

クラウド依存AIの終焉

Perplexity「Portable Computer」が提示する、
ローカルファーストへのパラダイムシフトと
エンタープライズ戦略

Architecture

**100%ローカルの「理想」と
ハイブリッドの「現実」**

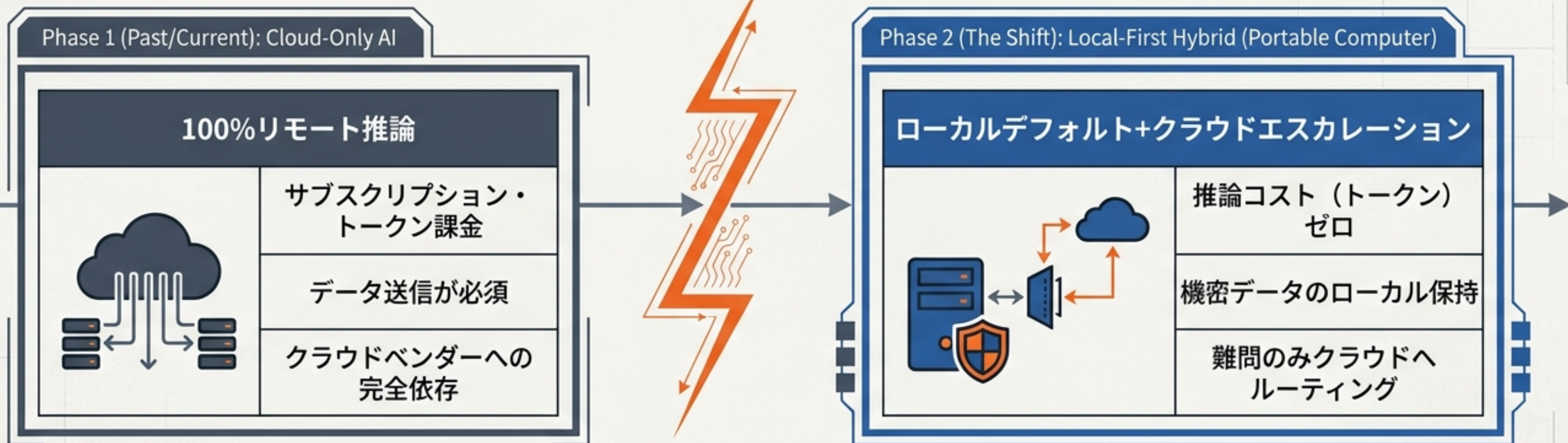
Engineering

**モデルサイズから「ハー
ネス工学」への主戦場移行**

Security

**OWASP基準で紐解く
エッジAIの真の脅威モデル**

AIコンピューティングの進化： 「APIでの智能調達」から 「ローカルハードウェアの所有」へ



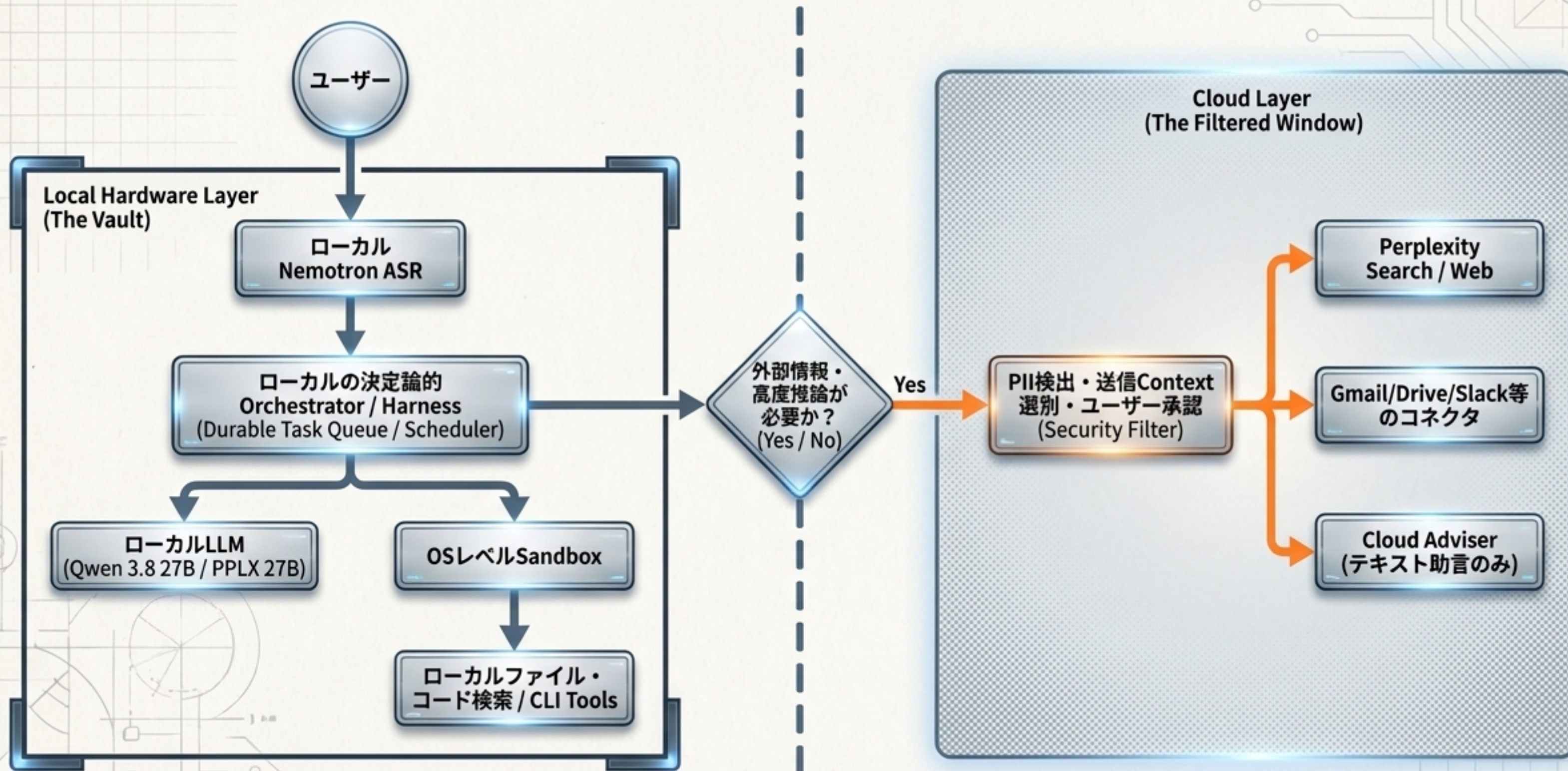
「クラウドかローカルか」の二元論ではなく、ローカルを通常経路、クラウドを例外経路とするハイブリッド型が実用段階に到達した。

マーケティングのレトリック vs. エンジニアリングの現実

Claim (X/PR)	
○	「100%ローカルで動作。クラウド依存なし」
○	「オーケストレーターLLMが全てを制御」
○	「トークンコストはゼロ」

Engineering Fact	
✓	オプションル・ハイブリッド。 基本はローカル（Qwen 3.8 27B / PPLX 27B）だが、高度な推論や外部情報が必要な場合はクラウドの「Adviser」へエスカレーションする設計。
✓	決定論的ハーネスコード。 LLMがシステムを支配するのではなく、決定論的なコード（Planner, Router）がループを管理し、LLMは「提案」に留まる。
✓	ローカル推論分のみ。 クラウド側のComputer機能（Adviser等）には別途クレジット体系が存在。

アーキテクチャ解析：ローカルとクラウドの境界線



性能の源泉：「モデルサイズ」ではなく「ハーネス工学」

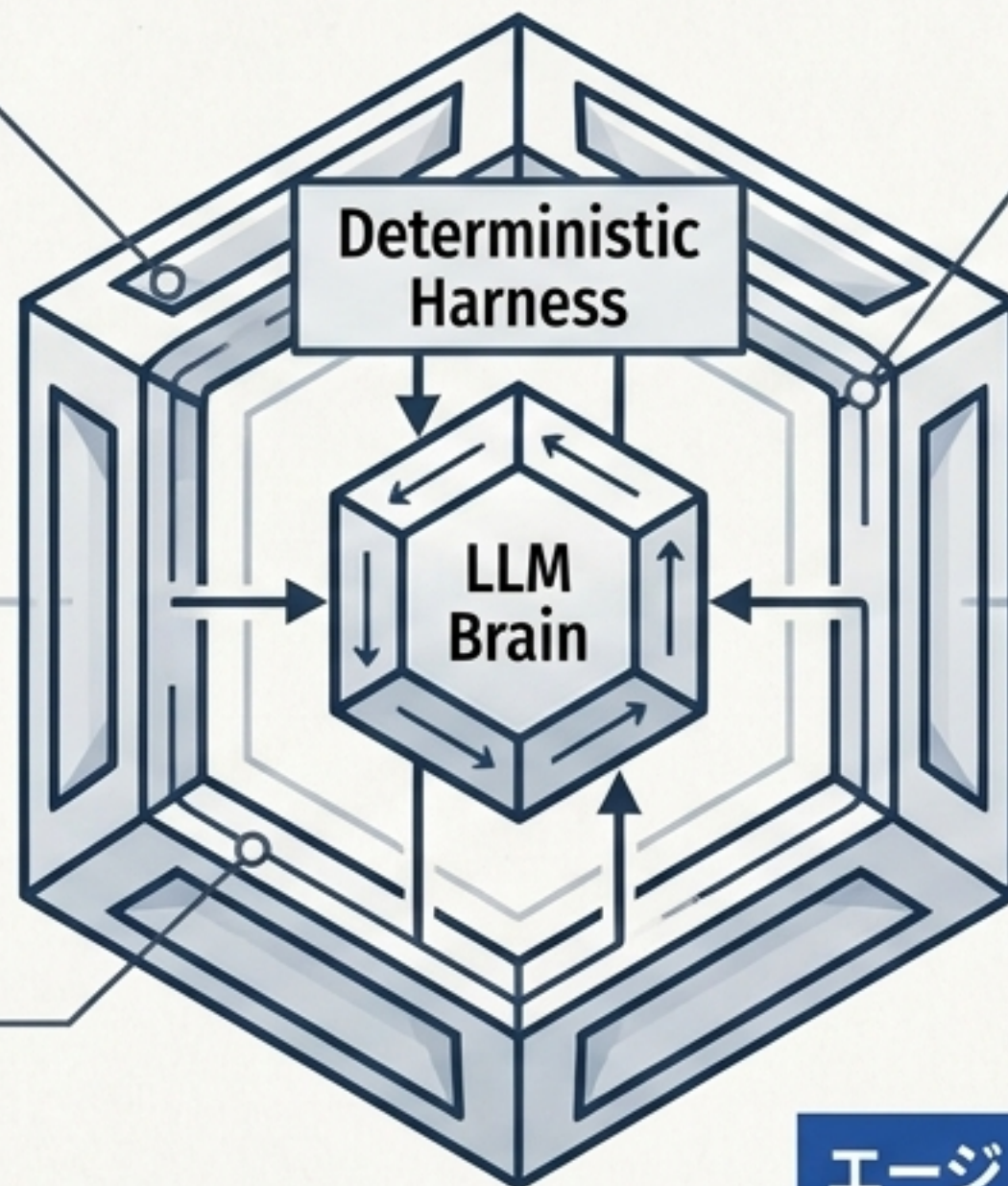
Core Concept: LLMは制御しない。決定論的コードが支配する。

Optimization 1: Context Compaction

260Kトークンの限界を回避。
100K超での性能不安定化を防ぐ
ため、極小のシステムプロンプト
とコンテキスト圧縮を採用。

Optimization 3: On-Demand Skills

必要なツール (Planner, Router)
のみを都度ロードし、メモリと
コンテキストを節約。

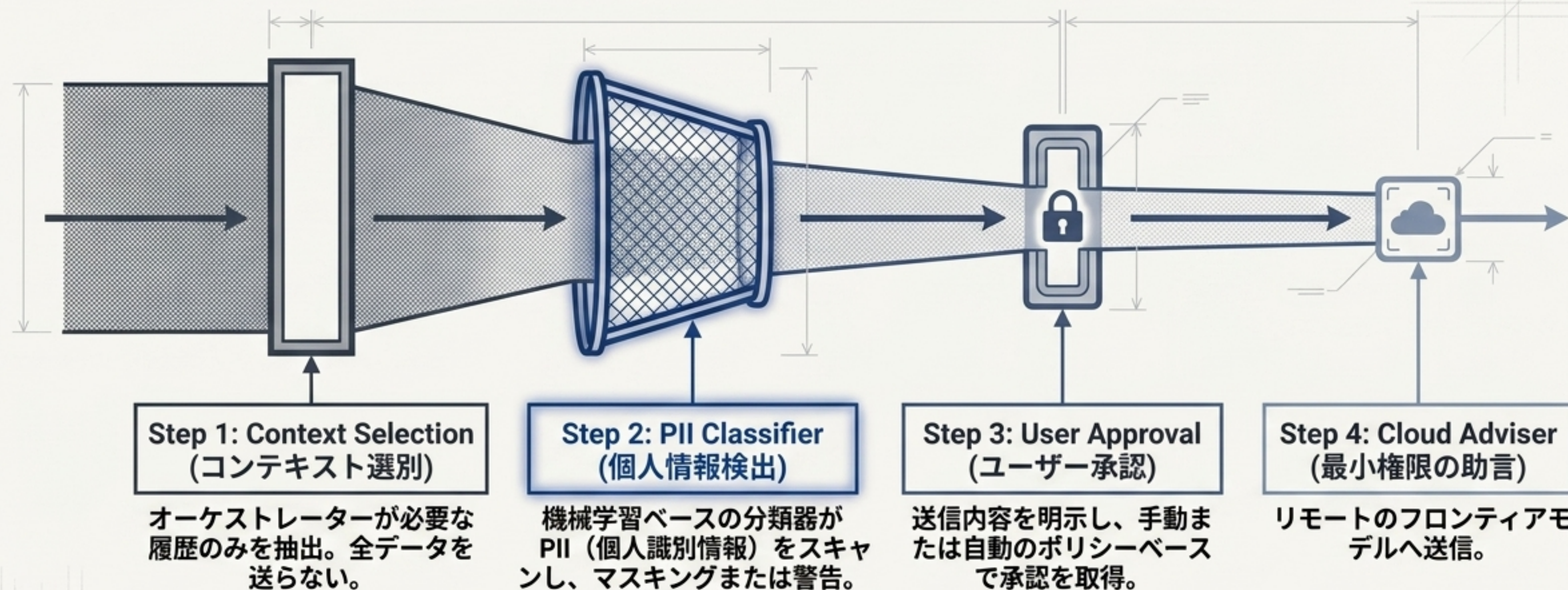


Optimization 2: Compact CLI Tools

巨大なMCPツール定義を避け、
頻用MCPをコンパクトなCLI
ツールへ変換。

エージェントの性能は、単なる27Bモデルの能力で
はなく、小型ローカルモデルに最適化された独自
ハーネスとの共同設計によって生み出されている。

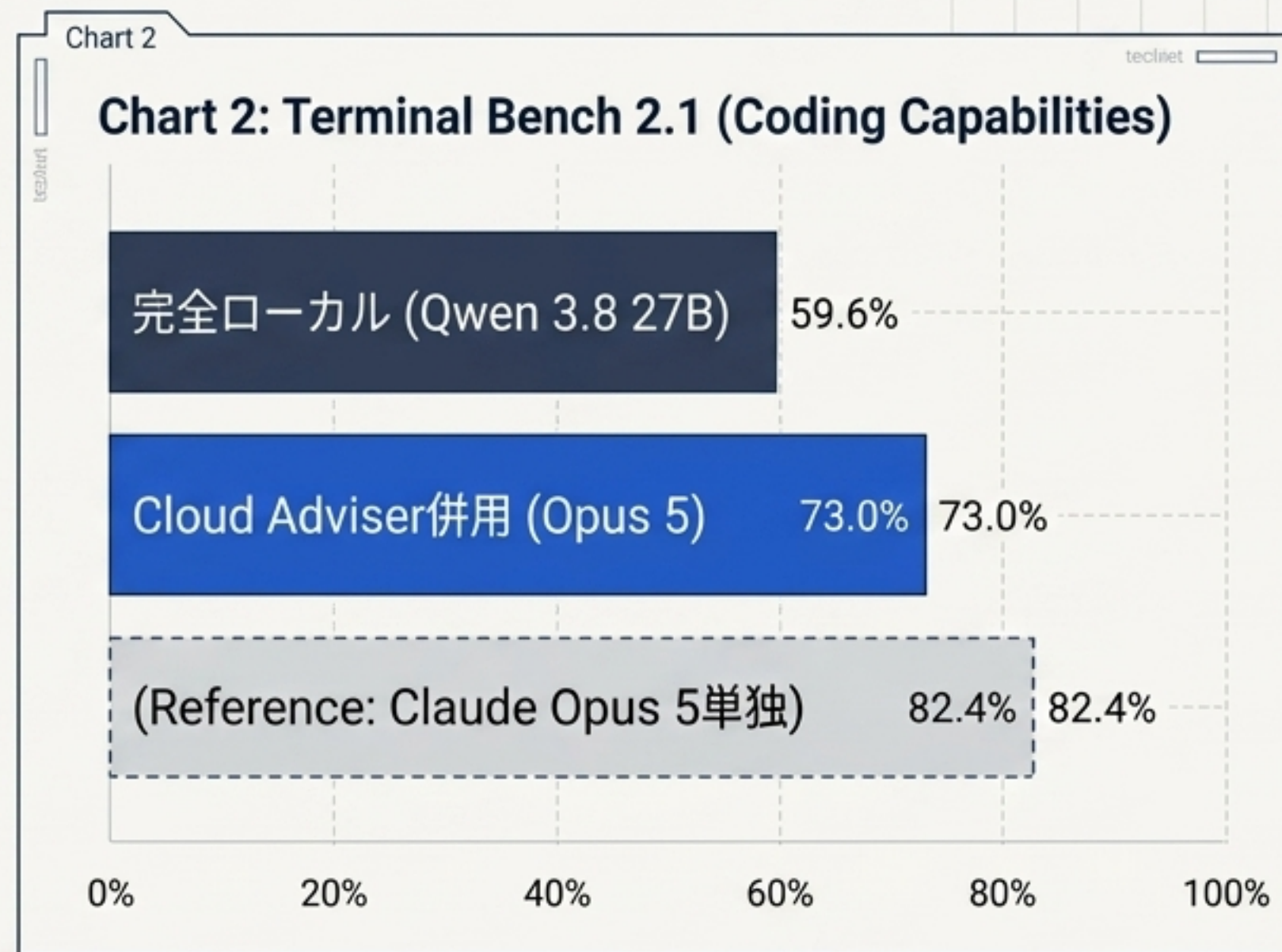
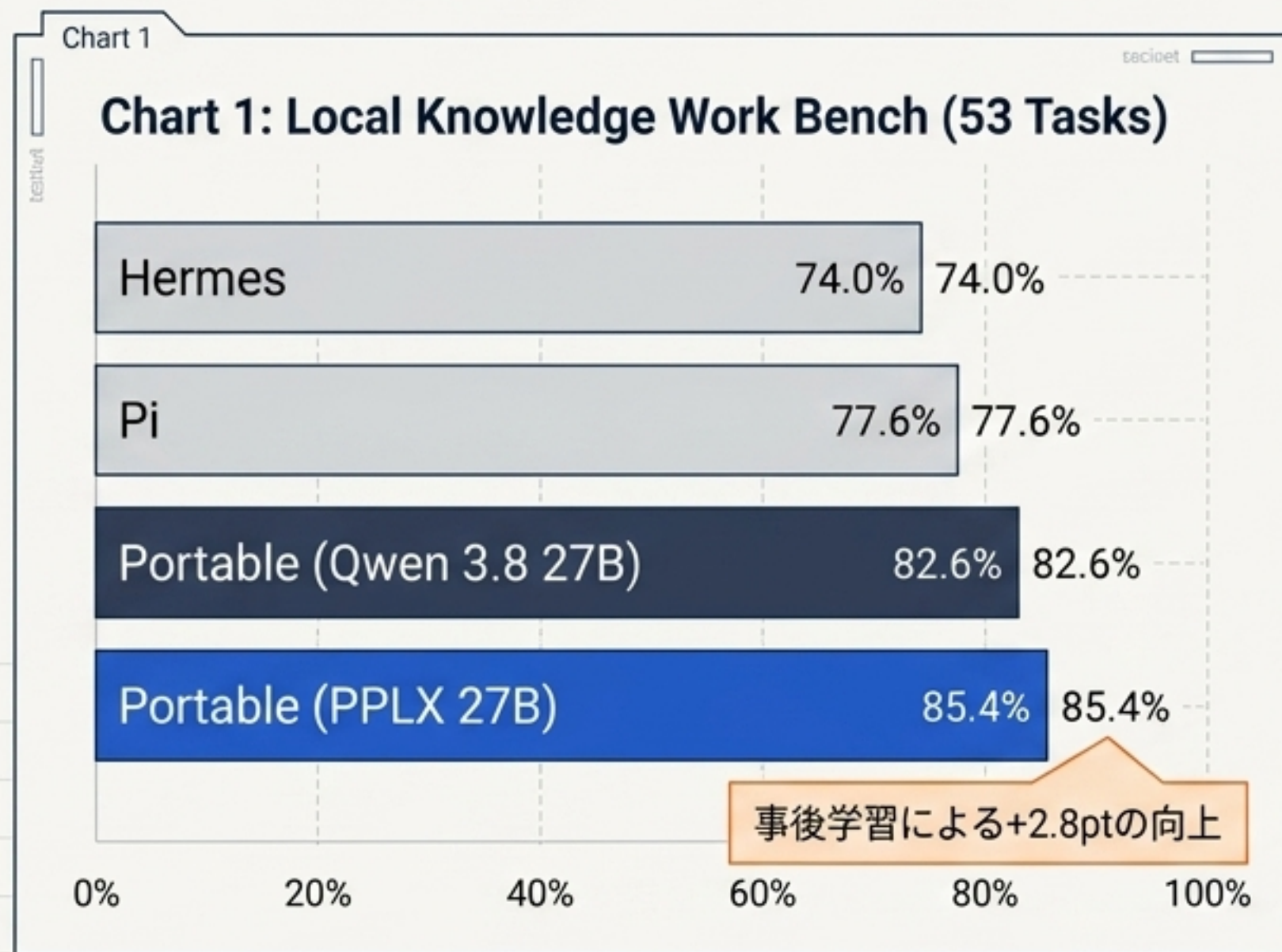
クラウド・エスカレーションの安全機構



Crucial Constraint

Cloud Adviserは「テキスト助言」を返すのみ。ローカルファイルやツールへの直接アクセス権限（直接操作）は一切持たない。

性能ベンチマークの現実：ハイブリッドの優位性

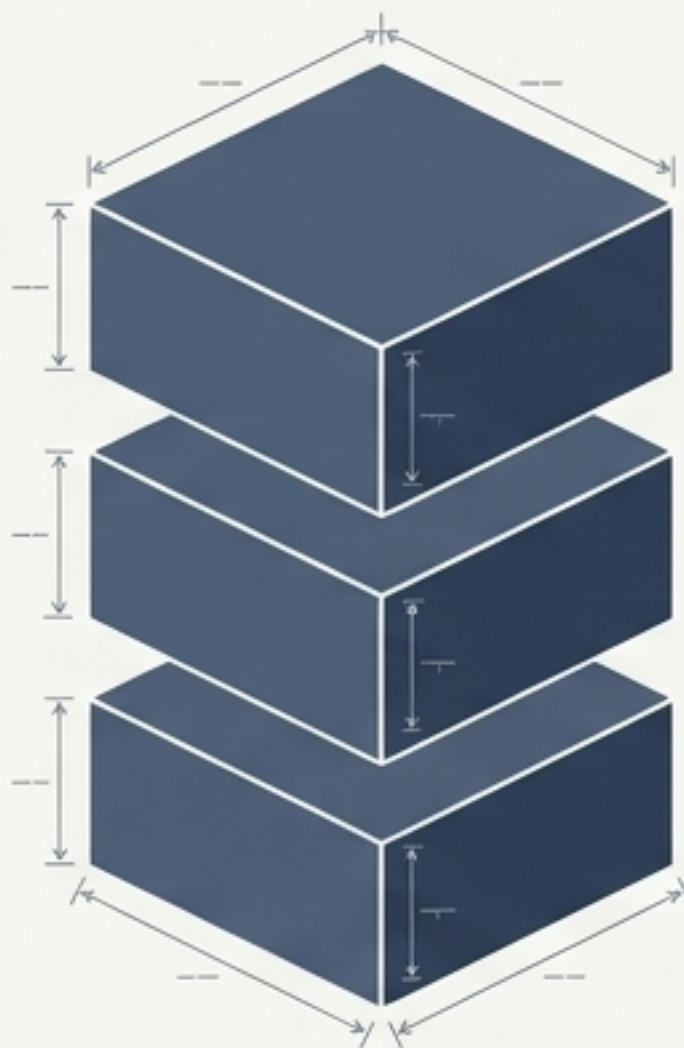


注意: 85.4%という数値はPerplexity内部の53タスク評価に基づく。同社は将来的なオープンソース化を予告しているが、現時点（2026年8月）で第三者による完全な独立検証は未完了。

TCO（総所有コスト）の真実：「ゼロトークンコスト」≠「無料」



CapEx + Fixed OpEx: Portable Computer Approach



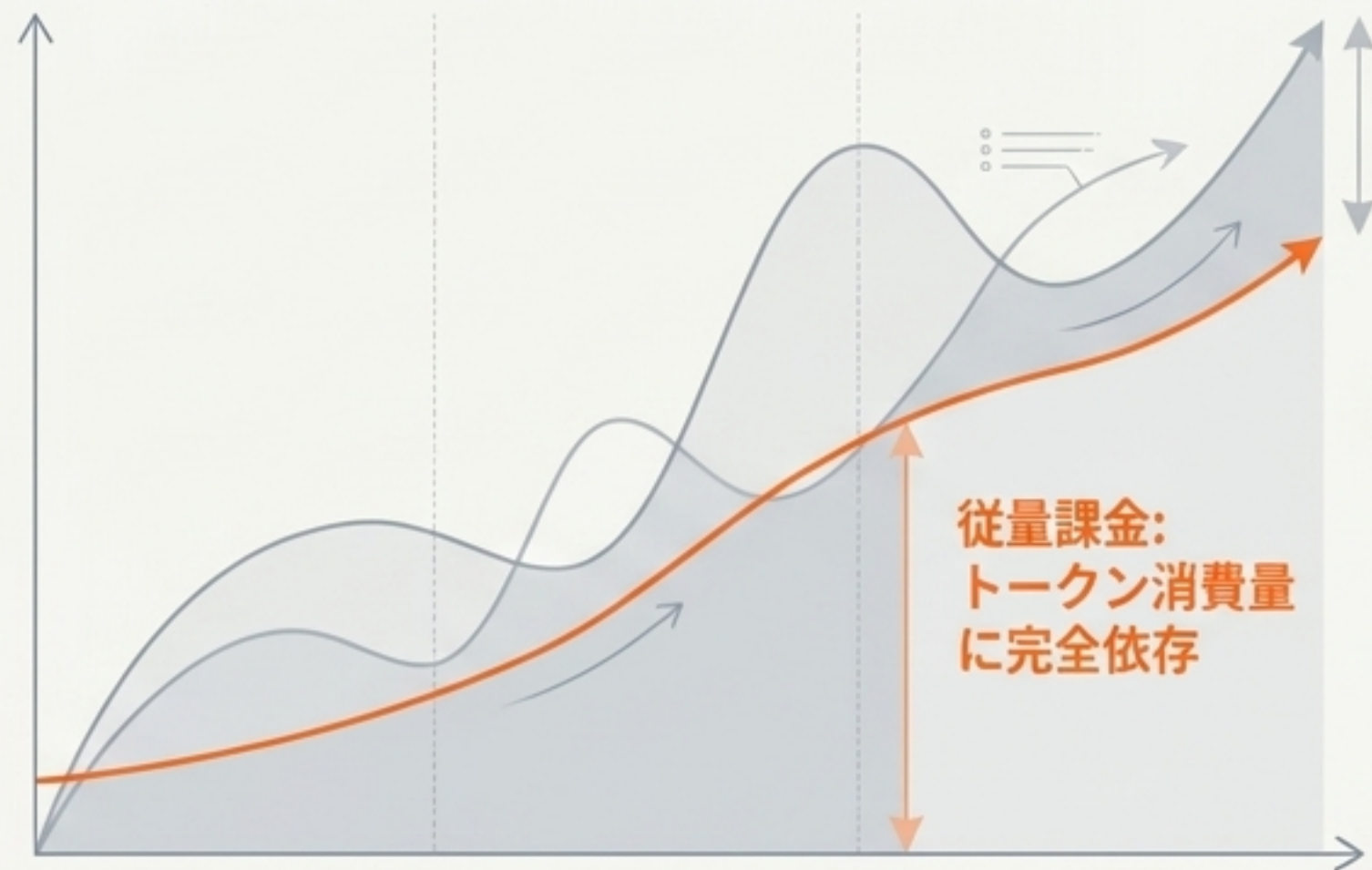
初期ハードウェア:
NVIDIA DGX Spark
(\$4,699 / 128GB Unified Memory)

必須サブスクリプション:
Perplexity Pro (\$20/月)
または Max (\$200/月)

隠れたコスト:
電力、GPU減価償却、
エンドポイント管理

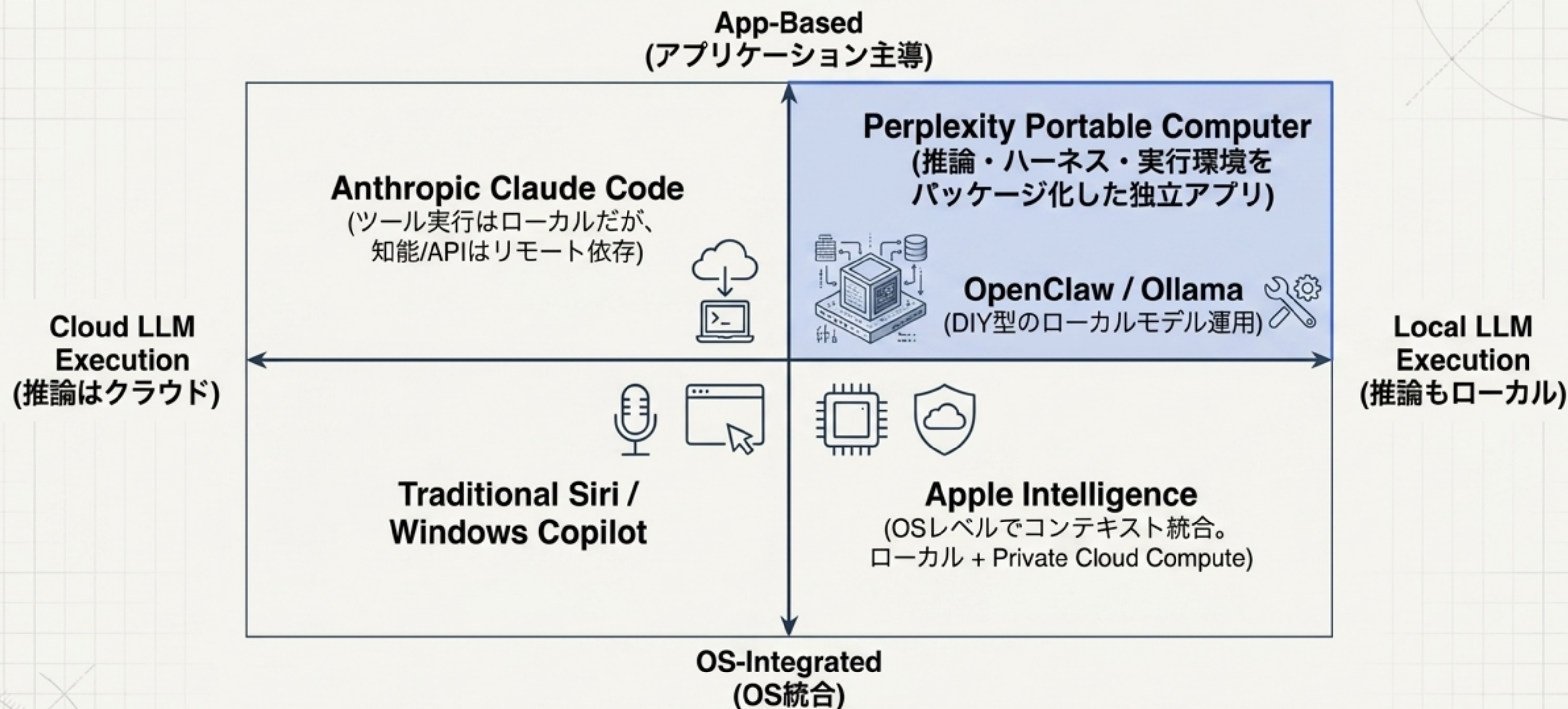


Variable OpEx: Traditional Cloud API



高稼働（24時間/365日）の知識労働エージェントとして利用するユーザーほどローカルの経済性が高まり、低稼働の場合はクラウドAPIの従量制が有利になる損益分岐が存在する。

エージェントAIの競争ランドスケープ

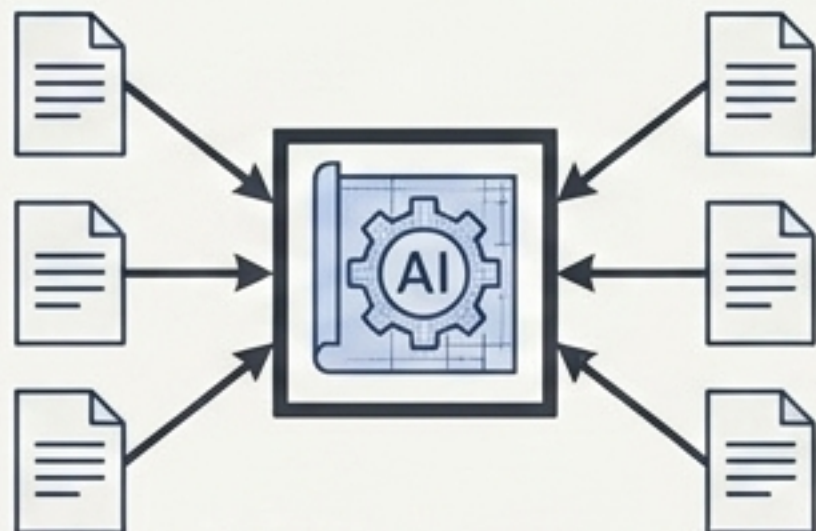


Insight: 競争軸は「モデルの賢さ」から、「どのレイヤーでエージェントを束ね、
どう権限を管理するか（知能の配置と統治）」へシフトしている。

ターゲットプロファイルと理想的なユースケース

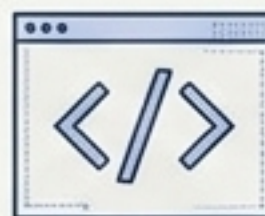


Card 1: Legal / Compliance
(法務・コンプライアンス)

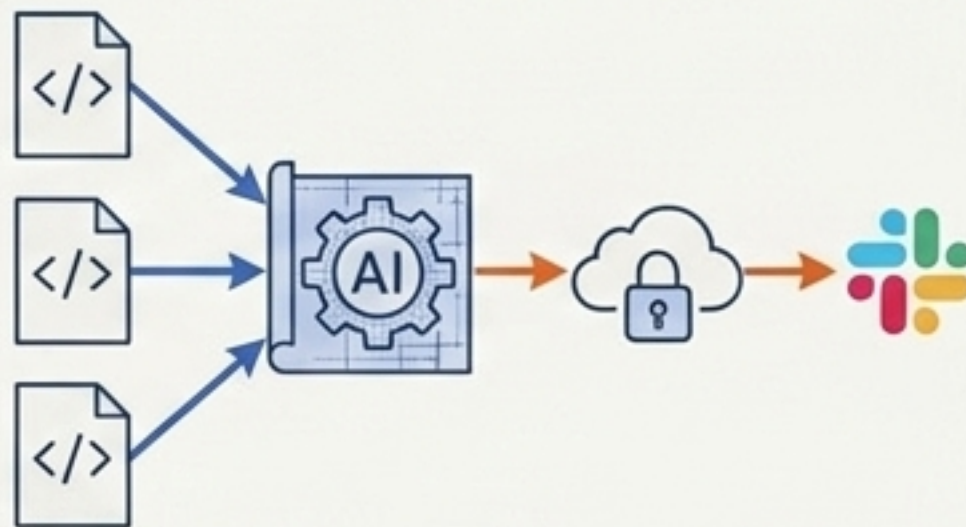


**Workflow: 機密PDF群の
完全オフライン横断分析**

Value: データ境界の維持。クラウドに一切
データを送信せず、契約書や監査資料の
矛盾点をローカルで抽出。



Card 2: Engineering / DevSecOps
(ソフトウェア開発)

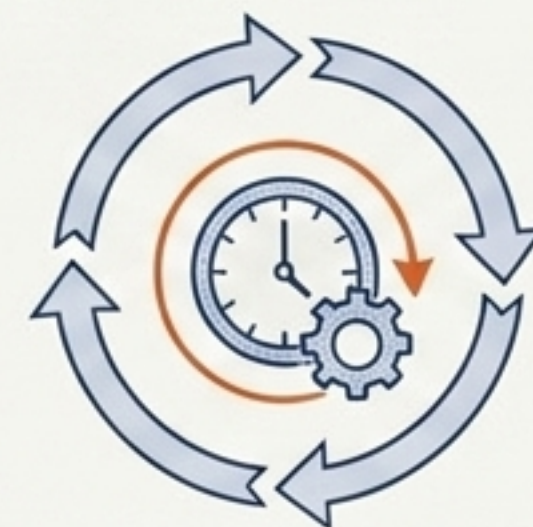


**Workflow: コードベース解析 →
クラウド助言 → Slack通知**

Value: 機密コードをローカルでトリアージ。
難解なバグのみPIIを除外してAdviserに相談
し、GitHub/Slackコネクタ経由で自動起票。



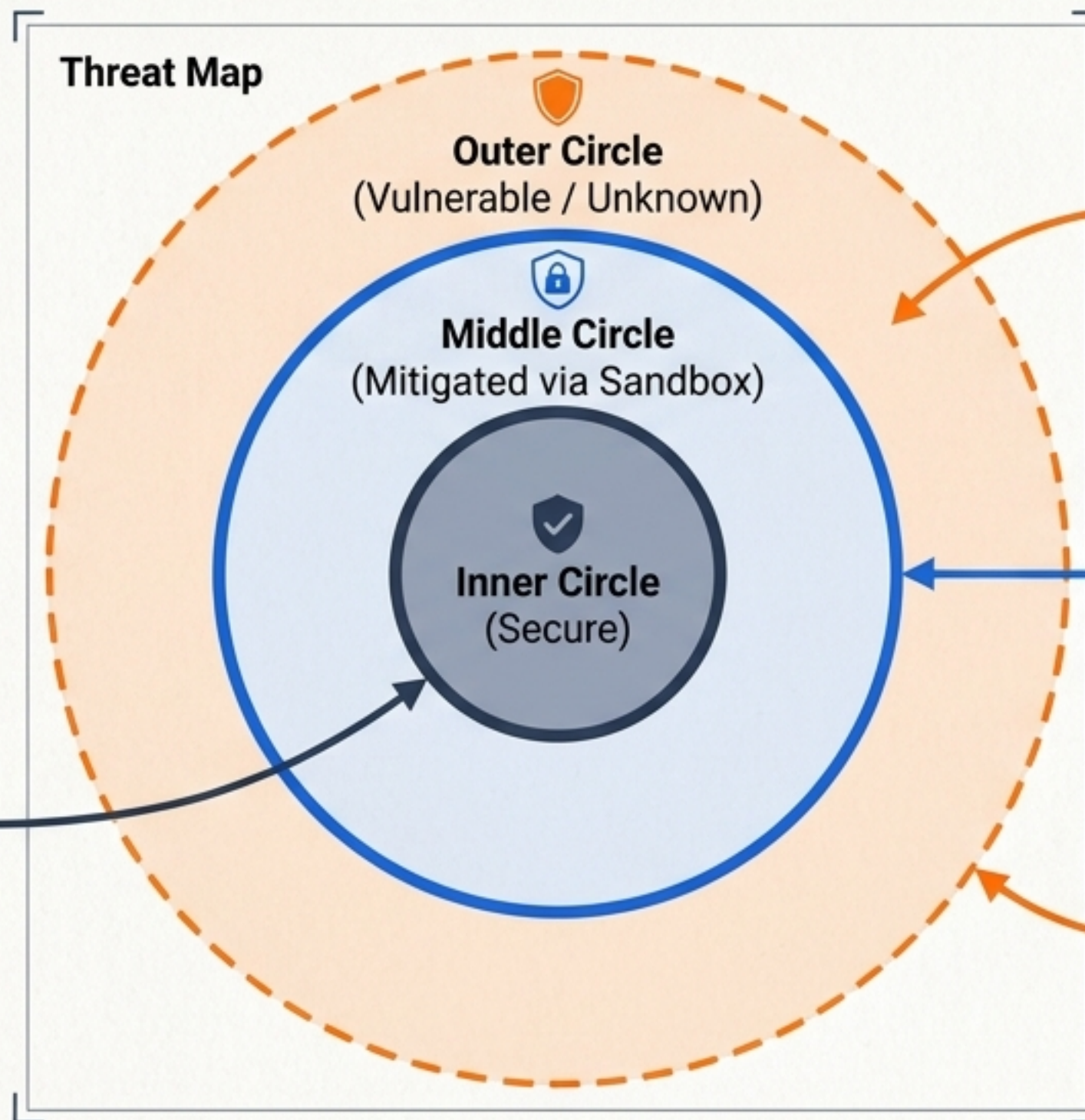
Card 3: High-Volume Knowledge Worker
(ヘビーリサーチャー)



**Workflow: 24時間稼働の
ディープリサーチ・エージェント**

Value: トークン課金を気にせず、ローカルの
財務データと公開Web検索 (Perplexity Search)
を組み合わせる連続的に調査タスクを実行。

エンタープライズ・セキュリティ：「ローカル」は免罪符ではない



データ処理の局所化。外部へのデータ流出（Data Exfiltration）の基本リスクは構造的に遮断。

PII分類器のFalse Negative（検知漏れ）によるクラウドへの誤送信。認証情報（OAuth）のローカル保管方式やTelemetryの不透明さ。

間接的プロンプトインジェクション（Indirect Prompt Injection）。悪意あるPDFを読み込んでも、OSレベルのSandboxがファイルシステムやネットワーク権限の暴走を抑制。

PII分類器のFalse Negative（検知漏れ）によるクラウドへの誤送信。認証情報（OAuth）のローカル保管方式やTelemetryの不透明さ。

Takeaway: 通常のコード実行基盤以上に厳しい脅威モデル（OWASP Agentic AI指針）のエンドポイント審査が不可欠。

コンプライアンスと法的要件：ローカル実行の限界

Japan (個人情報保護委員会 / PPC)

Rule: 生成AIサービス利用に関する注意喚起

Impact: ローカル処理は第三者提供を減らす
が、Gmail/DriveコネクタやCloud Adviserへ
データを流すワークフローを組む場合、利用
目的の特定や安全管理措置の確認義務は消滅
しない。

EU (AI Act / 人工知能法)

Rule: 2026年8月施行 Article 50 (透明性義務)

Impact: Portable Computerをローカルで動
かすこと自体がAI Actの包括的免除にはなら
ない。ユースケースとDeployerとしての役
割に応じたAI対話の透明性要件が適用され
る。

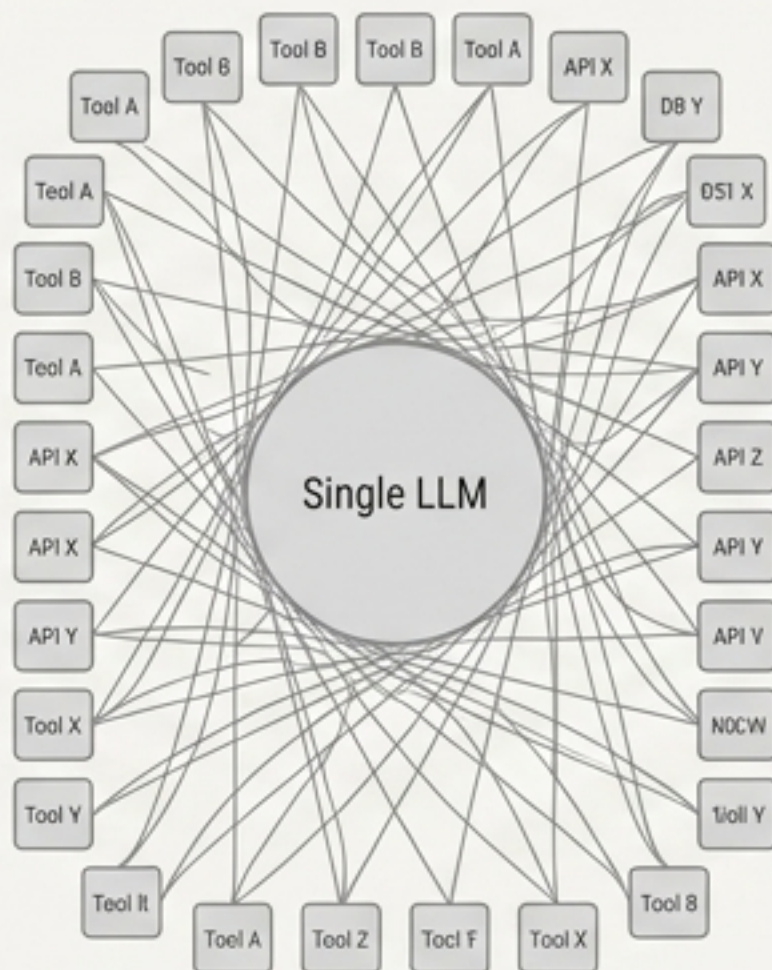
開発者へのパラダイムシフト：次世代エージェント・エンジニアリング

Old Skill (Before): MCPを繋ぐだけ

巨大なプロンプトに全ツール定義を詰め込む

モデルサイズ（パラメータ数）に智能を依存

単一のLLMが全プロセスを管理する設計

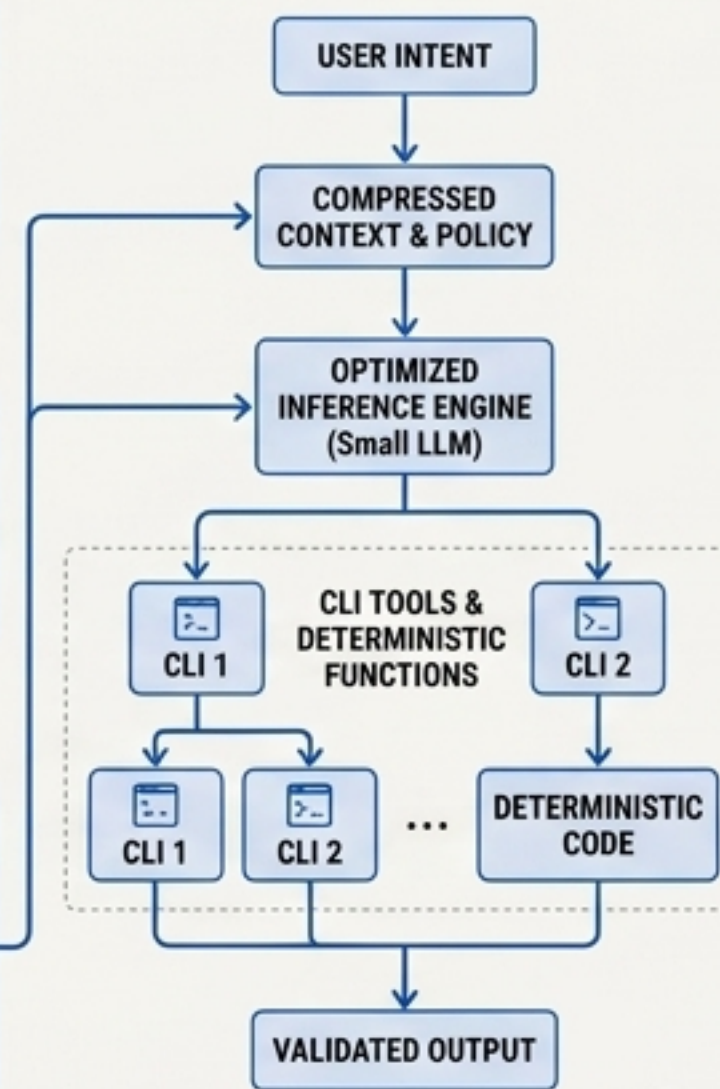


New Skill (After): コンテキストと権限の最適化

Context Compression:
限られたローカルウィンドウの節約技術

CLI Transformation:
巨大なMCPを極小のCLIツールへ変換・最適化

Deterministic Orchestration:
LLMを「推論エンジン」に降格させ、決定論的コードで実行ループを堅牢化



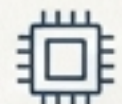
Insight: 競争力はモデルそのものから、「いかにローカルの制約内で効率的なハーネスを設計するか」に移行した。

市場インパクト：ハードウェア需要からコモディティ化へ

Short-Term (現在～直近): AI PCの「アプライアンス化」



AI実験用PCから、24時間稼働の「業務エージェント・アプライアンス」への進化。

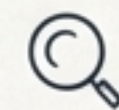


NVIDIA DGX Sparkの需要喚起。一般RTX/Windows環境への展開時期が真の普及転換点となる。

Mid-Term (中期的展望): エッジスタック のコモディティ化 vs. エコシステムの堀



OpenClawやOllamaの進化により、ローカル基盤自体はコモディティ化するリスク。



Perplexityの真の防御壁 (Moat) は「Search連携」「Connector UX」「高度に統合されたSandboxと事後学習モデル」にある。

戦略的総括：知能の配置と統治の時代

1. クラウド独占の終焉 (Intelligence Distribution)

「全トークンをクラウドで買う」時代は終わる。

固定資産（GPU）でベースの知能を所有し、クラウド推論をプレミアムサービスとして購入する新経済モデルが定着する。

2. モデルサイズよりエンジニアリング (The Triumph of the Harness)

勝敗を分けるのは巨大モデルへのアクセス権ではない。

決定論的オーケストレーター、コンテキスト圧縮、ツール最適化といった「エージェント・ハーネス工学」である。

3. セキュリティの幻想と現実 (The Security Reality)

「ローカル＝無条件に安全」は誤り。データ境界は強固になるが、エンドポイントの脅威（プロンプトインジェクション等）に対する厳格なポリシー運用と審査がエンタープライズ導入の鍵となる。