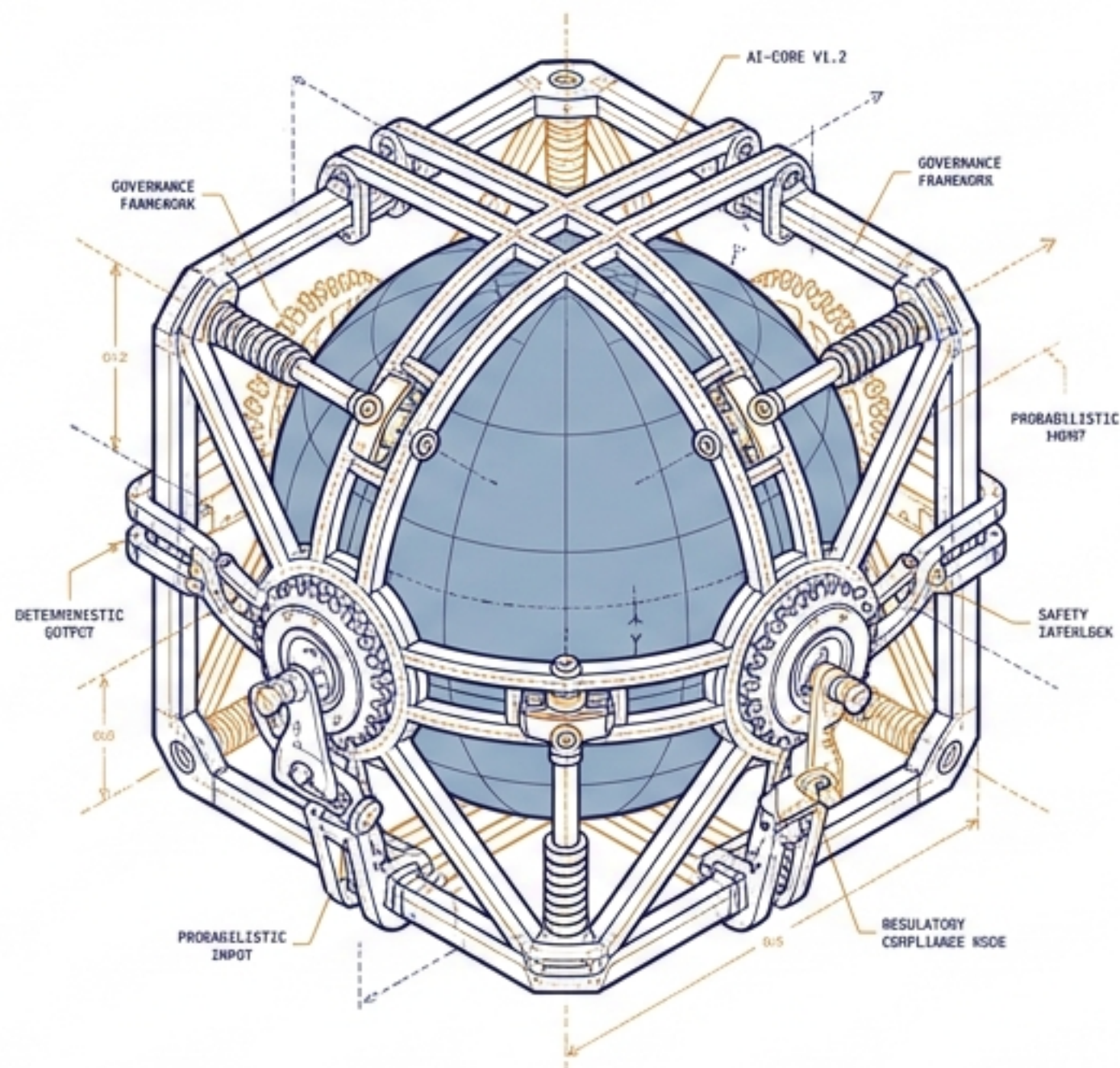


知財AIエージェントのハーネス設計とガバナンス・アーキテクチャ

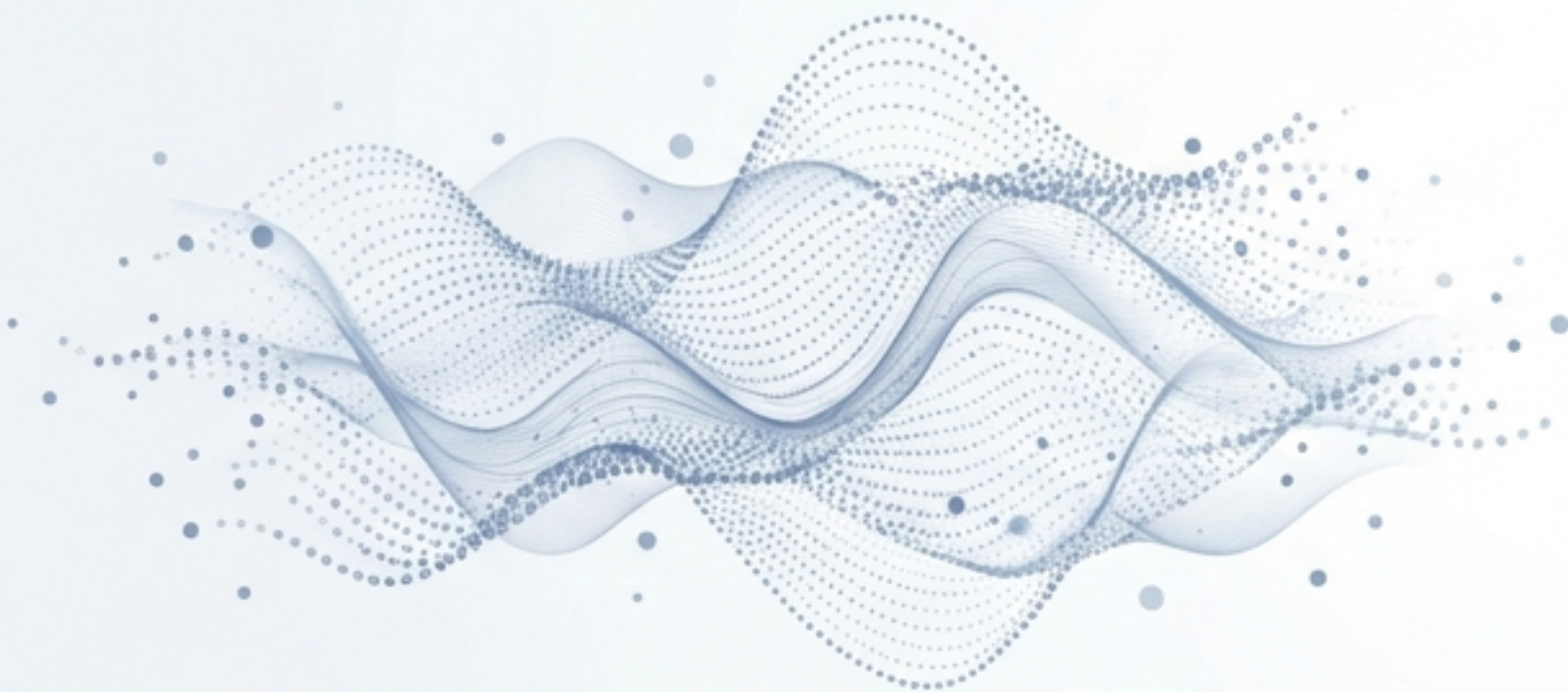
確率的推論を決定論的アクションへ変換する「精密工学」の青写真



Agent = Model + Harness

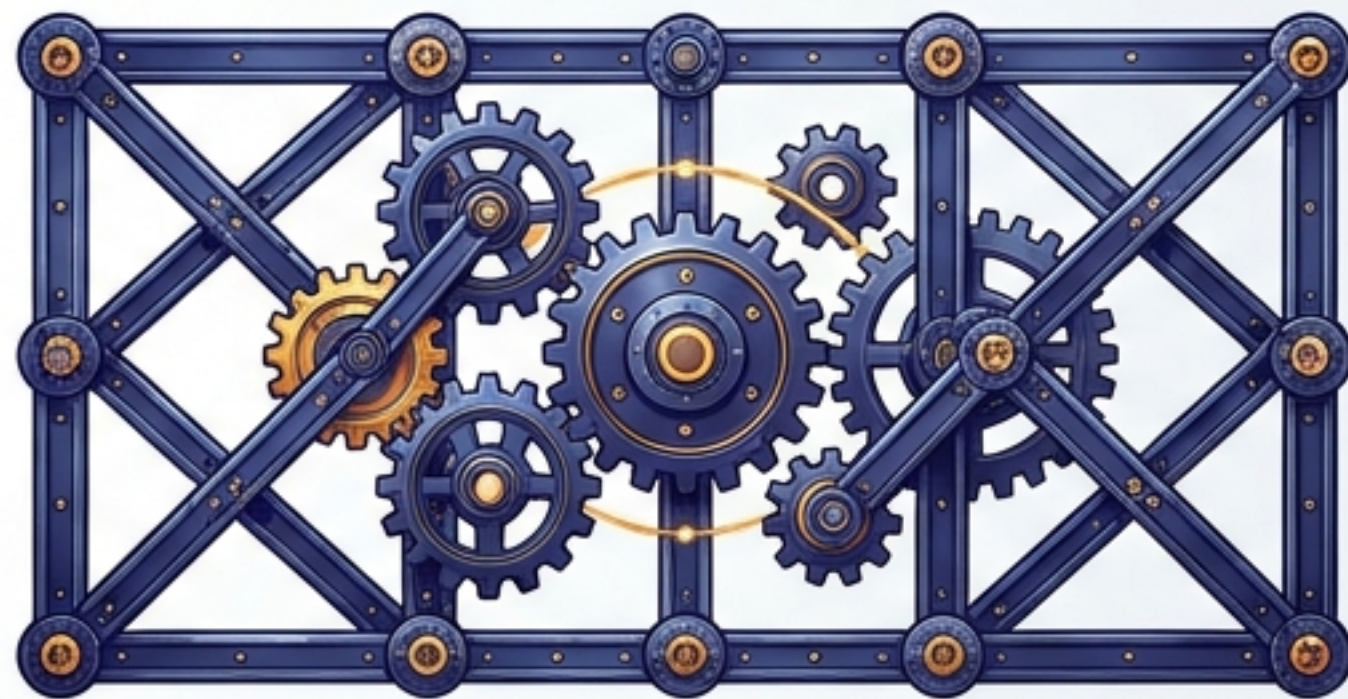
(Probabilistic Engine) (Cybernetic Governor)

Model (確率的推論エンジン)



本質的に確率的な推論エンジン。単体では特許法上の厳格な要件や善管注意義務を担保できない。

Harness (決定論的制御機構)



確率的推論を、再現性と法的整合性を備えた決定論的アクションへと変換するサイバネティクス制御器。

知財業務における成否は「モデルの規模」ではなく、「ハーネスの設計」で決まる。

外部API通信ゼロ。ローカルモデルが巨大フロンティアモデルを凌駕する

明細書・クレーム起案の総合品質スコア (0-100)

	総合品質	BERTScore	BERT-for-Patent
Patent-MAF (ローカル実行OSS + 役割分解ハーネス)	85.4	82.3	96.1
GPT-5 (クローズドモデル + マルチモーダル入力)	85.1	81.8	
Qwen3-VL-72B (オープンソース単体)	81.8	—	—

Patent-MAFは外部API通信ゼロ（ローカル実行）でありながら、商用フロンティアモデルと同等以上の推論能力を発揮。機密保持と性能のトレードオフは、ハーネス設計によって完全に打破できる。

知財法制が要求するハード・ガードレール

ガイドラインは「プロンプトの指示」ではなく「システムの強制」を要求する

新規性喪失と機密保持

第三者AIへのデータ送信(学習転用)による公知化・輸出管理規制違反リスク。



【システムの強制】

ゼロデータ保持APIの強制接続、完全ローカルLLMへのフォールバック、実行リージョンの固定。

善管注意義務と署名責任

架空判例の引用、未確認IDSの提出による懲戒処分リスク(37 CFR 11.18(b))。

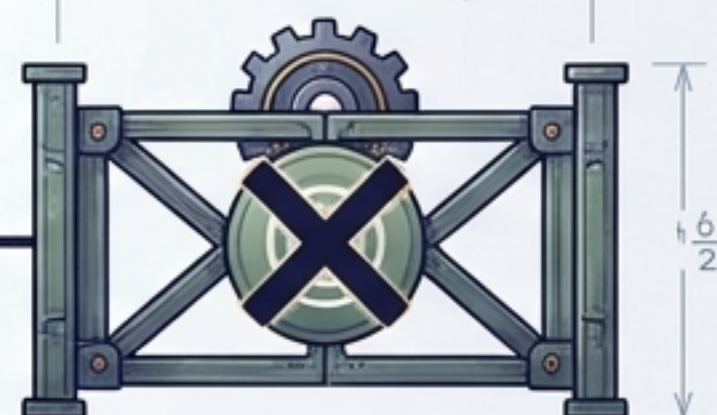


【システムの強制】

庁DB原文と突合する「決定論的検証フック」、起案者と承認者を物理的に分離したHITL承認ゲート。

庁ITシステムの保護

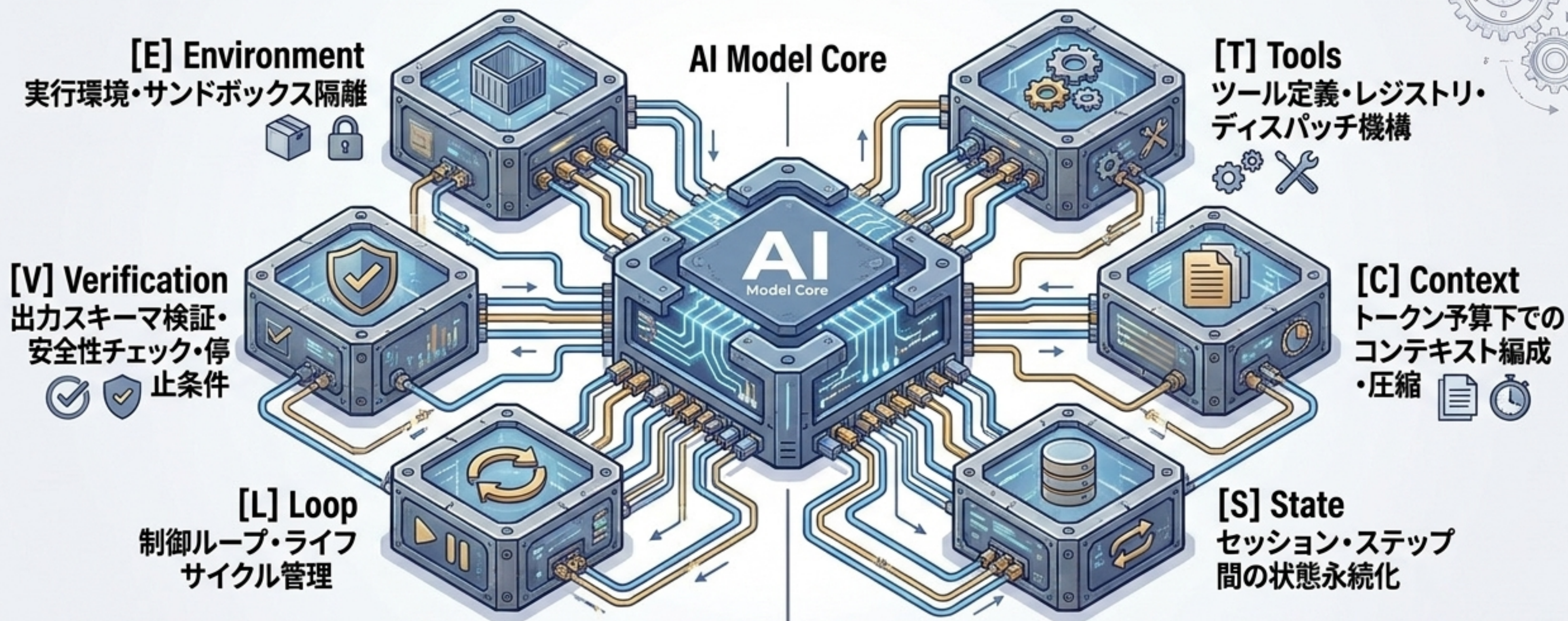
USPTO等のシステムへの自律的アクセスによるアカウント剥奪や法的制裁リスク。



【システムの強制】

エージェントからの直接提出(Filing)機能の完全排除、厳格なレートリミット監視。

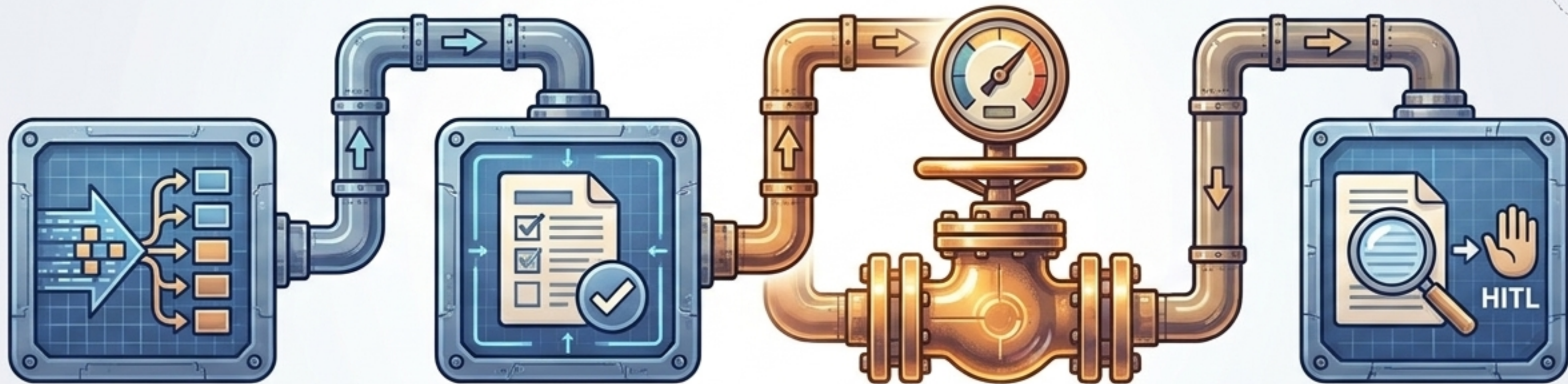
ハーネスの6要素解剖図 (Anatomy of a Cybernetic Harness)



拘束条件仮説(Binding Constraint Thesis): モデルの重みを固定したまま、この6要素のスケャ
フォールディングを変更するだけで、実務タスクの解決率が20%以上変動する。

Blueprint 1: 先行技術調査

コンテキスト劣化とAPI枯渇を防ぐデータパイプライン



Micro-batching

10万件の母集団を数件単位のマイクロバッチに分割し、長大コンテキストによる公報混同を防ぐ。

JSON Schema Validation

公報番号・対比結果・判定理由を定型フォーマットで強制出力。

Quota Control Valve

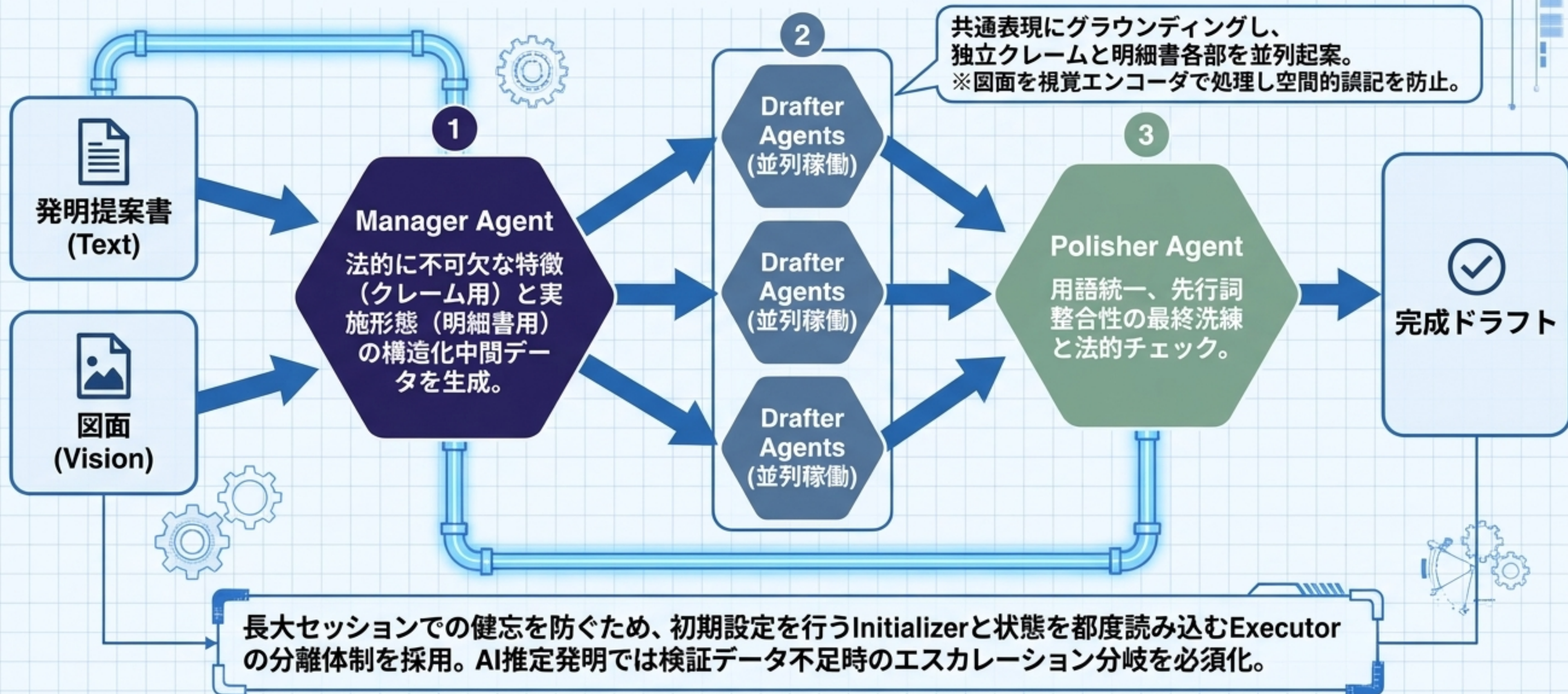
J-PlatPat (1日200回) やEPO OPSのAPIレスポンスヘッダを監視。キャッシュ優先とMax-Step上限による強制スロットリング。

Deterministic Audit

検索式・日付・ヒット件数のステート記録と、人間によるサンプリング監査 (HITL) への連携。

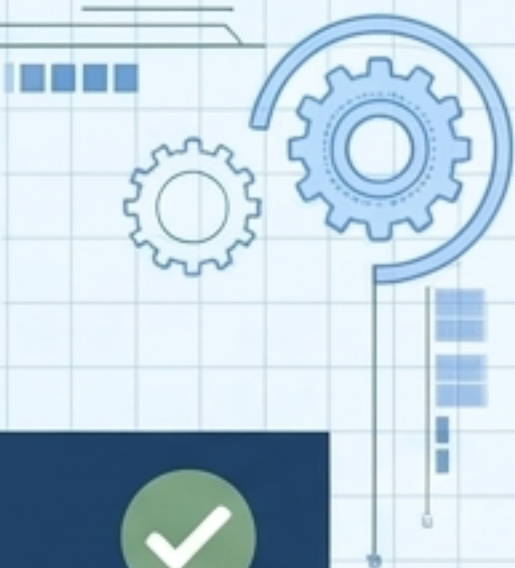
Blueprint 2: 明細書・クレーム起案

Patent-MAF型 マルチエージェントDAGとマルチモーダル統合



Blueprint 3: OA応答と5つの失敗モードの阻止

LLMの致命的欠陥に対する決定論的な処方箋



Failure Mode (症状) 	Deterministic Fix (処方箋) 
架空引用 (Fake Quotations)	引例原文 (OCR/XML) との完全一致文字列検索の強制
未クレーム特徴による反論	主張する相違点とクレーム構成要件の「1対1マッピングテーブル」の状態固定
新規事項の追加 (New Matter)	当初明細書内の段落番号・原文との直接的サポートリンクの必須化
拒絶理由の読み違い	審査官の拒絶ロジックの構造化マッピング出力
様式違反 (37 CFR 1.121等)	LLMを排除し、決定論的テンプレートエンジンによる下線・取消線の自動生成



二重検証ゲート (The Double Verification Gate)

決定論的スクリプトと確率論的AIの統合

第一層: 決定論的パテント・リンター (Deterministic Linters)

構文木ベースでの先行詞チェック、図面符号の一致、クレーム番号体系の整合性をコードによるハード検証で判定。

Pass (通過)

第二層: 確率論的 Judge (Probabilistic QA Judge)

第一層を通過したドラフトのみを処理。発明の本質的特徴の網羅性、設計回避への堅牢性、適法性をLLMによる意味検証で反復改訂 (平均+1.85pt品質向上)。



自動評価スコアの向上は、直ちに法的有効性の向上を意味しない。
記号・数値の厳密な一致はLLMに任せず、決定論的なコードで検査せよ。

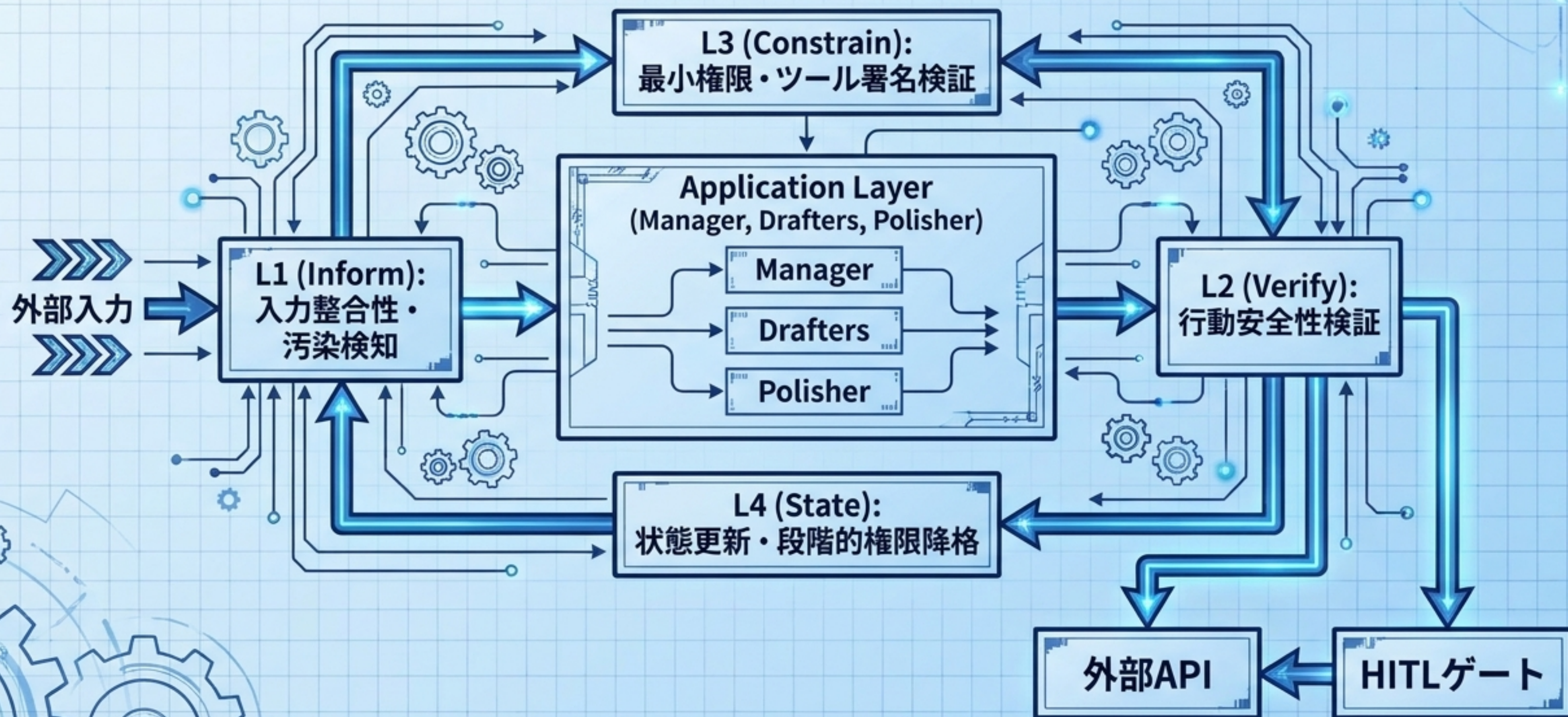
従来型セキュリティの限界

外付けガードレール (Bolt-on) の構造的欠陥



ライフサイクル統合型セキュリティ: SafeHarness

UBRを約38%、ASRを約42%低減する層間フィードバック型アーキテクチャ



L4 適応的権限降格 (Graceful Privilege Degradation)

システムをクラッシュさせず、セキュリティと業務継続性を両立する動的メモリ保護



Recovery Window: 安全な操作の連続により段階的に復帰 (Gradual recovery by continuous safe operations)

導入へのロードマップ (The Path Forward)

Step 1: 単一業務のモジュールスタートとDAG化

いきなり自律エージェントを全社展開しない。構成要素対比等の定型業務に絞り、固定ワークフロー (DAG) として検証・監査する。

Step 2: ローカル・マルチエージェント基盤の構築

外部API依存からの脱却。Patent-MAF型アーキテクチャを採用し、起案プロセスを役割分解 (マネージャー・ドラフター・ポリッシャー) する。

Step 3: 監査証跡の完全統合 (Observability)

AI生成プロセスと人間の承認プロセス (HITL) を分離。善管注意義務に対応する、再構築可能なイミュータブル・ログを確立する。

**ハーネス設計は単なるラッパー開発ではない。
確率的AIと決定論的知財実務を架橋する基幹インフラの構築である。**