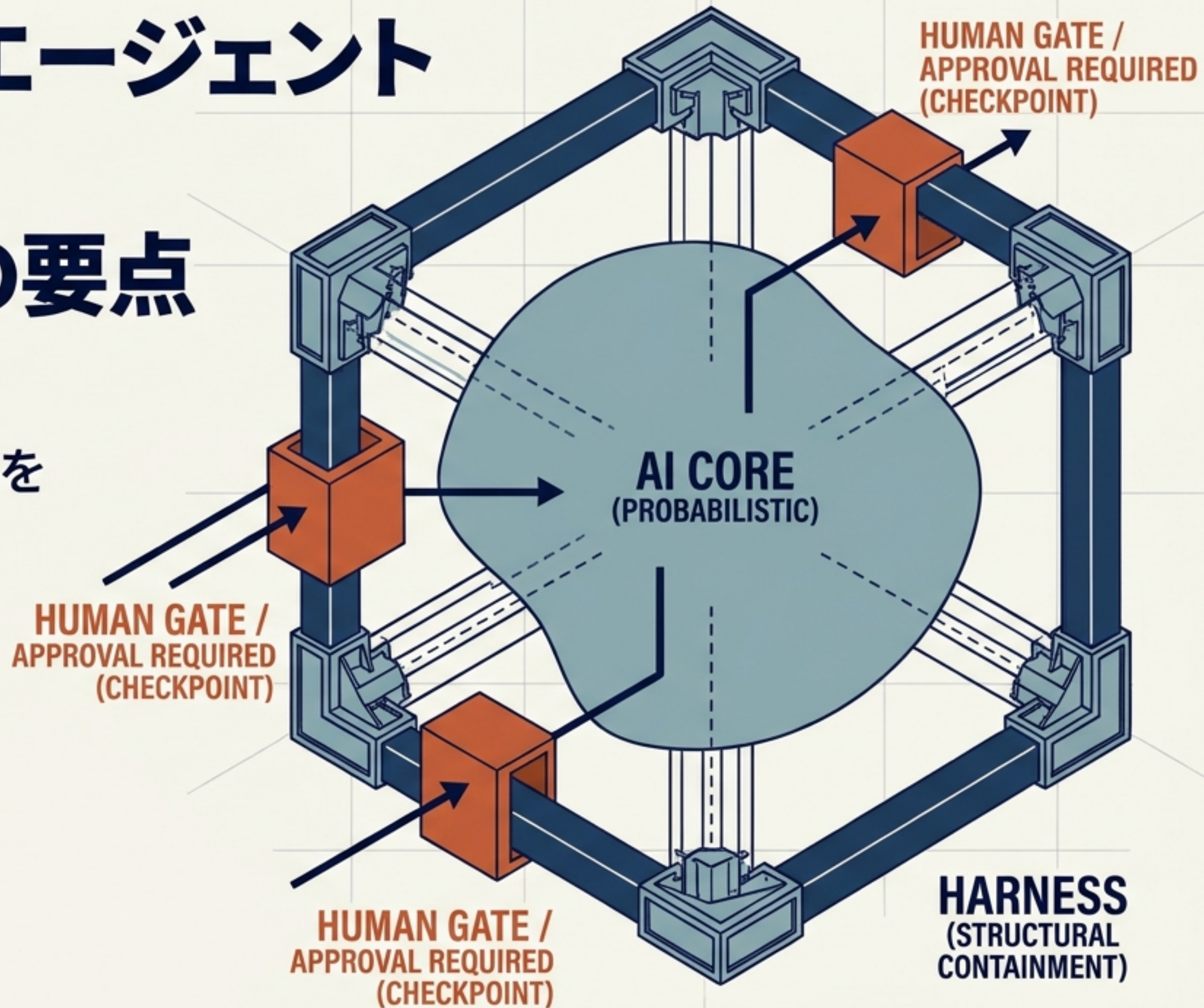


知財業務へのAIエージェント 適用における 「ハーネス設計」の要点

機密境界・出典検証・承認ゲートを
「構造」で担保する
実践的システムアーキテクチャ



プロンプトは「意図」を形づくるが、「許可範囲」は変えられない

モデル層に任せる領域(確率的誘導)

・クレーム解釈の観点や論理展開

・調査の網羅基準の提示

・文章のトーンや生成品質の調整

【限界】指示をすり抜ける可能性が常に残る

ハーネス層で強制する領域(決定論的統制)

・未公開発明の送信先と機密区分の統制

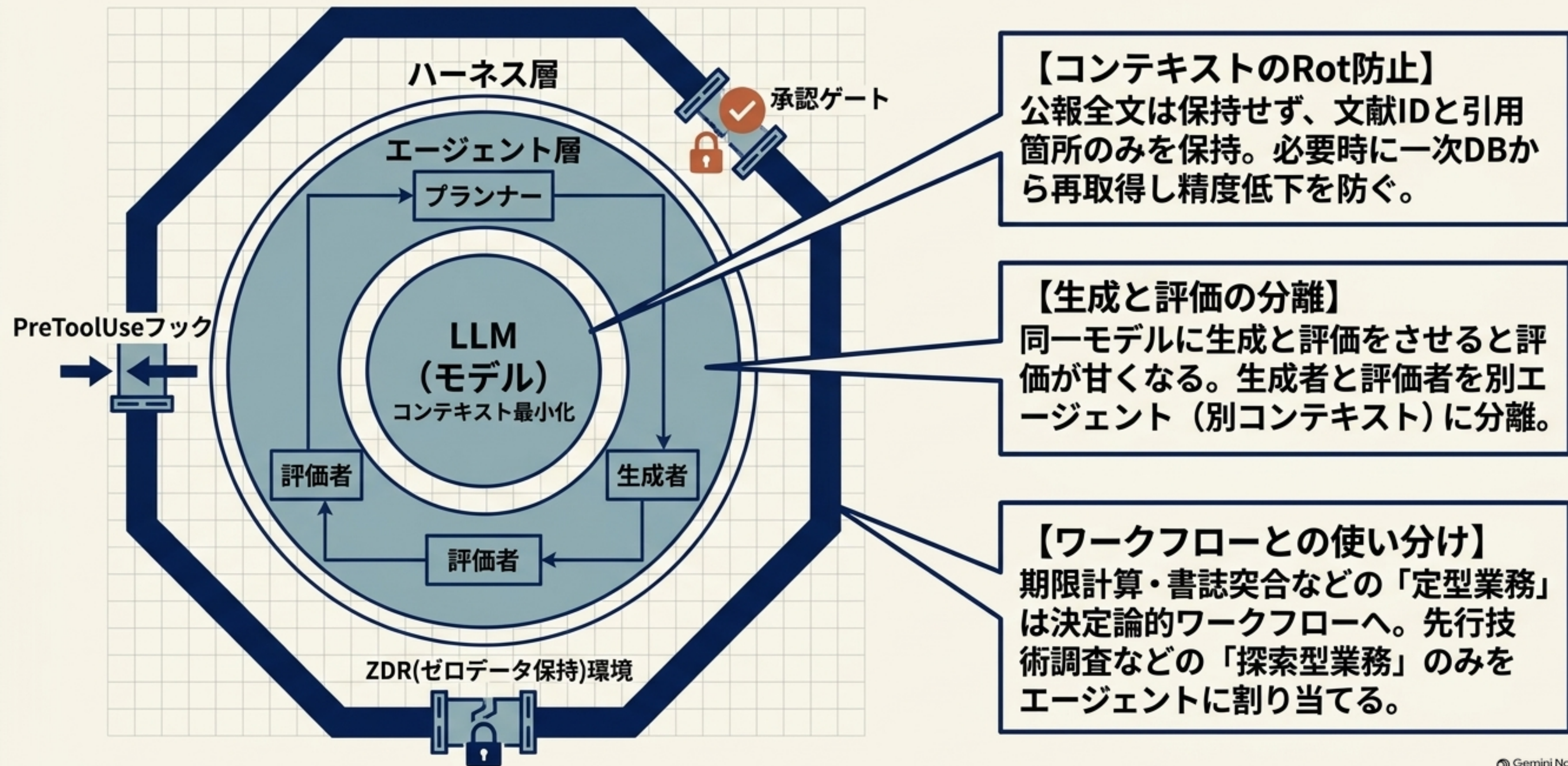
・出願・応答・外部送信などの不可逆操作

・特許庁APIや商用DBのアクセス頻度制限

【原則】守らなければ事故になる要件はすべてここで実装する

Anthropic社 セキュリティ原則 (2025-2026): 「システムプロンプトはエージェントの行動目標を定義するが、何が許可されるかの制約はOSレベルのサンドボックスやPreToolUseフックによる多層防御で担保しなければならない」

知財特化型AIエージェントの推奨アーキテクチャ構成



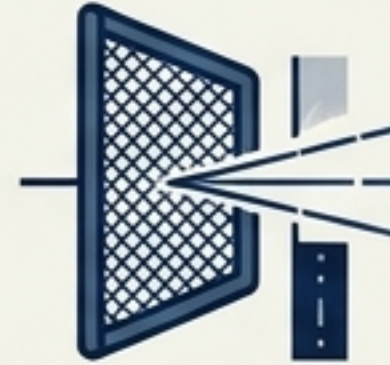
システム構造として解決すべき「4つの決定論的防衛線」

01 要件①：機密境界の統制 (Confidentiality)



- 防御対象: 未公開発明の漏洩
• 新規性喪失
- 構造的解決: ZDR契約ルーティングと機密語マスキング

02 要件②：出典の機械照合 (Accuracy)



- 防御対象: 存在しない架空特許・判例の提出（ハルシネーション）
- 構造的解決: 決定論的検証パイプラインの構築

03 要件③：外部データの統制 (Data & Legal)



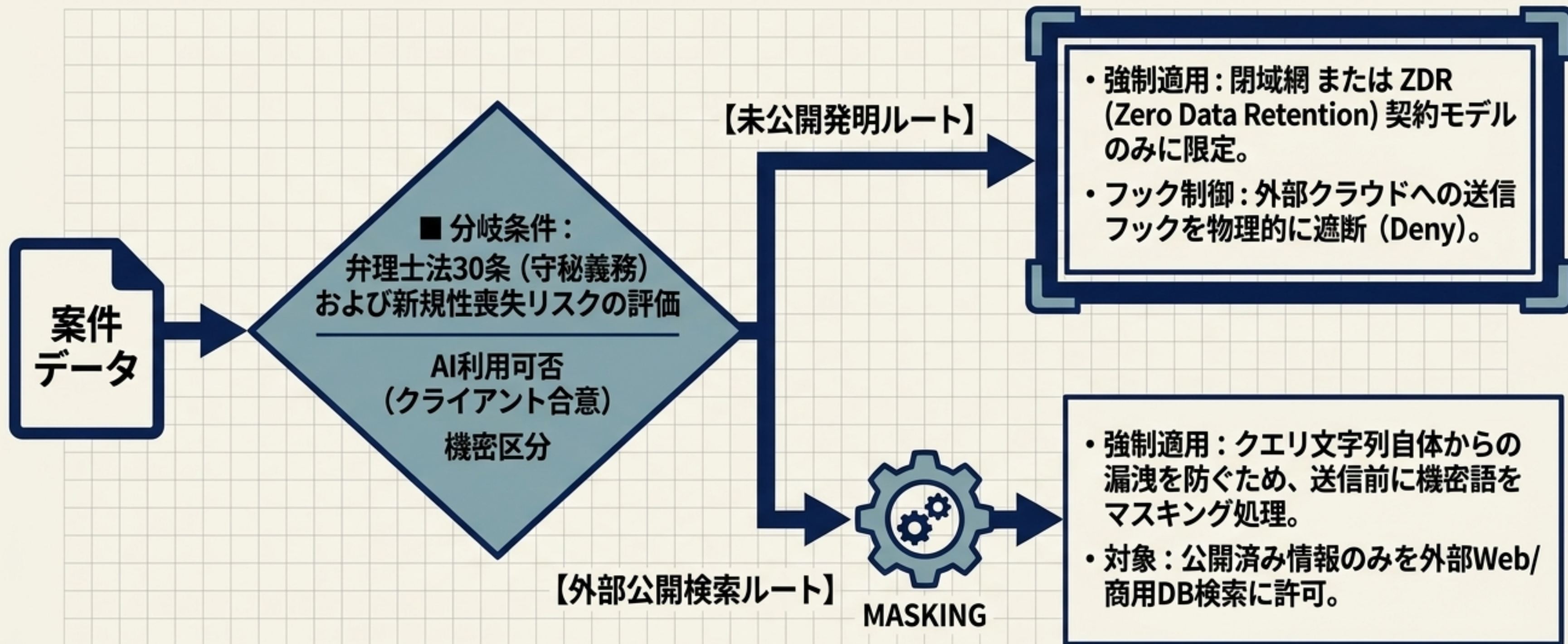
- 防御対象: DB規約違反（スクレイピング禁止）とプロンプトインジェクション
- 構造的解決: 正規APIの利用と読取り専用サブエージェントへの隔離

04 要件④：承認ゲートと責任分界 (Human Gate)



- 防御対象: 非弁行為（弁理士法75条）リスクと不可逆な誤操作
- 構造的解決: L0～L3のリスク別エスカレーションマトリクス

要件① 機密境界：案件メタデータに基づく強制的ルーティング



【設計の示唆】「クラウドかオンプレか」の二者択一ではなく、「データ分類 × 契約条件（学習不使用・保持期間） × 監査可能性」の組み合わせでアーキテクチャを決定する。

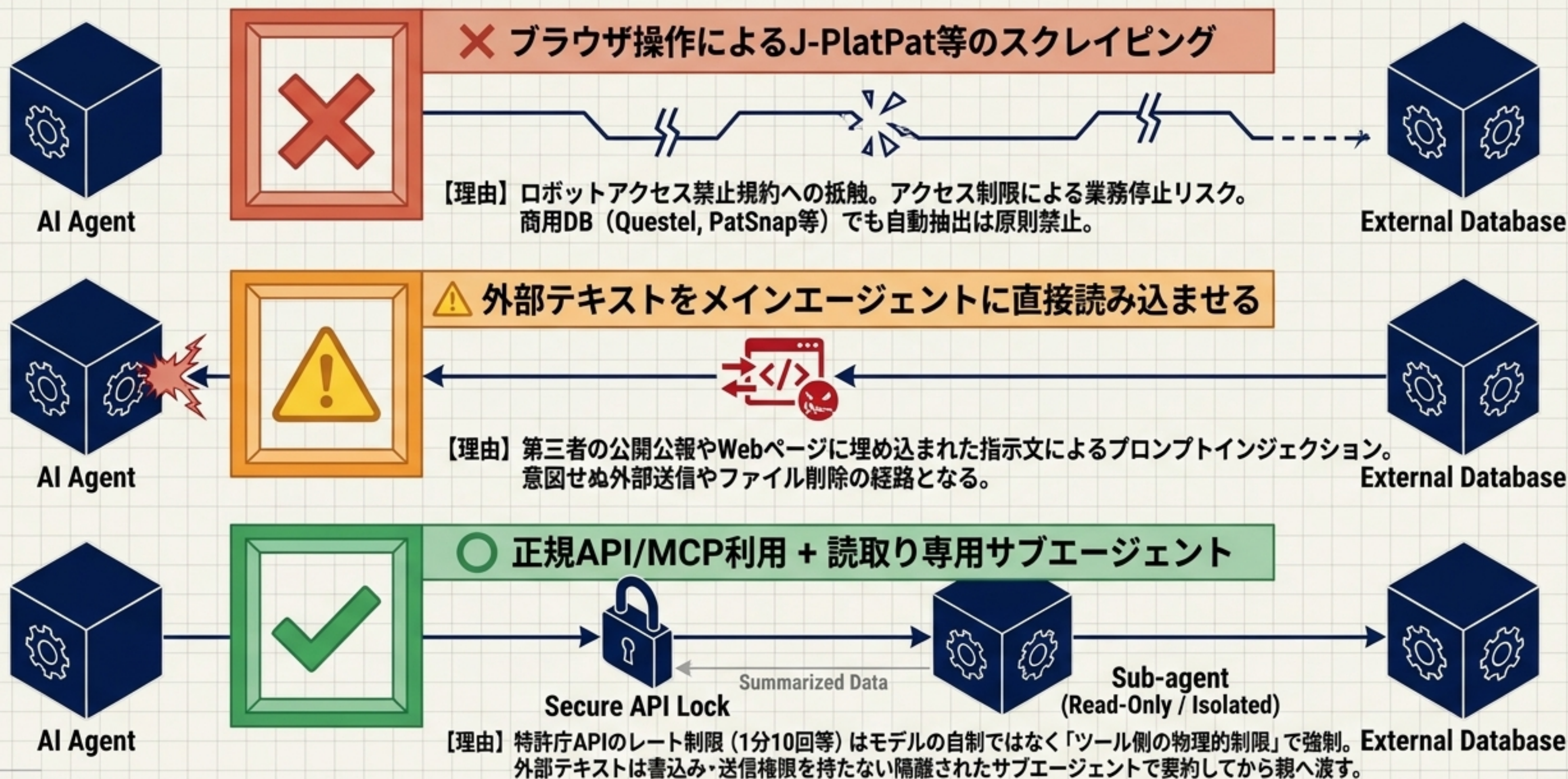
要件② 出典検証：

「AIによるAIのレビュー」を過信しない検証パイプライン

■ 背景：USPTO初のAI誤引用懲戒事例（2026年8月 In re Brian E. Mitchell）
起案とは別のAIでレビューさせたが架空の内部証拠引用を見抜けず戒告処分。

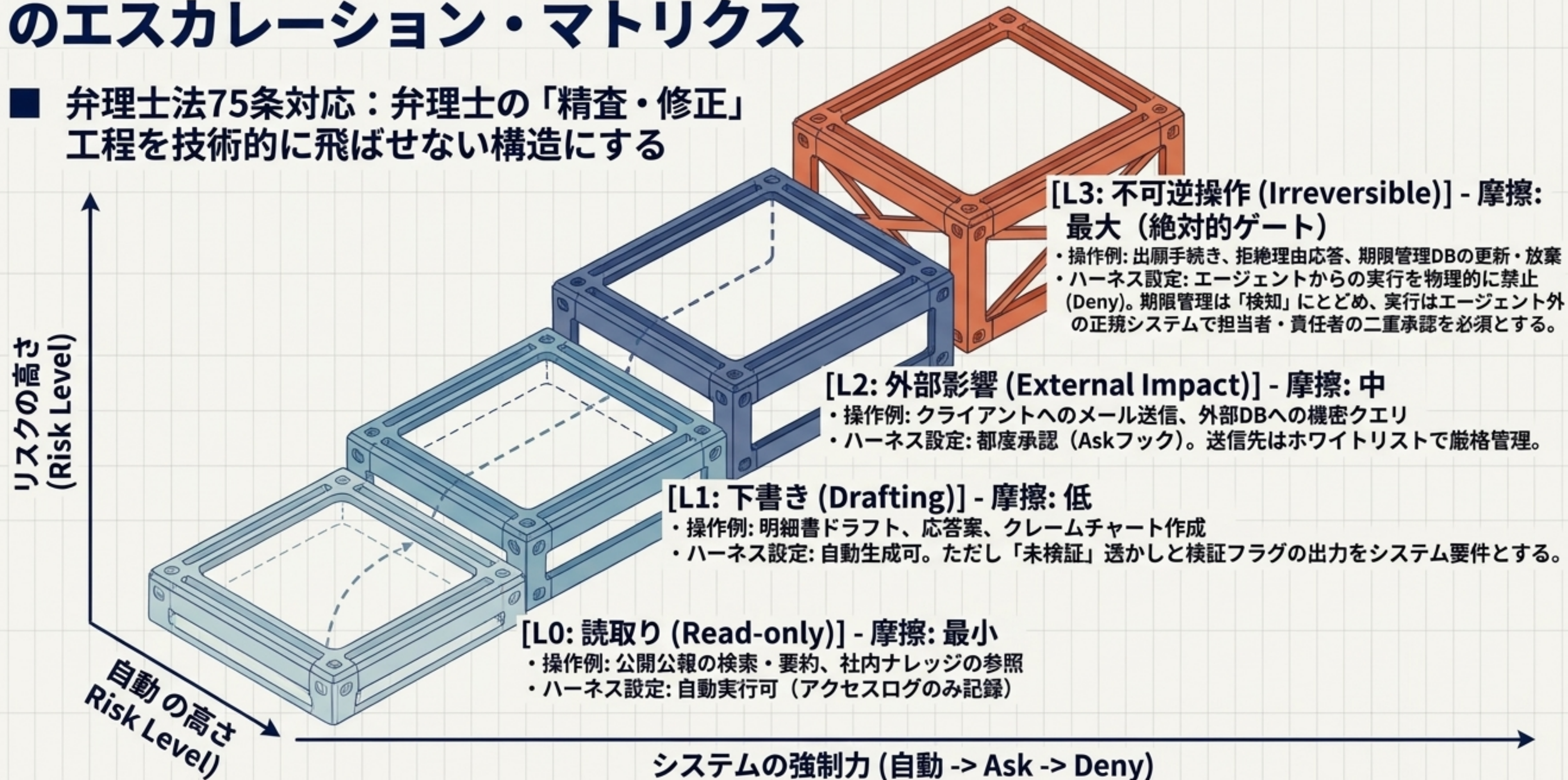


要件③ 外部ツール・データ連携：規約違反とインジェクションを構造で防ぐ



要件④ 人間の関与：リスクに応じた4階層のエスカレーション・マトリクス

- 弁理士法75条対応：弁理士の「精査・修正」工程を技術的に飛ばせない構造にする



監査とトレーサビリティ：立証資産と開示リスクのトレードオフ

■ ログは「残せば良い」わけではない。保存範囲とアクセス権限を法務と事前設計する。

防衛のための立証資産 (Proof Assets)

【FTO・鑑定業務における資産】

- ・ 網羅性の証明: 検索式、母集団、スクリーニング判断、採否理由を再現可能な形で保存。
- ・ 秘匿特権の要件: 「弁護士・弁理士の指示の下でAIを使用した」という明確な証拠の記録。

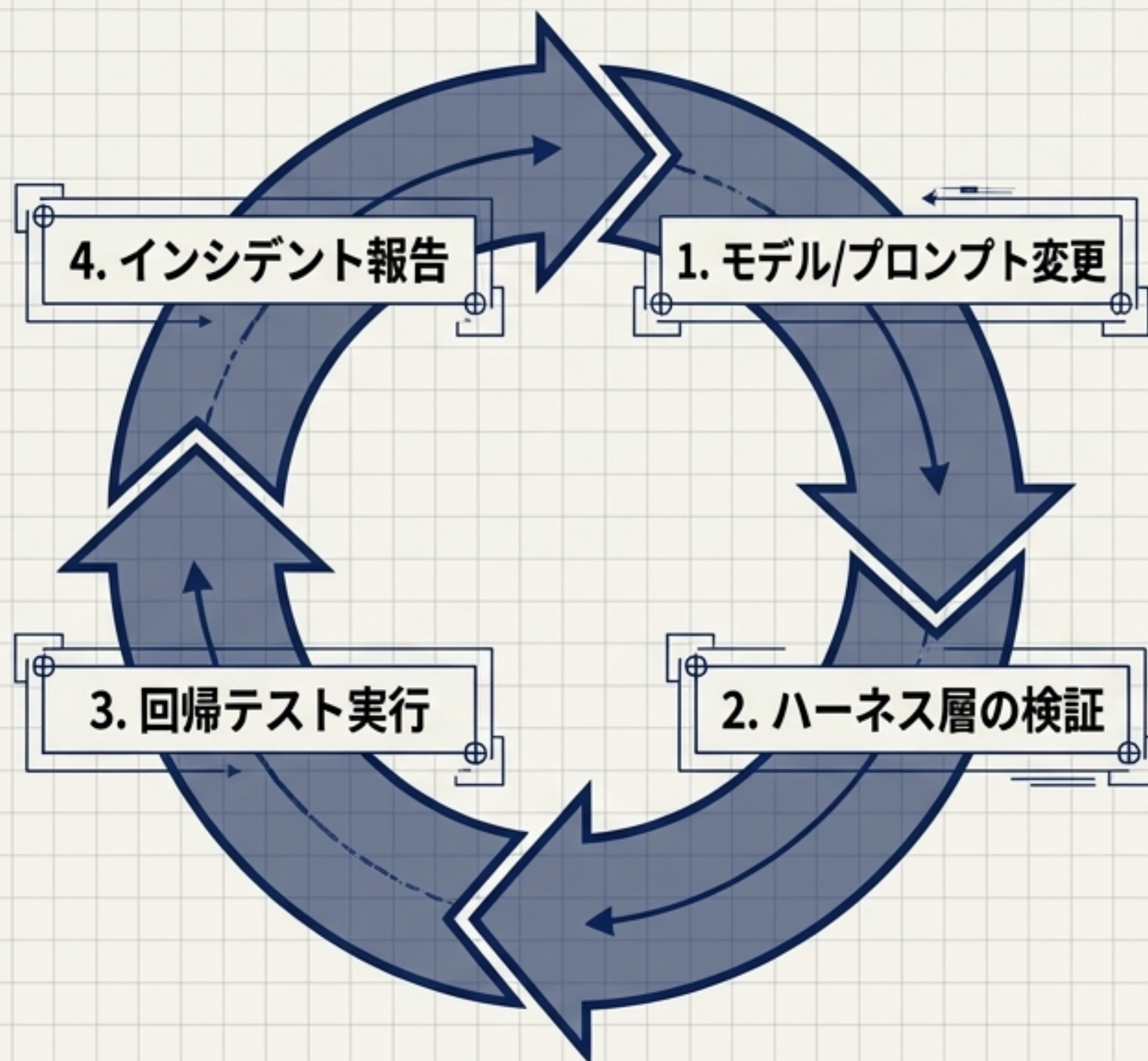
訴訟時の開示リスク (Discovery Risks)

【米国訴訟等でのリスク】

- ・ Heppner事件 (2026年2月 米国地裁): 弁護士の指示なく一般消費者向けAIを使った記録について、秘匿特権もワークプロダクト保護も否定された事例。
- ・ ディスカバリー対象化: エンタープライズ契約の有無、指示記録の欠如が致命傷になり得る。庁側AI (USPTO ASAP!) の検索結果が包袋に残る以上、出願人側の把握内容との差分管理が誠実義務の観点で問われる。

評価 (Eval) と継続的改善：公開ベンチマークは足切りに過ぎない

■ 評価対象は「AIモデル単体」ではなく、「モデル+ハーネスのセット」である。



1. 自社ゴールデンセットの構築

公開ベンチマーク (PatRe等) への過学習を避け、過去の自社・事務所案件から「正解引例 (審査官引用+社内後日発見)」付きの独自テストセットを構築する。

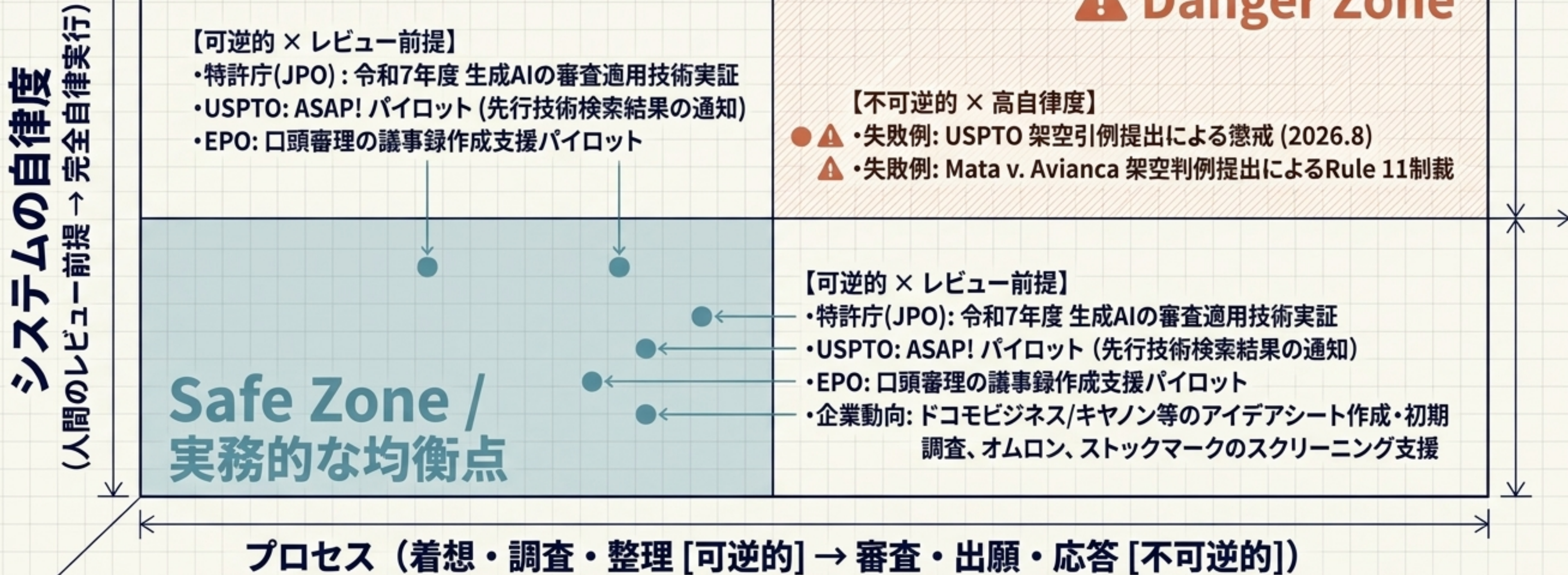
2. 適合率 (Precision) より再現率 (Recall)

調査業務においては、ノイズの多さよりも「致命的な見落としがないこと」を絶対的な主指標とする。

3. $\text{pass}@k$ ではなく pass^k を求める

k 回中1回成功する確率 ($\text{pass}@k$) ではなく、実務品質として「 k 回すべて合格 (pass^k)」するかを測り、出力のばらつきを極限まで抑え込む。

2025-2026年 実務適用ランドスケープと現在地



(※出願・応答の提出までを完全自律実行する公表事例は現時点では存在しない)

実務向け導入チェックリスト (ローンチ前ゲートキーパー)

■ A. 設計方針

- ☐ 業務を「ワークフロー(固定)」と「エージェント(探索)」に仕分けたか
- ☐ 「プロンプト(指示)」と「権限・フック(強制)」を分離したか

■ B. 機密・データ / C. 正確性

- ☐ 案件メタデータに基づく強制的ルーティング(ZDR/閉域)を実装したか
- ☐ 外部検索クエリのマスキング処理を組み込んだか
- ☐ 特許番号・判例の一次DB照合と、未検証出力のブロック機能があるか

■ D. 外部データ / E. 人間の関与

- ☐ J-PlatPat等のスクレイピングを禁止し、正規APIに限定しているか
- ☐ L3操作(出願・応答・放棄)をエージェント実行不可(Deny)としたか
- ☐ 弁理士の「精査・修正」工程を技術的にスキップできない構造か

■ F. 監査 / G. 評価

- ☐ FTO調査の証跡(検索式・母集団・判断)を再現可能な形で保存しているか
- ☐ 自社ゴールデンセットを用い、ツール変更のたびに回帰評価を回しているか



**優れた知財AIエージェントは、
卓越したモデル単体では成立しない。**

**「制約（ハーネス）」を精緻に設計できた企業だけが、
知財業務の革新を、安全かつスケラブルに実現できる。**