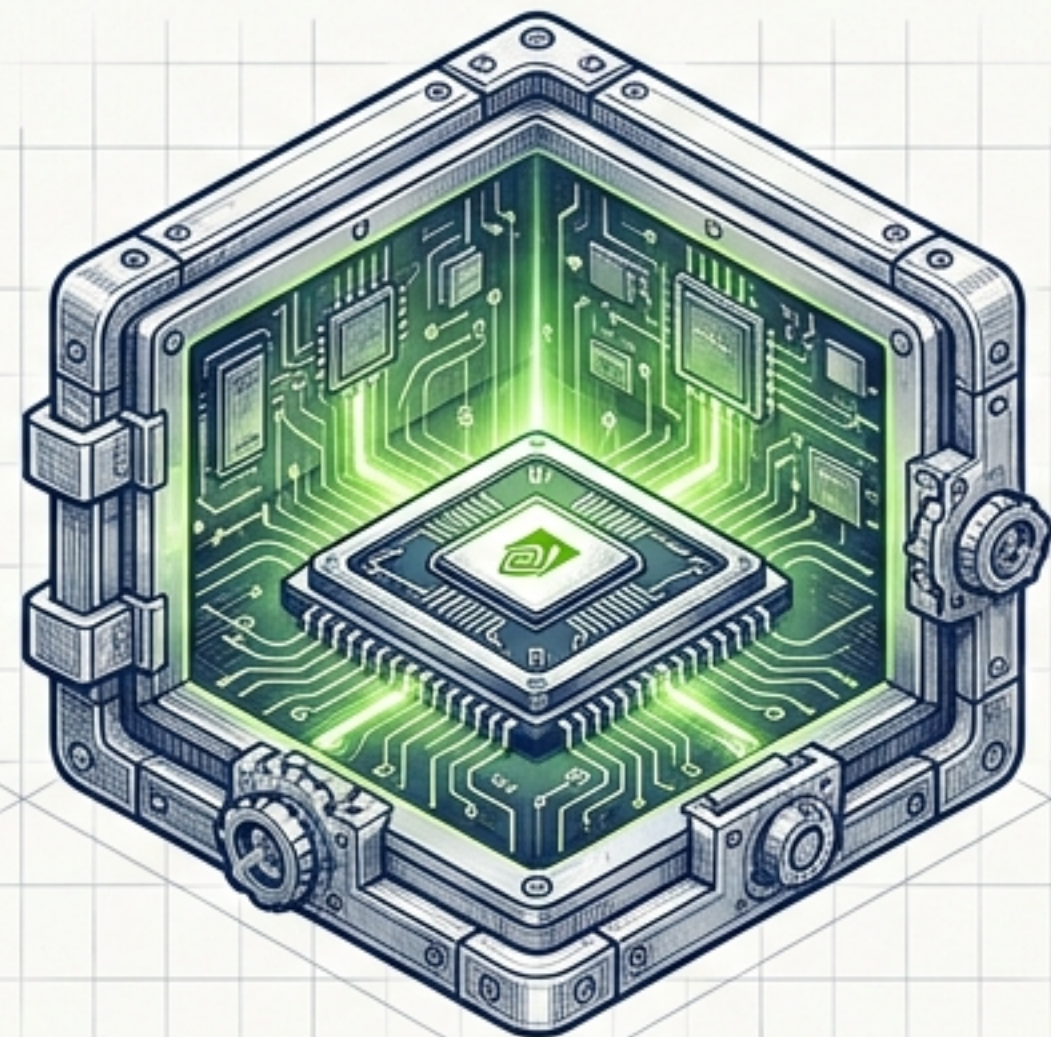


# NVIDIA RTX Spark (128GB) が実装する知財特化型 ローカルLLMの完全要塞化アーキテクチャ



機密保持と高度推論を両立するハードウェア選定・  
階層型ルーティング・法的リスクマネジメントの統合設計図

[Target: IP & Legal-Tech Architects]

[Status: Verified Blueprint]

[Classification: Internal Confidential]

# 知的財産業務におけるAI活用の構造的ジレンマ：クラウドの法的致命傷とローカルの物理的限界

## ！クラウドLLMの利用（法的リスクによる完全な遮断）



### RISK ALERT

新規性喪失リスク: 特許法第29条第1項。マルチテナント環境への未公開発明の送信は「不特定多数が知り得る状態」とみなされ、特許権が永久に失われる危険。

### CRITICAL

秘密管理性の崩壊: 不正競争防止法における営業秘密保護の無効化。

### VIOLATION

弁理士法違反リスク: 第30条に規定される秘密保持義務への抵触。

## ！既存ローカルPCの利用（物理的制約による破綻）



### HARDWARE LIMIT

VRAMの絶対的枯渇: 既存の独立型GPU（16～32GB水準）では、実務に耐えうる70B～120B規模の高度推論モデルをオンボード展開不可能。

### BOTTLENECK

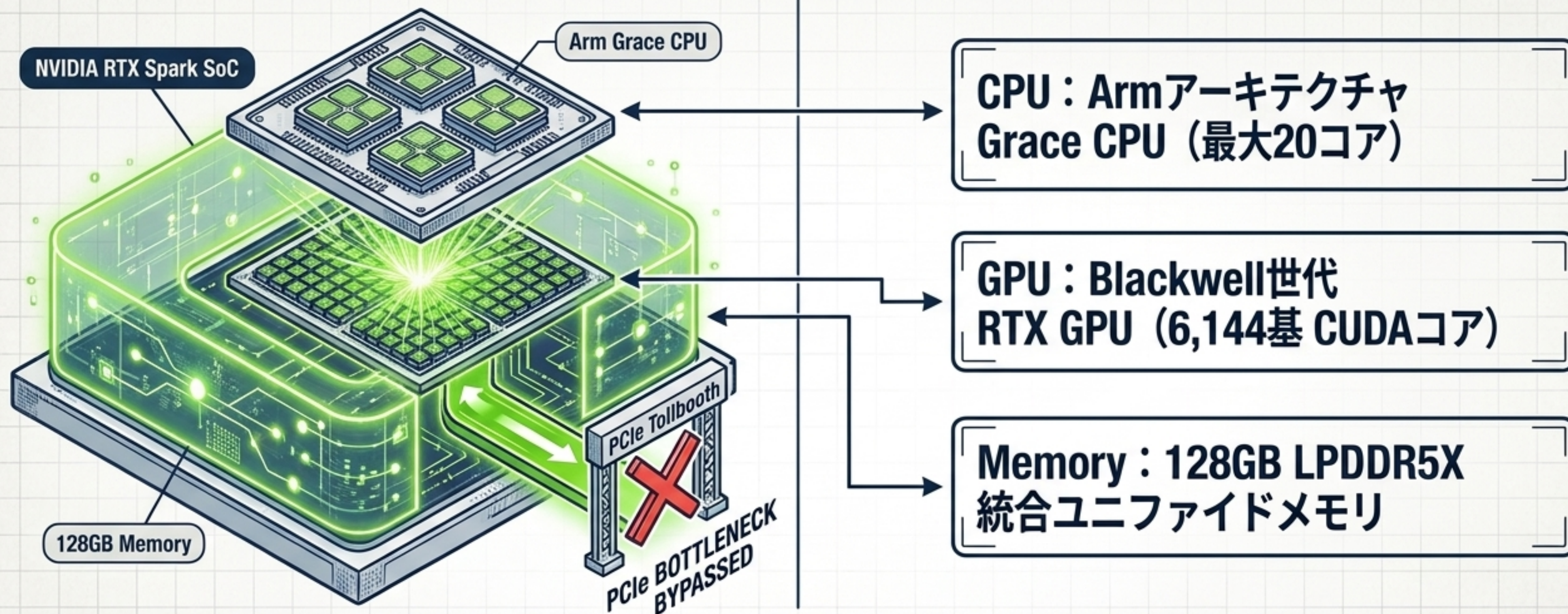
遅延のボトルネック: PCIeバス経由のオフロード処理による致命的なレイテンシ。

DEAD END

行き詰まり

結論: 外部通信を物理遮断しつつ、単一デバイスで128GBのメモリ空間を確保する新たな計算基盤が不可欠。

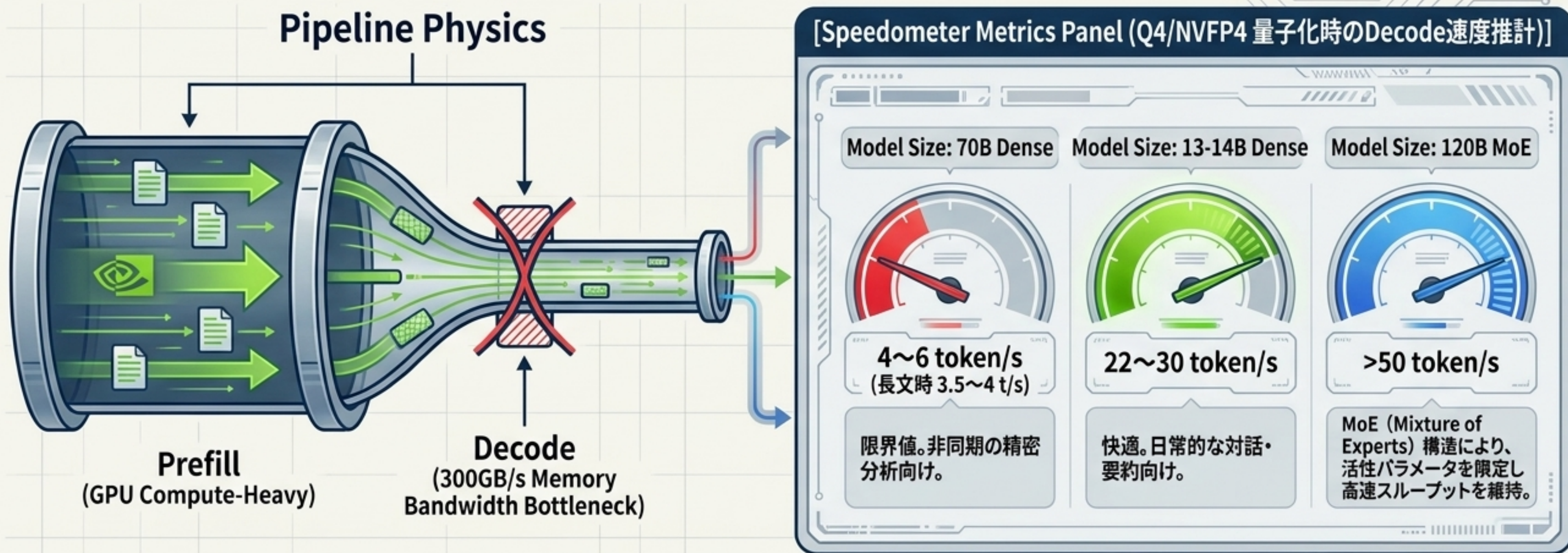
## ハードウェア基盤：128GBユニファイドメモリが実現するオフロード遅延ゼロの広大な推論空間



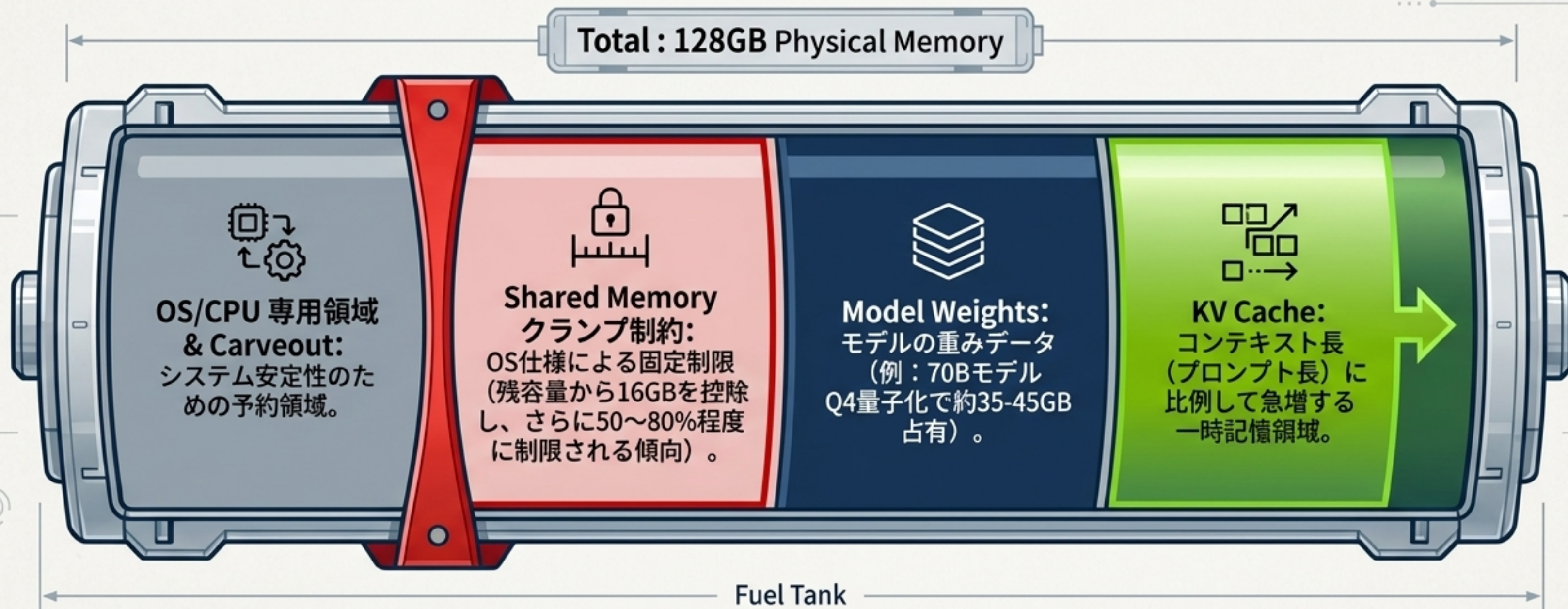
**単一SoC統合アーキテクチャの優位性：**「大容量モデルの展開」と「即応性」の両立。  
従来のデスクトップPCを悩ませていたPCIeバス経由のデータ転送ボトルネックを排除。70B～120Bクラスの超巨大モデルを単一のメモリ空間へ直接展開し、コヒーレントな共有を実現。知財部門のオンプレミス環境に、データセンター級のモデル収容力をもたらす。

# 推論の物理学：128GBの「容量」と300GB/sの「帯域幅」の厳密な区別

モデルが「収容可能」であることと、実務に耐えうる「対話速度」が出ることの境界線。プロンプトを一括処理するPrefill工程はGPU演算に依存するが、1トークンずつ出力するDecode工程（バッチサイズ1）はRTX Sparkのメモリ帯域幅（300GB/s）が物理的ボトルネックとなる。



# Windows on Armのメモリ管理構造：128GB物理メモリの実効確保可能領域



## 【Alert: System Crash Risk】



単一プロセスが128GB全域をVRAMとして自由に占有できるわけではない。特に知財文書（特許公報等の長大テキスト）を入力する際、KVキャッシュの肥大化によるVRAM上限超過（OOM）を防ぐため、システム設計段階での厳密なクランプ値の特定とFP8等によるKV KVキャッシュ量子化が不可欠。

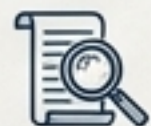
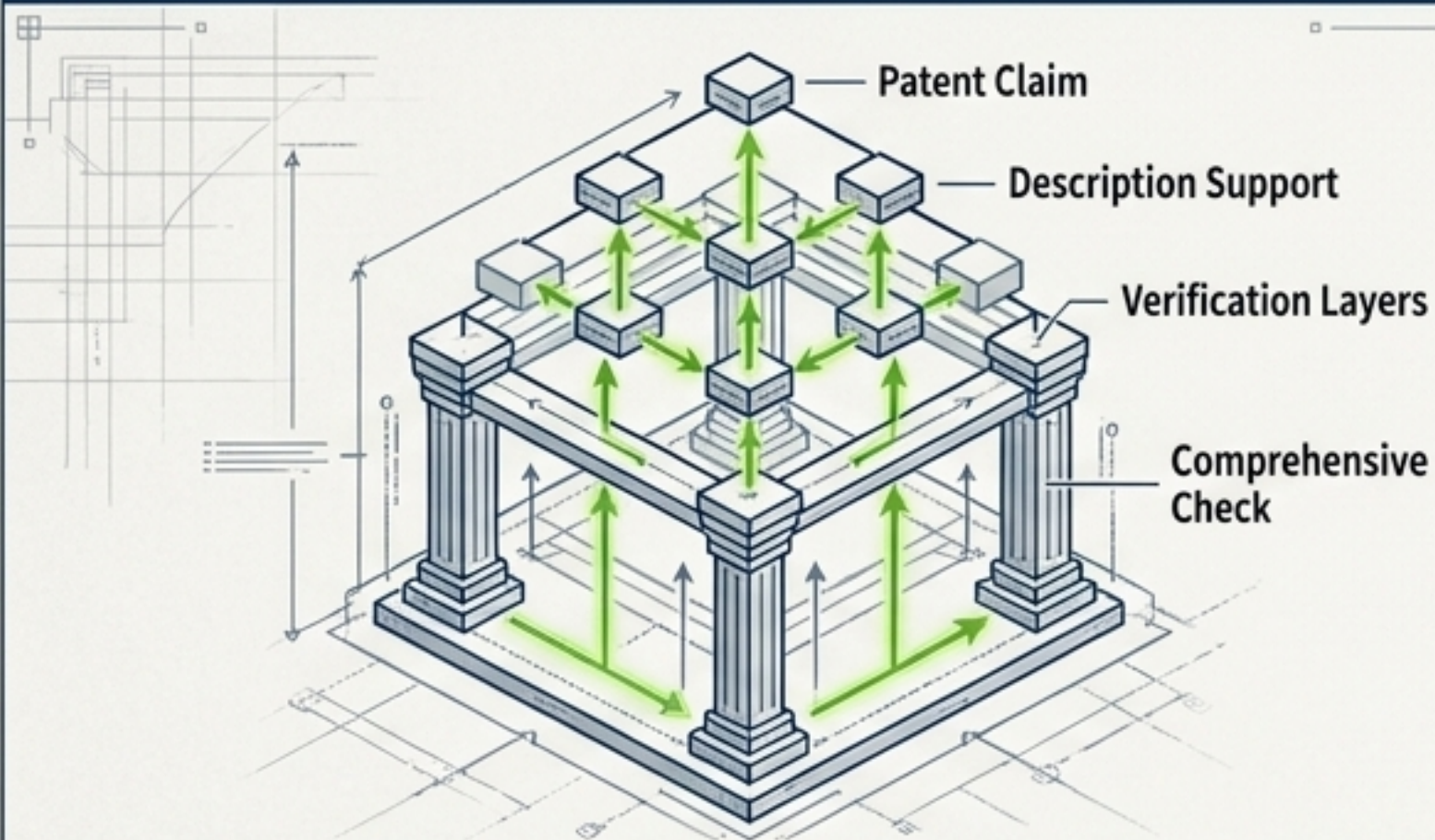
# 知財実務向けオープンウェイトLLM 選定マトリクス (ライセンス & 性能評価)

| Tier / Size       | Model Name                  | 推定Decode速度 | 知財言語適性・強み                                       | 商用ライセンス状態                                                                                                           |
|-------------------|-----------------------------|------------|-------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| [14-32B]<br>日常・論理 | Qwen 2.5<br>(32B/14B)       | 10-30 t/s  | 日: 優 / 英: 良<br>(日常的な要約、社内知財Q&A)                 |  Apache 2.0<br>(完全無償・商用可)        |
| [32B]<br>反論・構成対比  | DeepSeek-R1-<br>Distill-32B | 10-20 t/s  | 日: 優 / 英: 優<br>(CoT展開による進歩性・<br>構成要件の論理反論)      |  MIT<br>(完全無償・商用可)               |
| [70B]<br>英語・精密分析  | Llama 3.3<br>(70B-Instruct) | 4-6 t/s    | 日: 良 / 英: 極めて優<br>(英語特許読解、PCT対応、<br>外国審査応答)     |  Community<br>(条件付き無償)          |
| [72B]<br>クレーム整合性  | Qwen 2.5<br>(72B-Instruct)  | 4-6 t/s    | 日: 優 / 英: 優<br>(JMMLU上位。明細書とクレームの<br>厳密な整合性検査)  |  Apache 2.0<br>(完全無償・商用可)      |
| [104B]<br>RAG引用特化 | Command R+<br>(104B)        | 2-3 t/s    | 日: 良 / 英: 優<br>(正確なCitation付与による<br>ハルシネーション抑制) |  CC-BY-NC-4.0<br>(商用不可・別途契約必須) |

知財部門におけるオンプレミス運用では、法的安全性の観点からApache 2.0またはMITライセンスを最優先配備すべきである。

# 用途特化型プロファイル：「構造的整合性」のQwenと「論理反論」のDeepSeek

## Qwen 2.5 (72B / 32B) - 構造と網羅性の担保

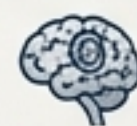
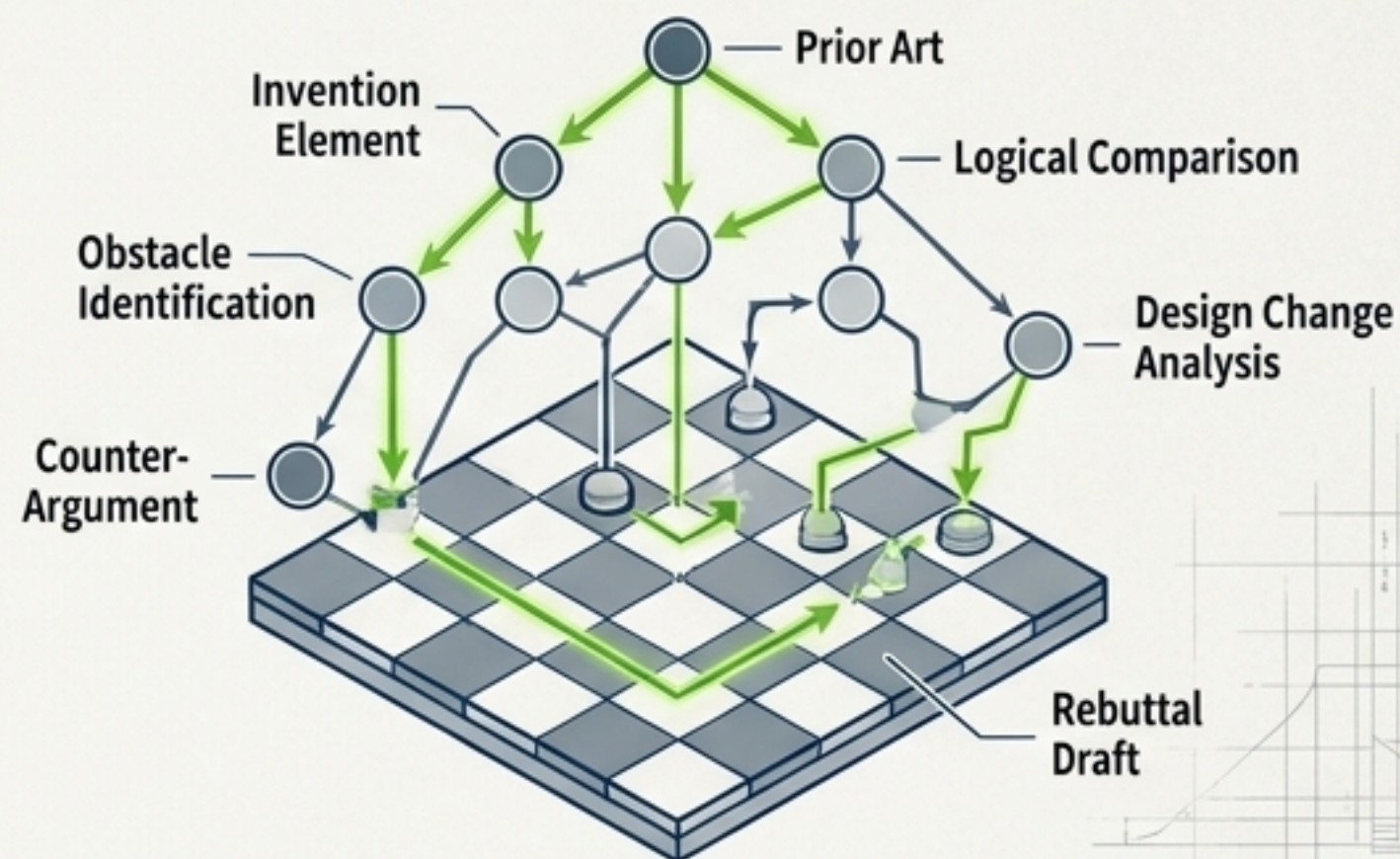


特性: 圧倒的な日本語読解力と最大128Kの長大コンテキストウィンドウ。



知財タスク適合: 特許法第36条（サポート要件・実施可能要件）違反の予防。数十ページに及ぶ公開公報を一括入力し、クレーム中の文言が明細書本文に網羅的に記載・定義されているかを多層的に照合する作業において、商用モデルに匹敵する精度を発揮。

## DeepSeek-R1-Distill (32B) - 思考プロセス (CoT) による論理構築

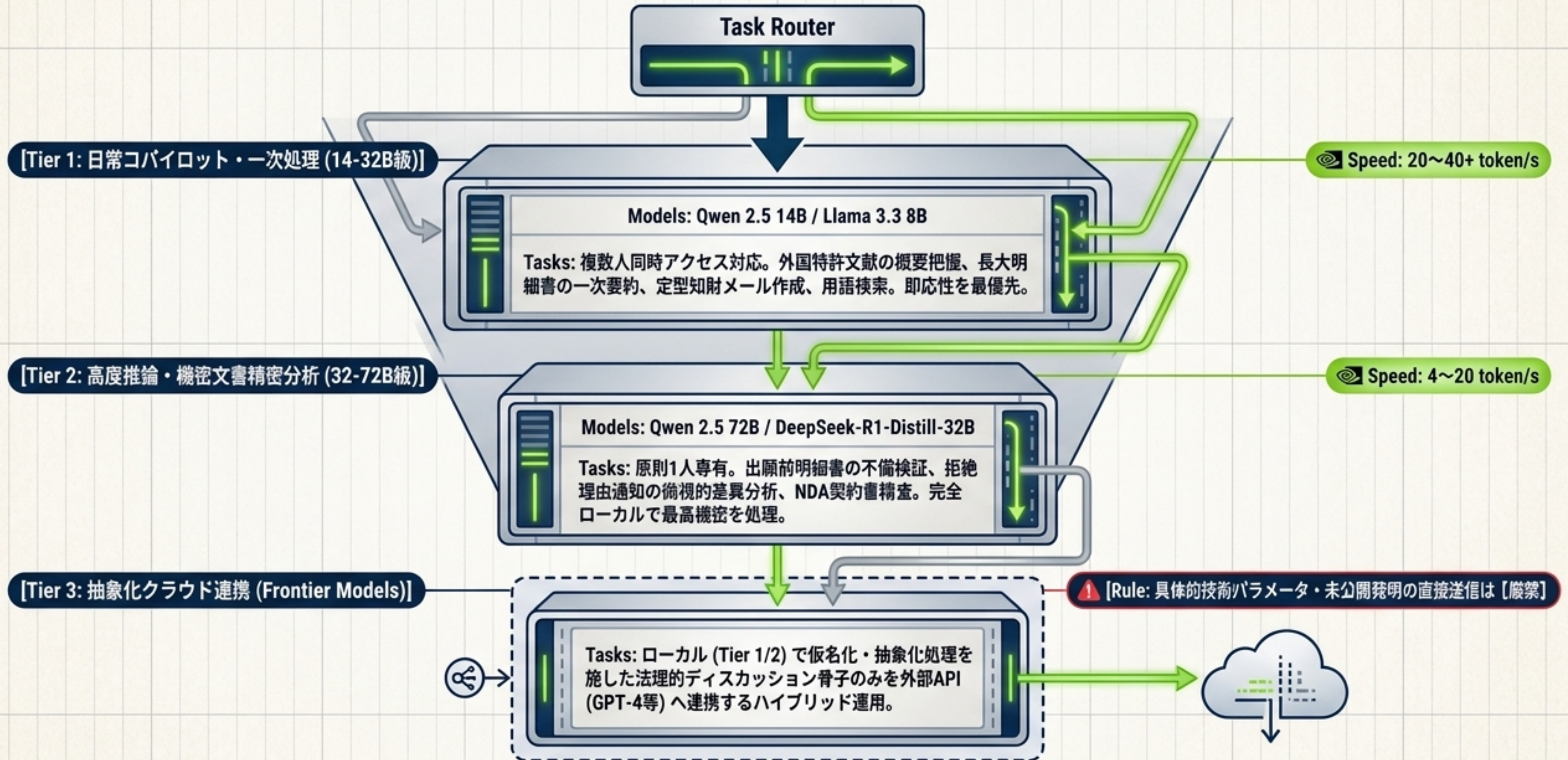


特性: 強化学習ベースの推論特化型蒸留モデル。結論に至るまでの思考プロセスを可視化。



知財タスク適合: 拒絶理由通知（OA）に対する反論構築。先行技術文献と本願発明の構成要件を対比し、「どの構成が公知技術からの設計変更に過ぎないか」「阻害要因はどこか」を論理的ステップに沿って深掘りし、意見書・補正書案の強力な補助輪となる。

# 階層型ルーティング (Tiered Routing) 設計：タスク要求に応じた計算リソースの最適分配



# 知財特化型Local RAGの構築：ハルシネーションを根絶する「請求項対応型チャンキング」

## [汎用的な固定文字数チャンキング (非推奨)]



特許明細書を機械的なトークン数で分割すると、請求項における構成要件の係り受け構造や、実施例と作用効果の対応関係が寸断される。これが、トークンと同等構造的な構造の為、対応関係が寸断される。これが架空の特許番号や誤った審判決例を確信をもって生成する「ハルシネーション」の根本原因となる。

## [Claim-aware Chunking (請求項対応型チャンキング)]

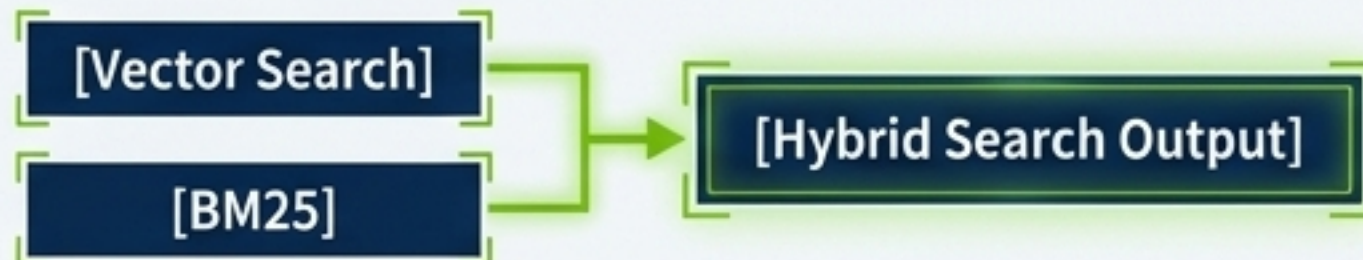


ドメイン固有の工学的配慮。

**セマンティック境界の維持:** 請求項、発明を実施するための形態、図面の簡単な説明といった論理境界を厳密に維持して分割。

**ハイブリッド検索の実装:** ベクトル検索単独では化合物命名法や特許分類記号のRecall (再現率) が不足する。

BM25 (完全一致検索)とベクトル検索を統合し、IPC/CPC分類や優先日メタデータフィルタリングを併用する。



# 長大コンテキストの陥穽：TTFT遅延の回避とRe-rankerによる文脈最適化

最新モデルの128K～1Mトークン対応を過信し、数十件の特許公報や明細書全文を直接プロンプトへ投入する運用は実務上破綻する。



## TTFT (Time to First Token) の壁:

プロンプトを読み込むPrefill工程は多大な計算リソースを消費する。80B～120Bモデルに128Kトークンを一括投入すると、最初の1文字目が出力されるまでに70秒～90秒以上の待機時間が発生し、対話型業務の効率を著しく阻害する。

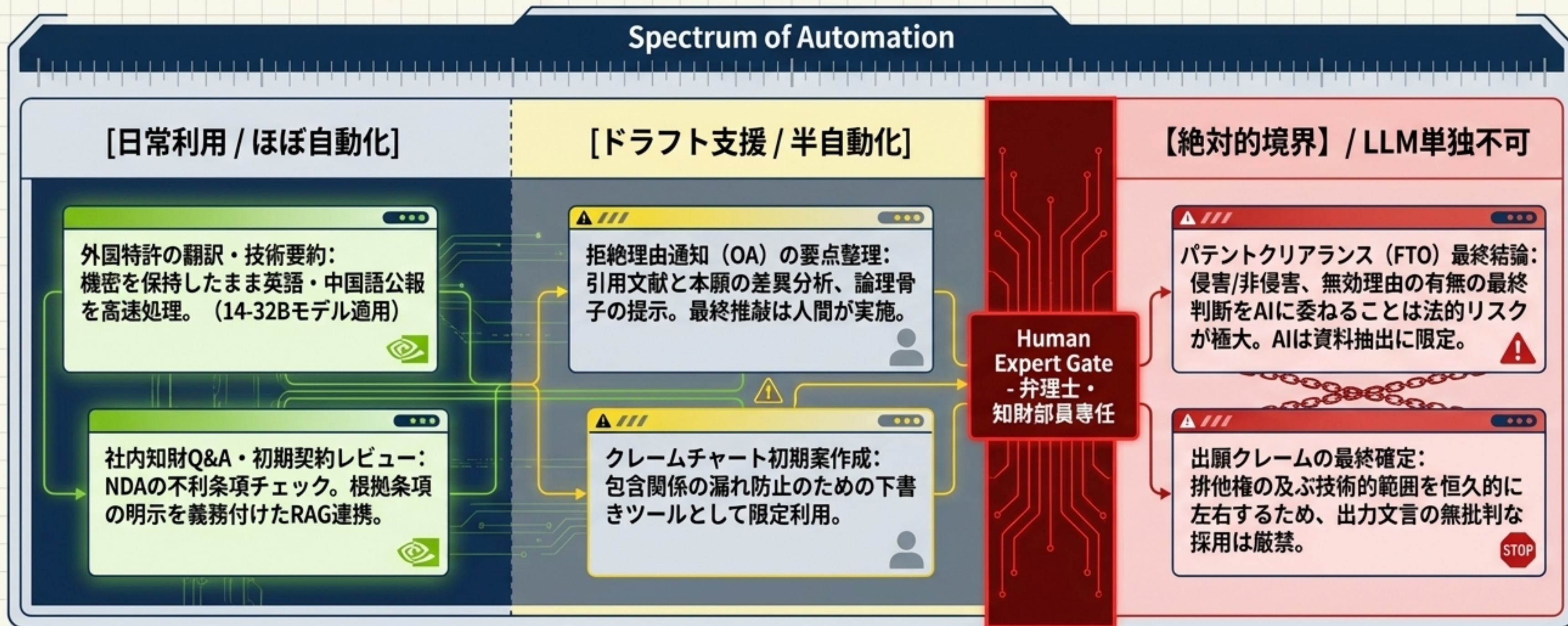


## 解決策 (BGE-Rerankerの導入):

検索された上位チャンクを無差別にLLMへ送るのではなく、交差エンコーダ型再ランキングモデルを用いて情報の適合率をスコアリング。本当に関連する数千～数万トークンのみをプロンプトに注入することで、応答速度の維持とKVとキャッシュ枯渇の防止を両立する。

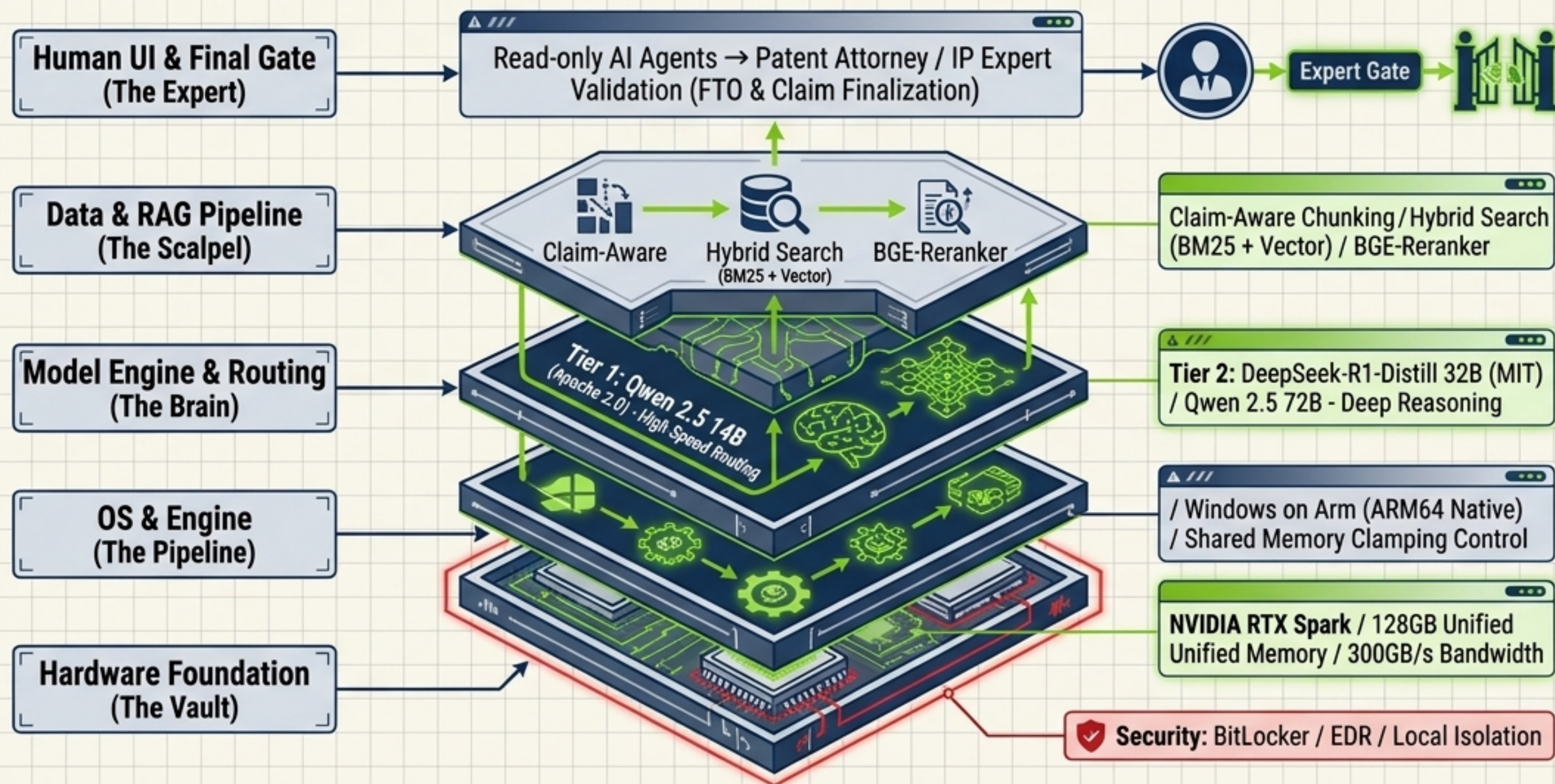


# Human-in-the-Loop リスクマップ：自動化スペクトラムと法的責任の境界線



【Bottom Security Note】 端末の物理的脆弱性を補うため、BitLocker暗号化、EDR常時稼働、およびAIエージェントの「Read-only」権限制限を必須ポリシーとする。

# 統合アーキテクチャ・ブループリント：専門家・ハードウェア・AIの完全連携モデル



RTX Sparkは単なる大容量PCではない。企業の未公開発明を強固に防御しつつ、出願・権利化業務の生産性を飛躍させる「知財要塞」の中核インフラである。