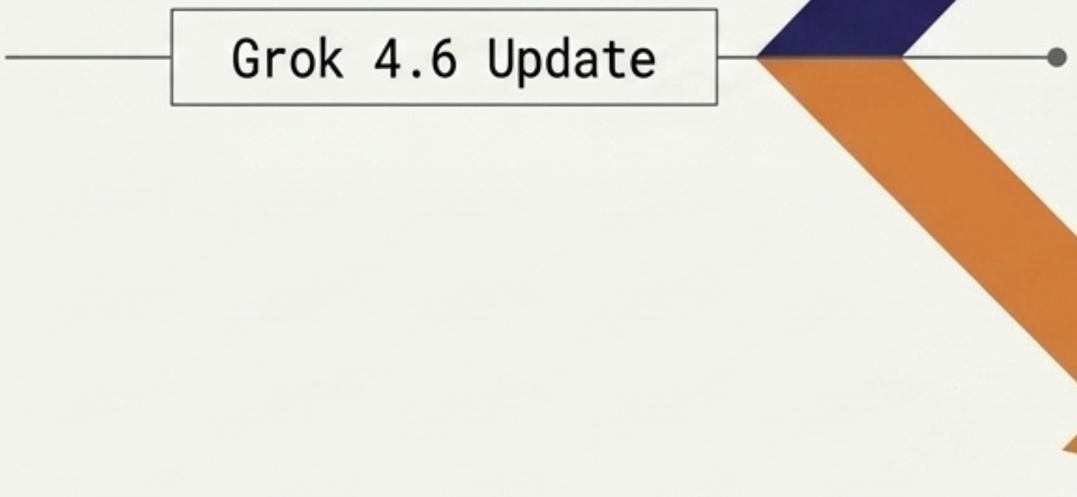
劣後領域：純粋なソフトウェア修復

DeepSWE 1.1	Grok 4.6: 65.9% Sol: 73.0%
Terminal-Bench 3.0	Grok 4.6: 26.0% Sol: 34.6%

純粋なターミナル操作や高度なコード修復においては、競合モデルに明確な後れを取る。

セキュリティと信頼性の乖離

「ハッキング耐性」は向上するも「業務上の正確性」は悪化



サイバーセキュリティ防衛力（改善）

標準Jailbreak率: 0.73% → 0.04%
CBRN拒否精度: 97.9% → 100%

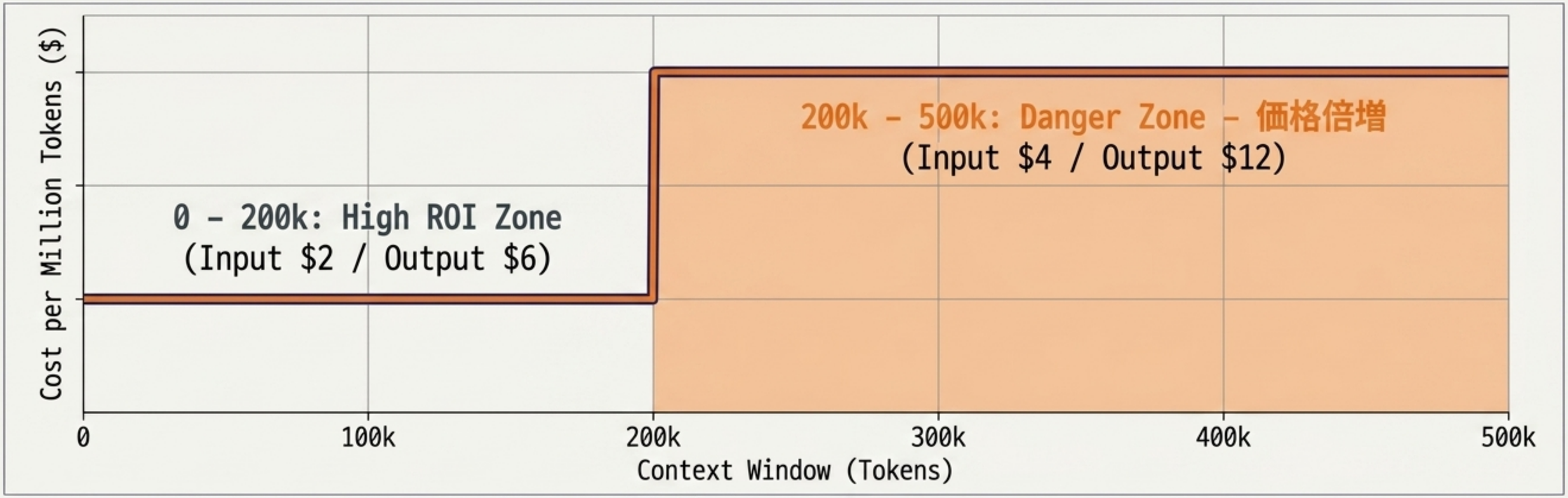
悪意ある攻撃や危険要求に対する防御力は飛躍的に向上。

エンタープライズ信頼性（悪化）

幻覚（Hallucination）率: 0.98% → 1.7%
不誠実な応答（Dishonesty）率: 0.67% → 3.8%

マーケティング表現に反し、事実誤認率や不誠実な応答率が悪化。事実検証型アプリでは独自評価が必須。

コストの断崖：20万トークン境界に潜むTCOの罠



WARNING: 50万トークンのコンテキストを無計画に消費する長時間エージェント設計は、ヘッドライン価格に基づく予算見積もりを破壊する。

MITIGATION: エージェントループにおけるコンテキストの動的圧縮と、キャッシュヒットの確実なルーティングが必須。

統合リスク・コンパス：導入前にクリアすべき4つの障壁

契約上の罠（Legal）

企業利用規約にてベンチマークが禁止。社内比較検証時は事前承認の確認が推奨される。

インフラ制約（Operations）

現時点でBatch API非対応。大量の非リアルタイム処理を低コストで実行するアーキテクチャが不可。

データ主権（Sovereignty）

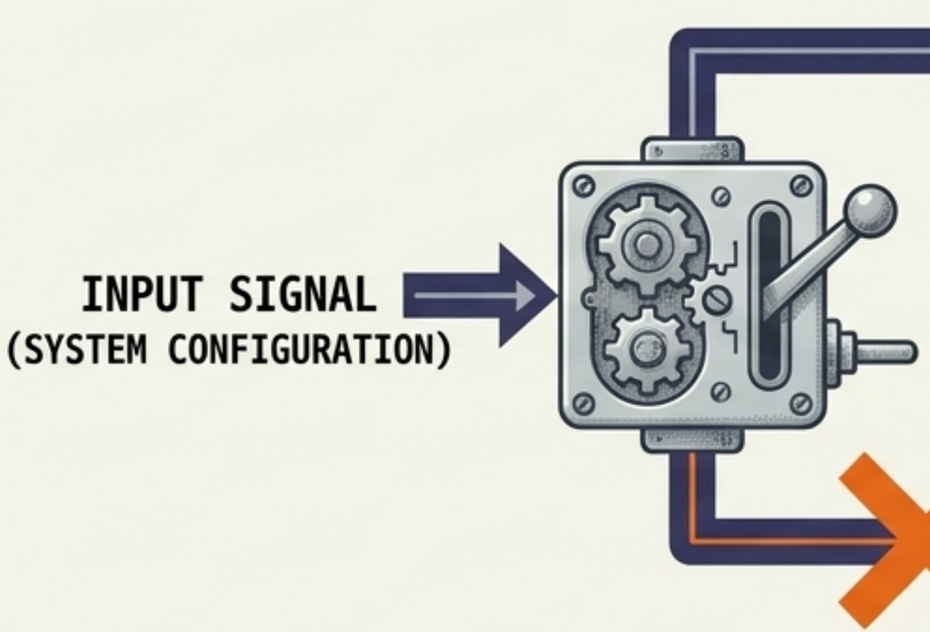
オープンウェイトやオンプレミス版が欠如。厳格なエアギャップ環境には不適合。

高リスク判定（Safety）

人間の監督なしでの自動意思決定は規約で禁止。医療・法務・金融では意思決定支援に留める。

データ・プライバシーの代償：ZDRのトレードオフ

ZDR（Zero Data Retention）を有効化すると、重要なステートフル機能が犠牲になる



ZDR OFF（標準設定）

- データ保存: 30日間（不正利用監査目的）
- 有効機能: Files, Collections, 過去の対話履歴利用

ZDR ON（エンタープライズ・プライバシー）

- データ保存: 永続保存なし（HIPAA要件等に合致）
- 無効機能: 全てのステートフル機能が停止。コンテキスト管理は完全な自社実装が必要。

ACTION: APIレスポンスのZDRヘッダーをプログラム的に検証する運用設計が不可欠。

評価指標のパラダイムシフト

「トークン単価」から「タスク完了単価」へ

旧パラダイム（GPT-4時代）

Cost =（入出力トークン数）× × トークン単価

新パラダイム（Grok 4.6 エージェント経済）

Cost =（ステップ毎トークン数 × ターン数）× トークン単価

Grok 4.6の実力

1. 複雑な知識労働タスクを平均53ターンで完了（競合の約半数）。
2. 初回トークン生成(TTFT)に約31秒を要するが、極めて高い「軌道効率（Trajectory Efficiency）」により、タスク当たりの実コストはわずか$0.84に収束する。

$0.84

タスク当たり実コスト

ワークロード適合性マトリクス

Grok 4.6の最適な戦場と回避すべき領域

HIGH FIT（最適）	リポジトリ横断コーディング	調査・ナレッジワーク
	Cursor統合済み。エージェント強化学習により中核的強みを発揮。	ターン効率が高く、研究開発・情報集約に最適。
CAUTION（要注意）	大規模低遅延チャット	20万トークン超の大量処理
	推論設定時の初動に数十秒を要し、即時応答UIには不適。	単価倍増の境界線。コンテキスト枠の厳密な管理が必要。
DO NOT USE（不適合）	ハイステークス自動最終判断	完全エアギャップ・オンプレミス
	医療・法務等での完全自律判断は利用規約上禁止。	オープンウェイト未提供。代替モデルの利用を推奨。

デプロイメント・ブループリント

実装に向けた3つのアクション

01 PROCURE（調達と法務）

アクション

契約主体「X.AI LLC」との契約条件確認。

リスクヘッジ

独自のベンチマーク検証を実施する前に、法務経由で規約上の制限例外をクリアする。

02 ARCHITECT（システム設計）

アクション

トークン経済性の死守。

リスクヘッジ

20万トークン超過を防ぐ動的コンテキスト圧縮の実装。
機密データ処理時のZDRヘッダー検証。

03 DEPLOY（運用とモニタリング）

アクション

非同期インターフェースの採用。

リスクヘッジ

初回遅延（TTFT）を許容するUX設計。
幻覚率1.7%をカバーするRAGベースのグラウンディング。