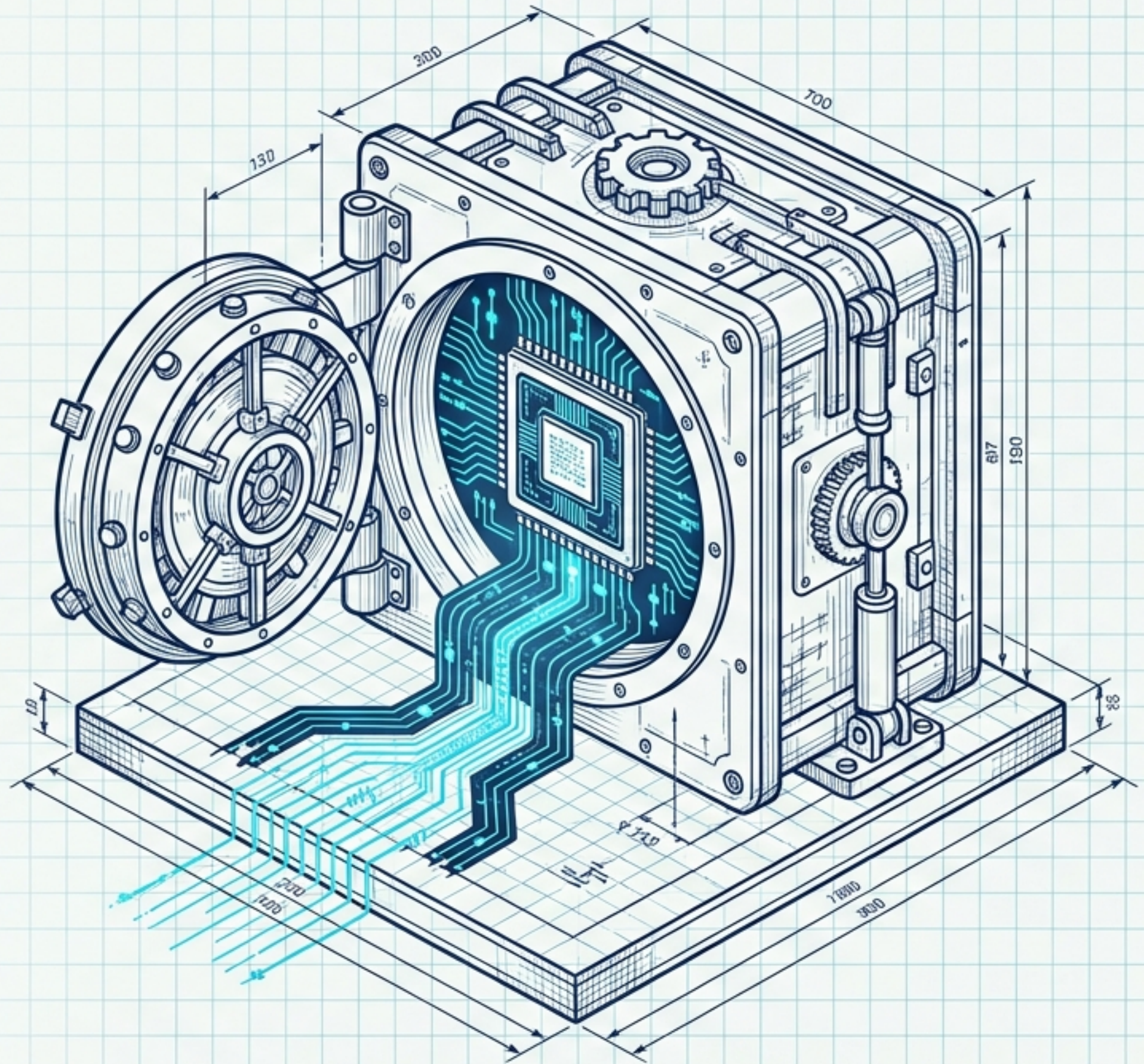


RTX Spark環境に おける知財業務向け ローカルLLM構築の 最適解

128GB統合メモリを制する
「国産LLM」の選定戦略と、
セキュアなRAGアーキテクチャ設計図



Executive Summary: 意思決定の要旨

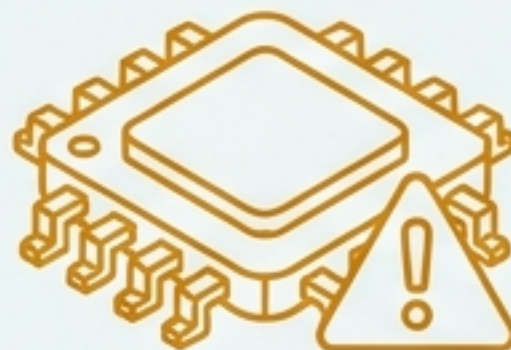


【結論】最適解は「三段構え」

主軸はバランスに優れる「Qwen3-Swallow-32B (4-bit)」。

難案件の精査・第二意見には「Stockmark-2-100B」。

安定稼働と完全フルスクラッチの安心感に「CALM3-22B」。



【制約】128GBの現実

統合メモリ128GBは有限。OS、KV Cache、RAGが競合する。

172Bモデルは実用外。複数並行処理のスweetspotは22B～32Bクラスに存在する。



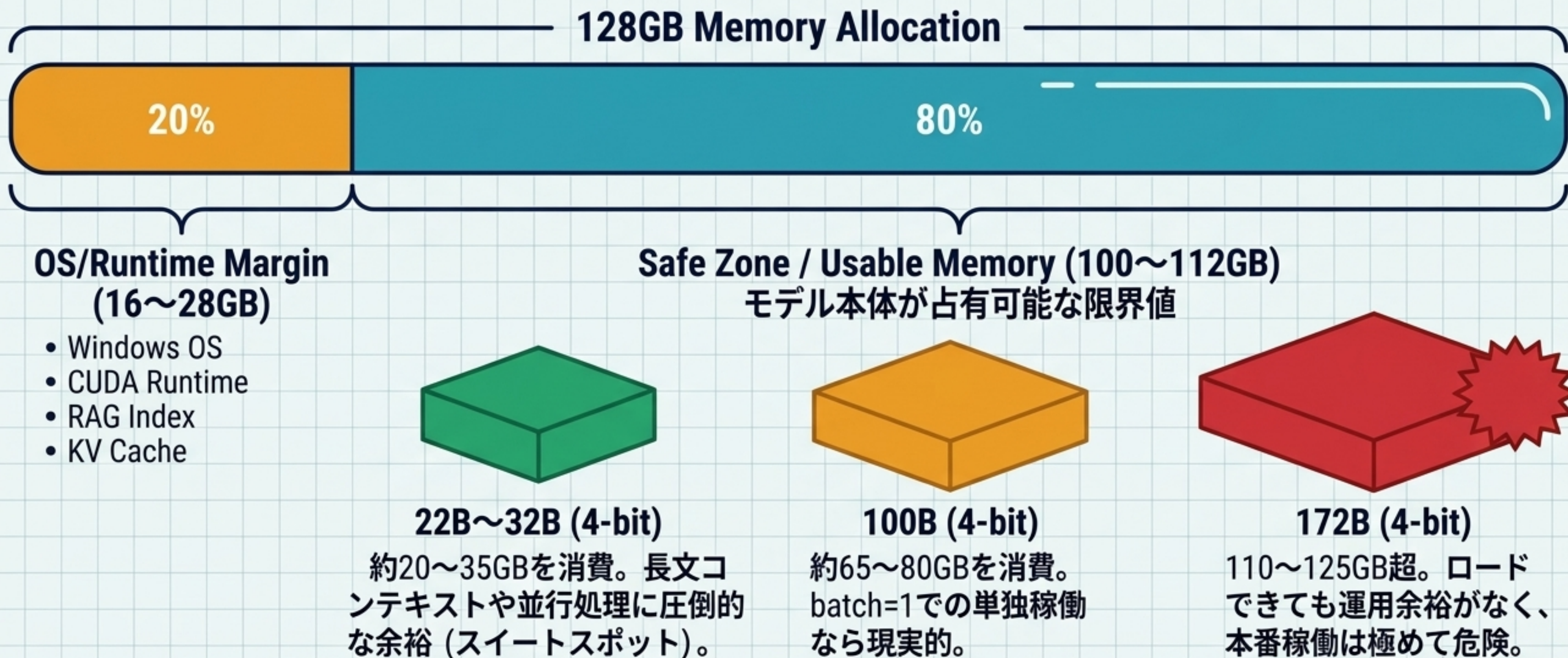
【統制】ガバナンスとRAG

「ローカル＝安全」「国産＝権利クリア」という認識は危険。

J-PlatPat等を用いた証拠分離型RAGと、出力の機械的検証（監査ログ）の構築が必須条件。

128GB統合メモリの真の姿：競合する有限のリソース

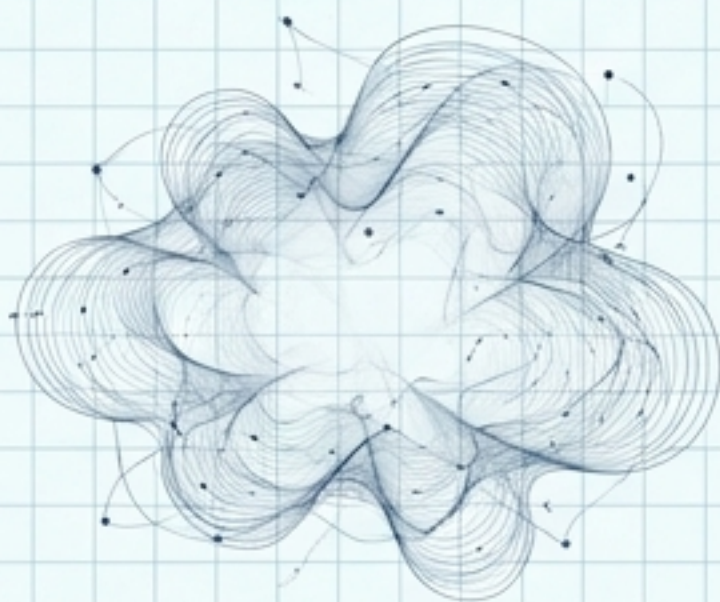
一般的な「RAM + GPU VRAM」とは異なり、RTX Sparkでは全ての処理が単一のメモリ空間を奪い合う。



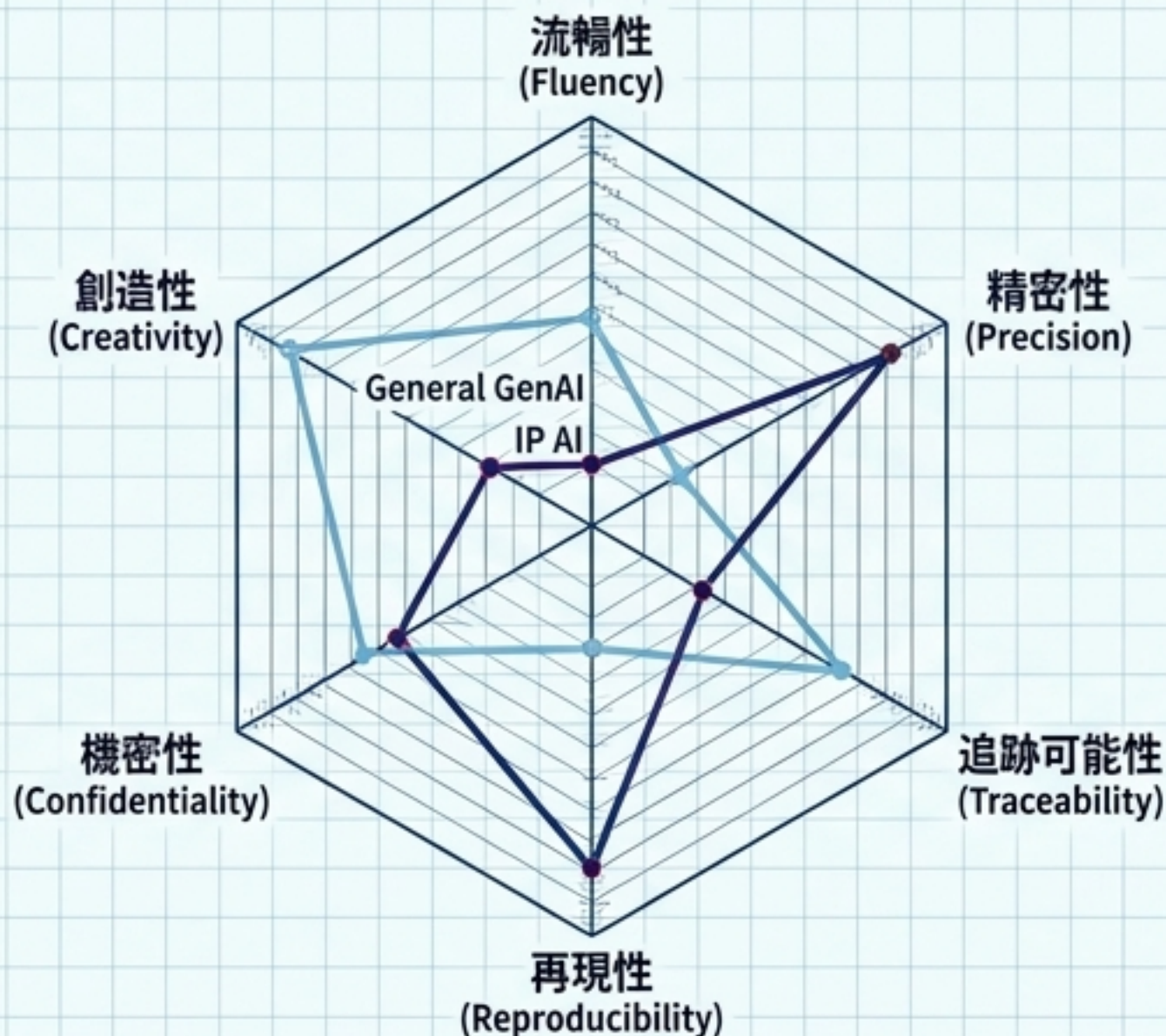
知財業務がAIに求める特異な要件

知財領域において、LLMは「クリエイター」ではなく、厳密な「チェッカー／分解者」でなければならない。

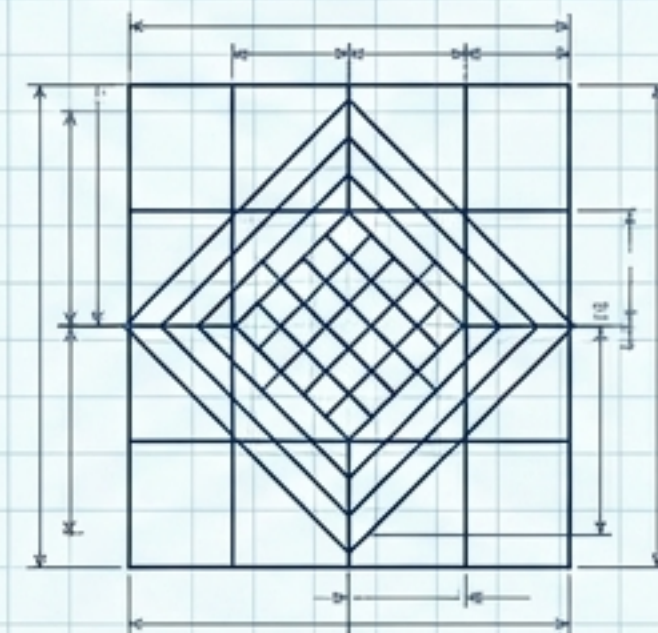
一般的なLLM用途 (Chat / Creation)



- 流暢な文章生成
- モデル内部知識への依存 (ゼロからのアイデア出し)
- 自由度・創造性の重視



知財業務の要件 (IP Operations)



- 精密な読解: 長文中の用語や参照関係の厳密な把握。
- 証拠の追跡: J-PlatPat等、一次データ (権威ある引用元) の絶対視。
- 再現性と統制: 同一入力に対する安定した出力と、誤りを検出する機械的ワークフロー。
- 機密保持・ライセンス: NDA資料の保護、法務利用基準のクリア。

「国産LLM」の再定義：ソブリン性と派生リスク

上流ライセンスや海外依存度に応じ、企業は「どこまでの純度」を求めるか判断が必要である。

国内適応型
(Domestic Adapted)

国内フルスクラッチ型
(Sovereign / Full-Scratch)



※文化庁の整理に基づく注意喚起：「フルスクラッチだから著作権リスクがゼロ」ではない。ライセンスと生成物の侵害責任は別次元の問題として法務評価が必要。

候補モデル総合比較マトリックス（RTX Spark 128GB環境）

モデル名 (Model)	推論メモリ目安 (4-bit時)	商用ライセンス (License)	知財適合度 (IP Fit)	特記事項・懸念点 (Notes)
Qwen3-Swallow-32B	24-35GB	Apache 2.0	◎	AWQ INT4公式対応・最高バランス。 GPTQ版の反復履歴に注意、AWQ必須。
Stockmark-2-100B	65-80GB	商用可	◎	特許・技術文書に強み。 メモリ逼迫のため batch=1 運用推奨。
CALM3-22B-Chat	17-25GB	Apache 2.0	○～◎	高速・安定。2024年世代。 最新推論モデルとの精度比較要。
RakutenAI-47B (MoE)	34-45GB	Apache 2.0	○～◎	MoEカーネルの対応要検証。 総重量は47Bである点に注意。
LLM-jp-172B	110-125GB超	条件付き可	△	ロード限界。無資格者の法的サービス 提供制限等、独自規約あり。

推奨構成：適材適所の「三段構え」アーキテクチャ

単一の巨大モデルに頼るのではなく、用途と負荷に応じたモデルのルーティング（振り分け）が最適解である。

Tier 1: 常用・主軸モデル (The Workhorse)

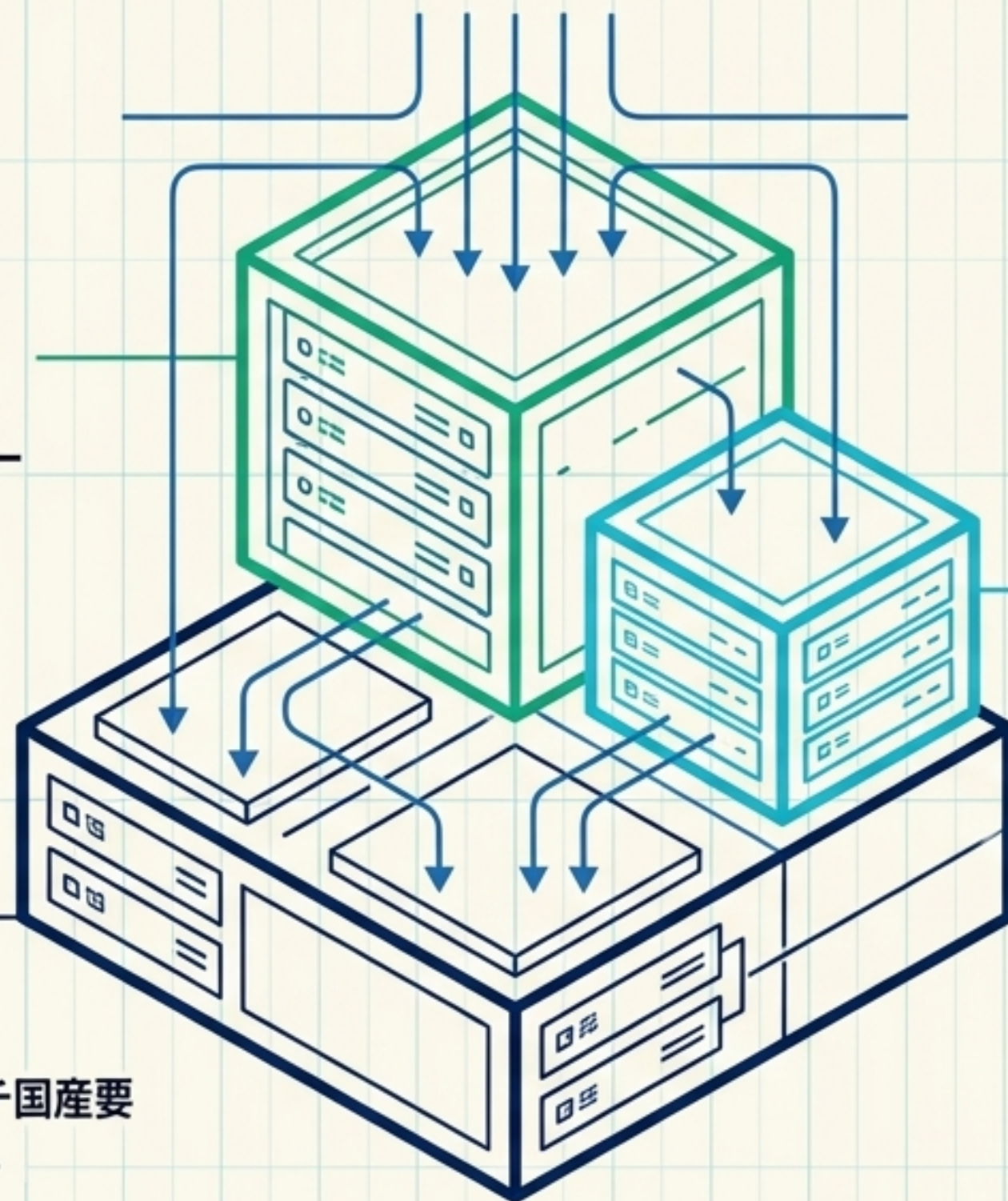
Qwen3-Swallow-32B (AWQ 4-bit)

- 日常的な対話、明細書の要約、クレーム構成要件の分解。
- メモリに圧倒的な余裕があり、複数ユーザーの処理やRAGデータベースの常駐に最適。

Tier 3: 安定・ソブリンモデル (The Sovereign Fallback)

CALM3-22B-Chat

- 高速な大量スクリーニング、フルスクラッチ国産要件が求められる極秘案件のフォールバック。



Tier 2: 難案件・高精度モデル (The Specialist)

Stockmark-2-100B-Instruct (4-bit)

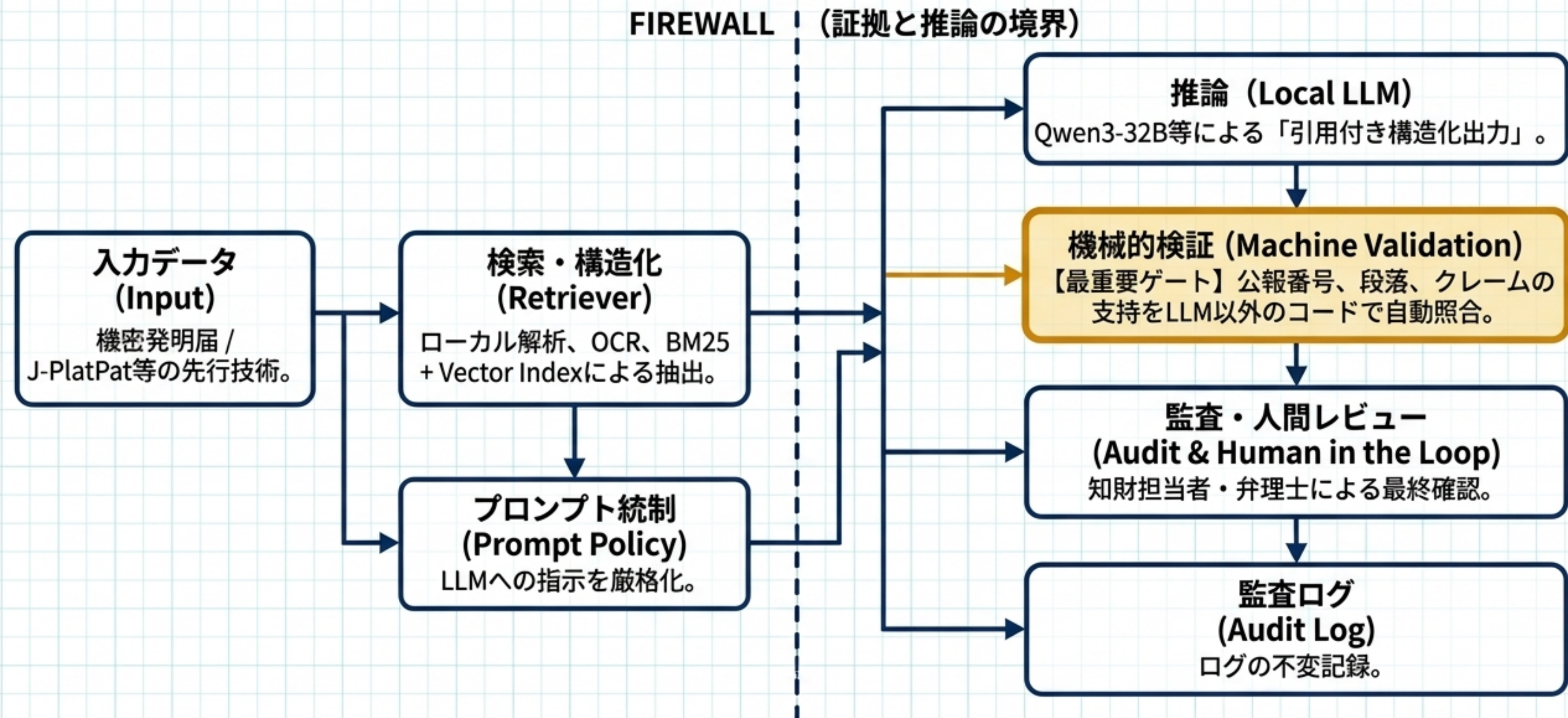
- 複雑な技術文書の比較、長文の高度な文脈解析、第二意見の生成。
- ※夜間バッチ処理や、batch=1での専有環境での稼働を推奨。

知財実務タスクごとのモデル適合性マッピング

	小型（3B）	主軸/中型（22-32B）	大型（100B）
先行技術の大量スクリーニング （Screening）			
	- 最適なアプローチ: 小型～中型モデル（3B～32B）で高速処理。 - 理由: 類似度判定に高度な推論は不要。ルーティングと抽出が鍵。		
クレーム（請求項）の構成要件分解 （Decomposition）			
	- 最適なアプローチ: 主軸モデル（32B）。 - 理由: 流暢さより、正確な論理分割とフォーマット遵守が求められる。		
複雑な文献対比と新規性評価の補助 （Deep Analysis）			
	- 最適なアプローチ: 大型モデル（100B）。 - 理由: 複数の長文間の微細な差異（差分）を自然言語で説明する高度な読解力が必要。		
契約書（IP条項）レビュー （Contract Review）			
	- 最適なアプローチ: 主軸モデル（32B）＋社内Playbook RAG。 - 理由: 法律判断をモデルにさせず、承認済み基準との「差分抽出」に特化させる。		

Target Architecture：知財特化型ローカルRAG構成図

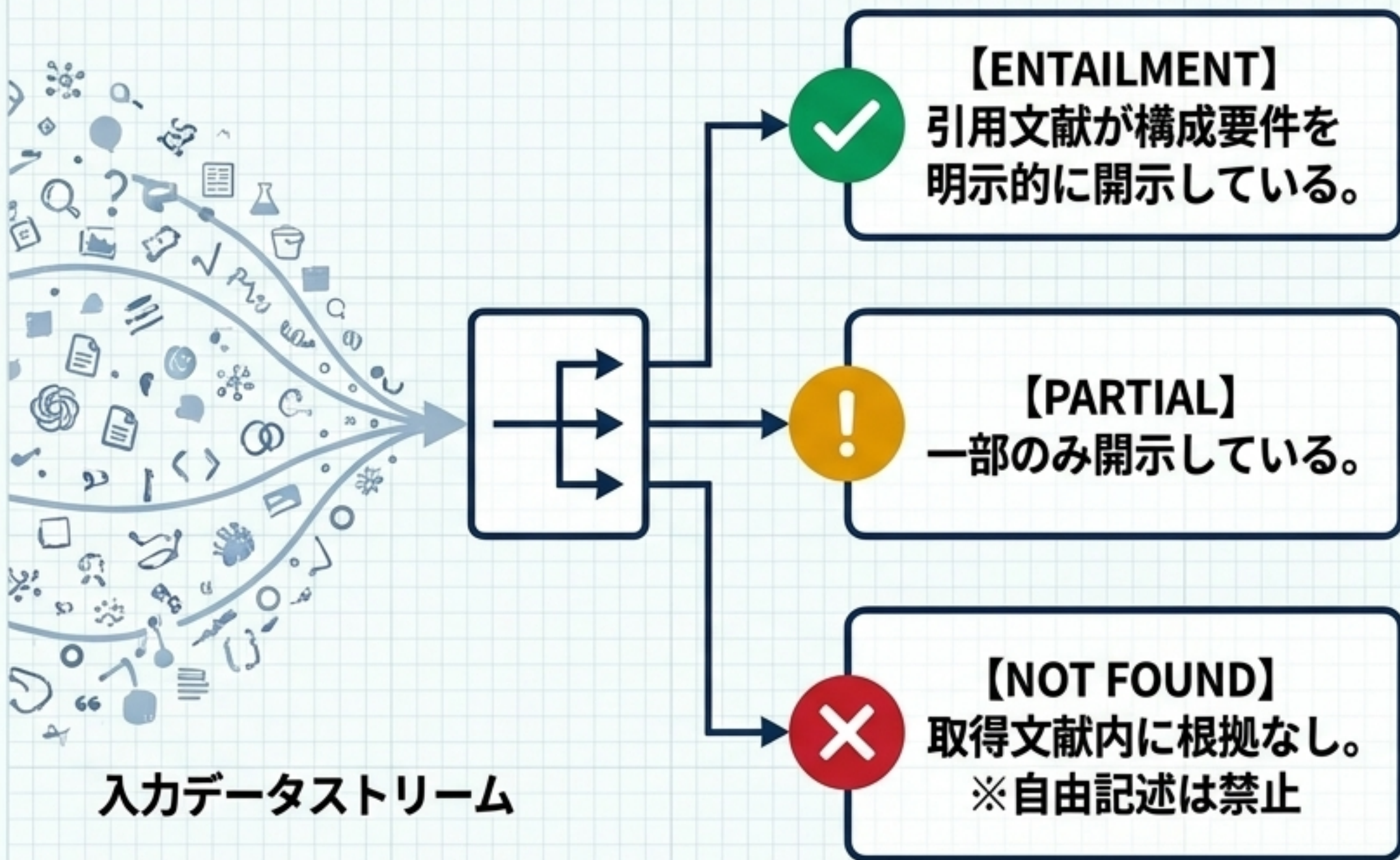
モデル内部の「知識」と、公報から得た「証拠」を完全に分離・隔離する。



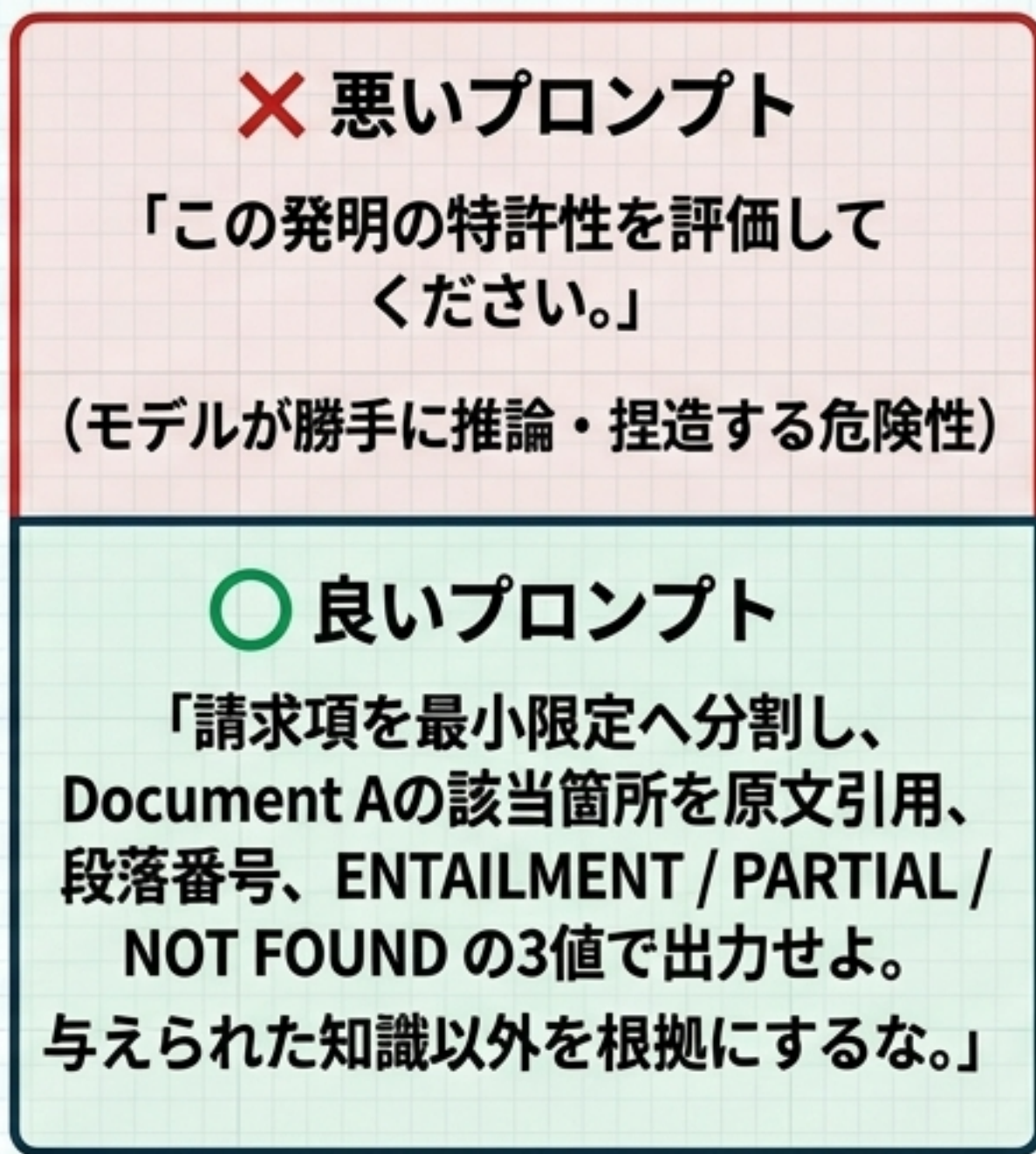
ハルシネーションの構造的排除：3値分類ゲートウェイ

LLMに「特許性」を判断させてはならない。事実を証拠（Evidence）に拘束するプロンプト設計。

Triaging Filter



Before/After Prompt Policy



クレーム生成 (Claim Drafting) における機械的検証

流暢なだけのクレーム生成は法的リスクの温床。生成温度を下げ、対応関係を強制する。

構成要件 (Claim Element)	明細書の支持段落 (Supporting Paragraph)	図面符号 (Drawing Reference)	信頼度スコア (Confidence)
【構成要件1】 ...	【明細書段落0025, 0028】	【図1, 2】	【98.5%】
【構成要件2】 ...	【明細書段落0027, 0028】	【図1, 2】	【98.5%】
【構成要件3】 ...	【明細書段落0028, 0028】	【図1, 2】	【98.5%】
...	...	...	...
...	...	...	...

Mechanical Guardrails (LLM外での自動検知項目)
⚙️・アンティセデント（先行詞）の不整合 ⚙️・従属関係の論理破綻 ⚙️・明細書に存在しない「新語」の混入 ⚙️・同一用語の表記揺れ、数値範囲の勝手な変更

法務・機密・セキュリティ統制（4つの防壁）



1. 機密保持 (Confidentiality)

「ローカル＝漏洩ゼロ」ではない。キャッシュ、RAGのDB、Windows/WSLのファイルシステム暗号化、監査ログの分離保護が必要。



2. ライセンスと著作権 (License & Copyright)

Apache 2.0は「著作権侵害の免責」を意味しない。特定モデル（NIM版とHF版）で利用規約が異なる点に嚴重注意。



3. 輸出管理 (Export Control)

海外子会社からのリモートアクセスやモデル共有は、外為法・米国輸出管理法令の対象となり得る。

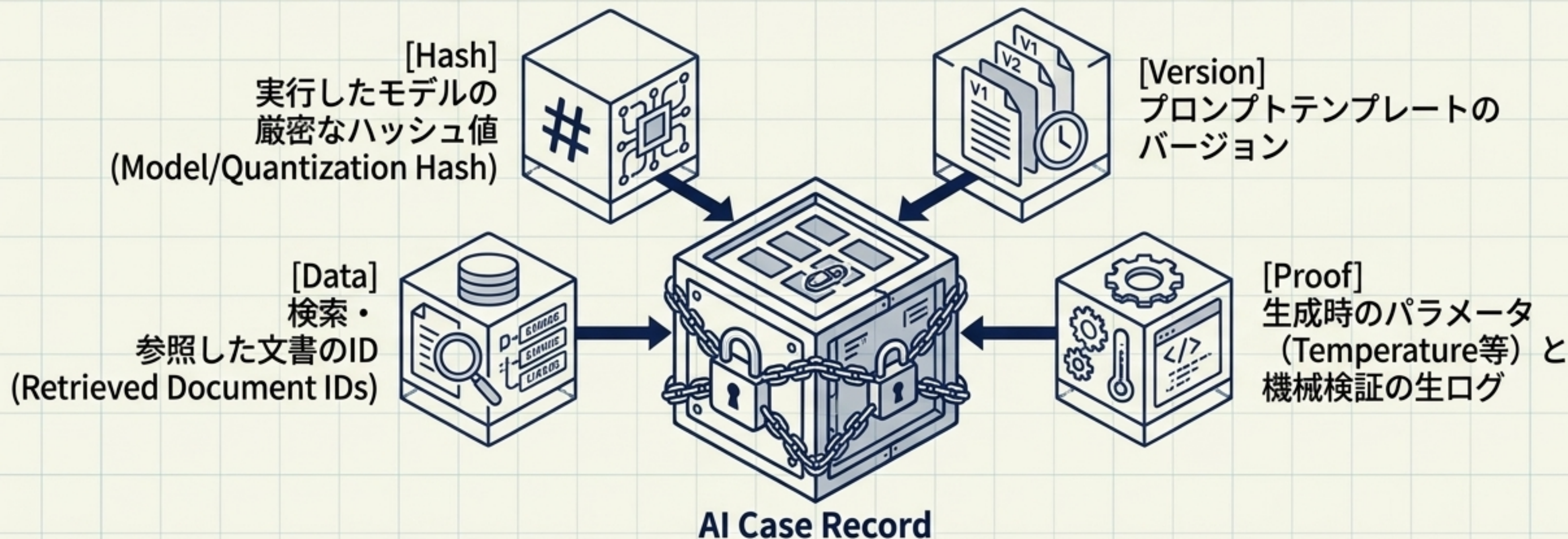


4. 独自規約のトラップ (Specific Terms)

特定モデル（LLM-jp等）の規約には「無資格者による法的サービス提供制限」が存在。知財ツールを社外SaaS化する際は法務確認が必須。

監査証跡 (Audit) と再現性の確保

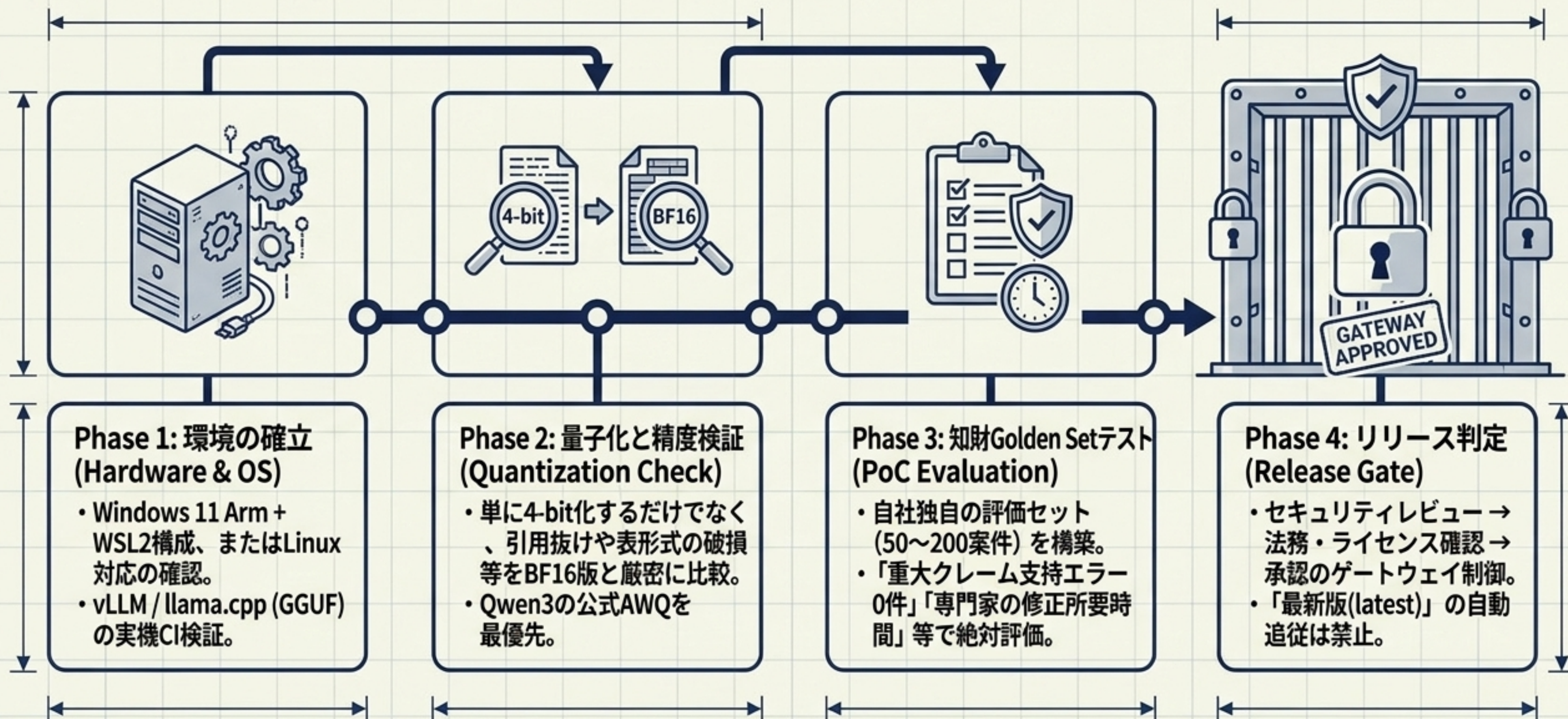
AI事業者ガイドライン準拠。「出力結果」だけを保存する運用は、監査に耐えられない。



Governance Note:

監査目的であっても機密プロンプトの永久保存は新たな情報漏洩リスクを生む。プレーンテキスト、ハッシュ値のみ等、機密度に応じた保存期間 (Retention Period) の明文化が必要。

本番導入への実装ステップ (Implementation Checklist)



結論：ハードウェアのポテンシャルを解放する「堅牢な設計」

- ✓ **RTX Spark 128GBの強み:** 正しいモデルサイズ（32B/100B 4-bit）を選択すれば、知財業務に革命的な並列処理・長文解析をもたらす。
- ✓ **AIを「法廷」に立たせる覚悟:** LLMの自由回答を封じ、証拠（J-PlatPat等）に基づく機械的検証と監査ログをシステムに組み込む。
- ✓ **Next Action:** Qwen3-Swallow-32B（主軸）とStockmark-2-100B（難案件）を用いた、自社データ（Golden Set）でのクローズドPoCフェーズへ移行する。



Intelligence constrained by evidence is the only intelligence suited for Intellectual Property.
(知財における真の知性とは、証拠によって統制されたAIにのみ宿る)

