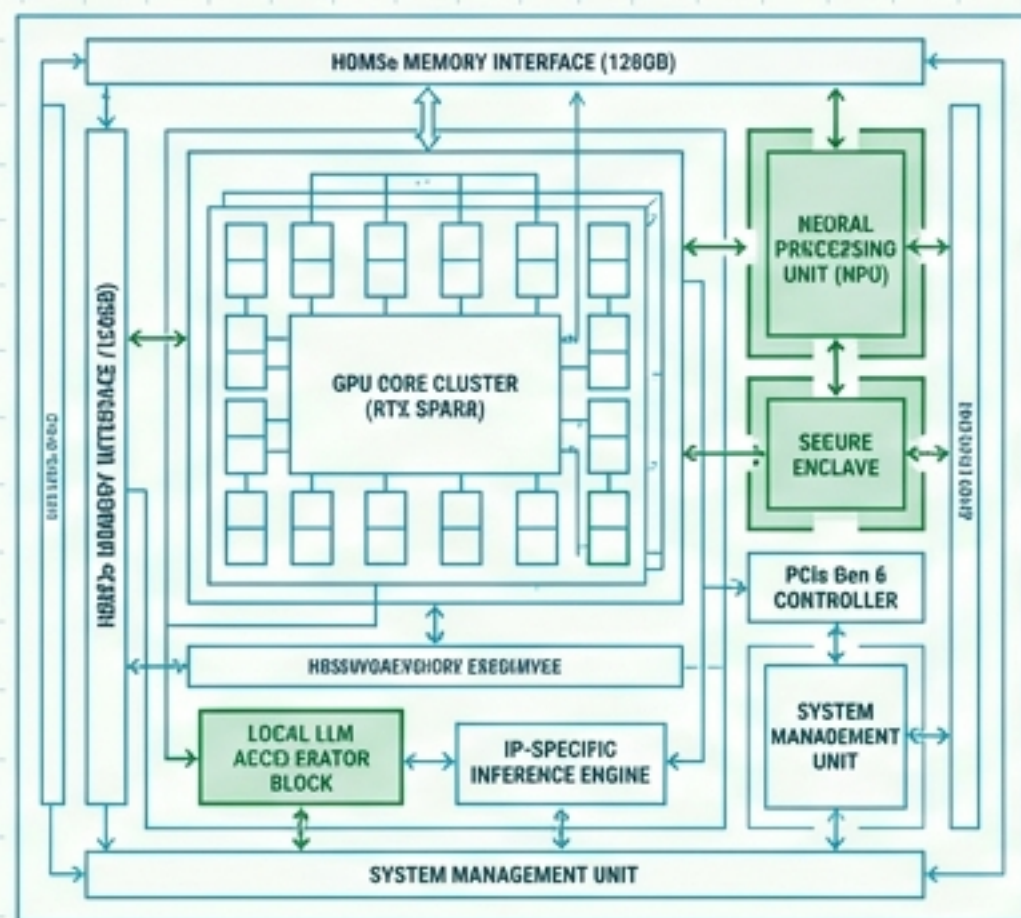


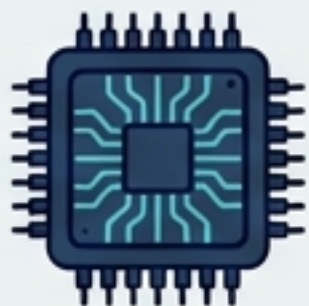


NVIDIA RTX Spark (128GB) 導入設計図：知財業務向け ローカルLLMの選定と実装

物理的制約・経済安全保障・弁理士法
に基づく戦略的アーキテクチャ

版数:	2026年10月最新版
作成:	Claude Opus 5.5
分類:	社内秘 / 経営陣・知財部門責任者向け情報

エグゼクティブ・サマリー



基盤（ハードウェア）

RTX Spark 128GB (Windows on Arm / N1X)

メモリ帯域300GB/sの制約下では、計算力よりも「MoE (Mixture of Experts) 構造」と「NVFP4量子化」の採用が実用速度の鍵となる。



推奨モデル構成

性能重視 vs 経済安保重視

主力モデルとして「Qwen3.5-122B」または非中国製の「gpt-oss-120b」を採用。これに高速な補助モデル（LLM-jp-4.1等）と特許RAG用埋め込み（ruri-v3）を組み合わせる。



ガバナンス（法務・倫理）

弁理士法と完全閉域網

弁理士会ガイドラインに基づく善管理注意義務（出力のファクトチェック必須）と秘密保持を担保するため、外部通信を完全遮断したローカル環境を構築する。

ハードウェア基盤：RTX Spark vs DGX Spark

RTX Spark (N1X)



OS: Windows 11 on Arm (Copilot+ PC)

Specs: 20-core Grace CPU,
6,144 CUDA cores,
1PFLOP (FP4)

Target: 個人の業務PCと同居させる
ハイエンド環境。

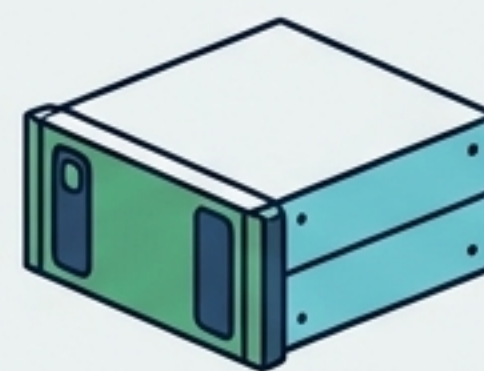
帯域幅:
300GB/s

帯域幅:
273GB/s

帯域幅:
300GB/s

帯域幅:
273GB/s

DGX Spark (GB10)



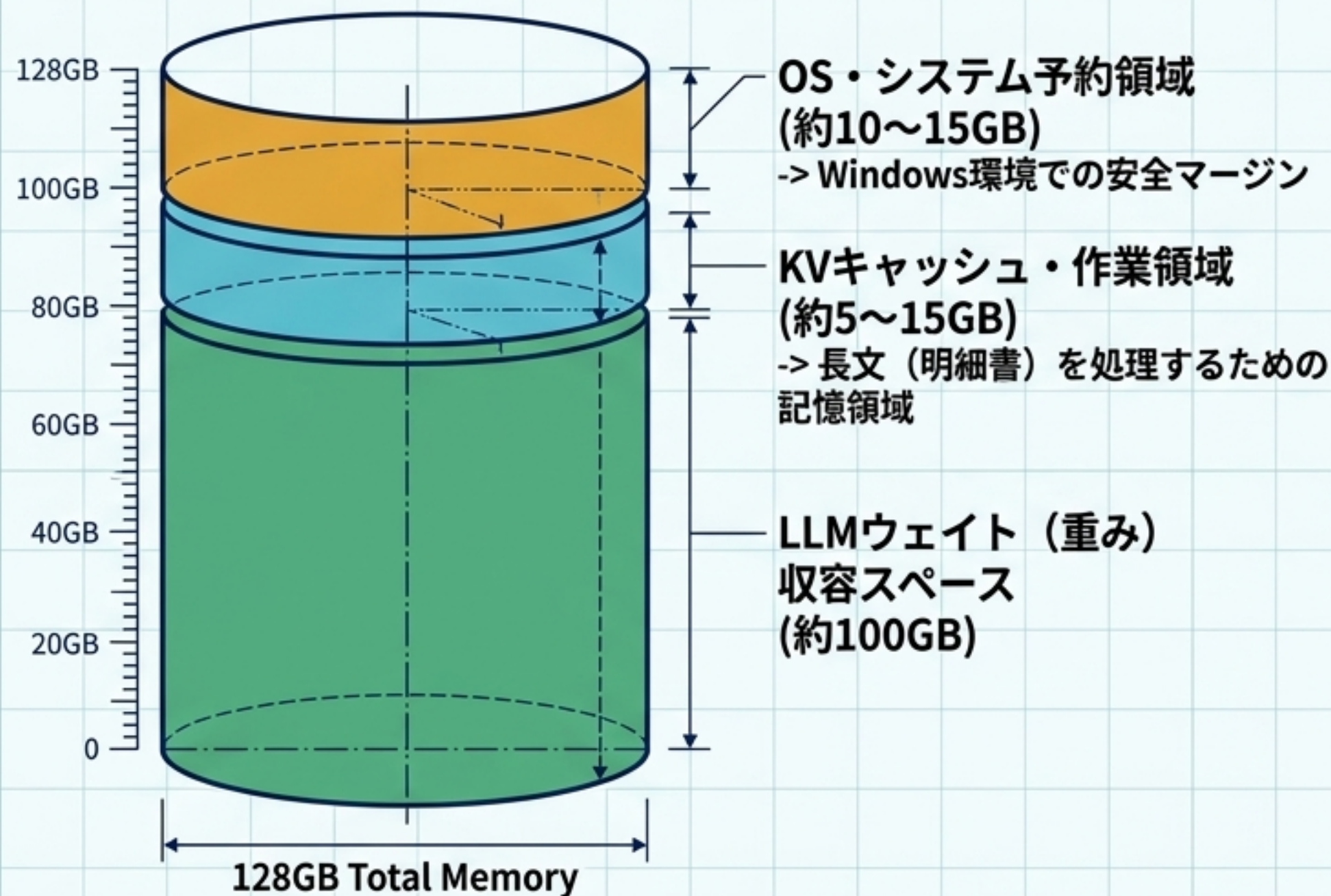
OS: DGX OS (Ubuntu 24.04 base)

Specs: 20-core Arm CPU,
Blackwell Tensor Cores

Target: 複数人での共有・vLLM等の
Linuxツール群をフル活用
するサーバー環境。

最重要インサイト: 生成速度は「帯域幅 ÷ 1トークンのパラメータ読込量」で決まる。プロンプト処理は数万トークンでも十数秒で完了するが、出力速度の確保が最大のボトルネック。

物理的制約の理解：128GBメモリの数学



量子化（Quantization）の必然性

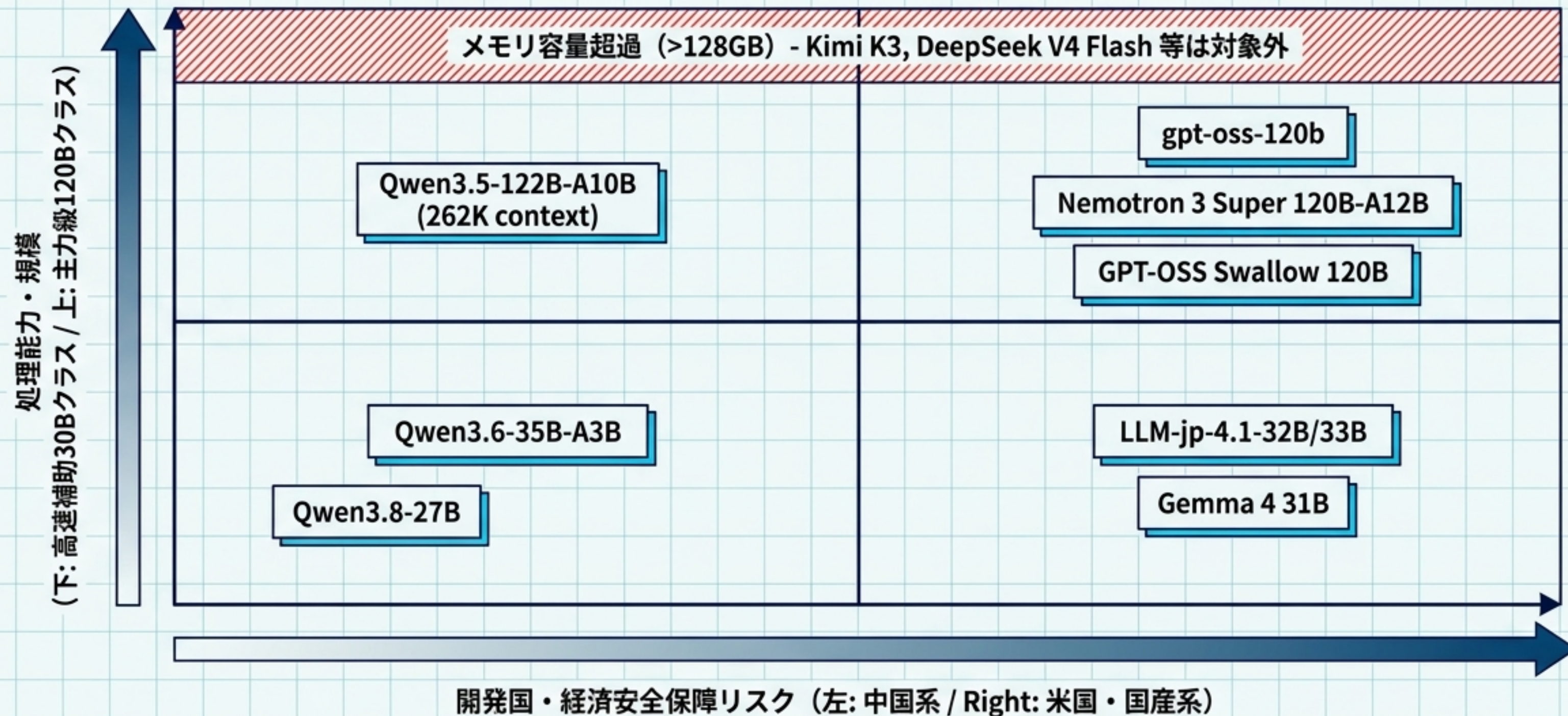
FP16 (2 bytes) から NVFP4/MXFP4 (約0.53 bytes) への圧縮。これにより120Bモデルを約65GBに収めることが可能。

MoE（Mixture of Experts）の必然性

密（Dense）モデルの70BはKVキャッシュが128Kトークンで約40GBを消費するが、Qwen3.5-122B（MoE）は約6~7GBで済む。

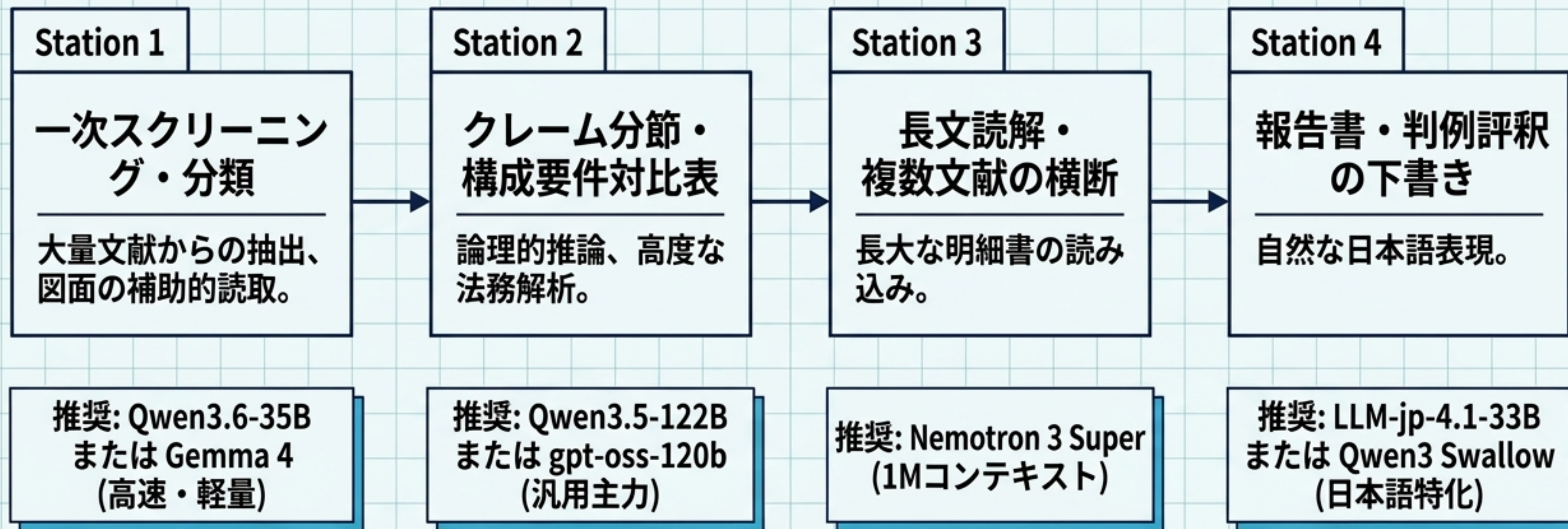
結論: 128GBという「物理的限界」に、120Bクラスの知能と数十万字の特許明細書を同時に押し込むには、ハイブリッドMoEと高度な量子化が「必須条件」となる。

候補LLM群の俯瞰マトリクス



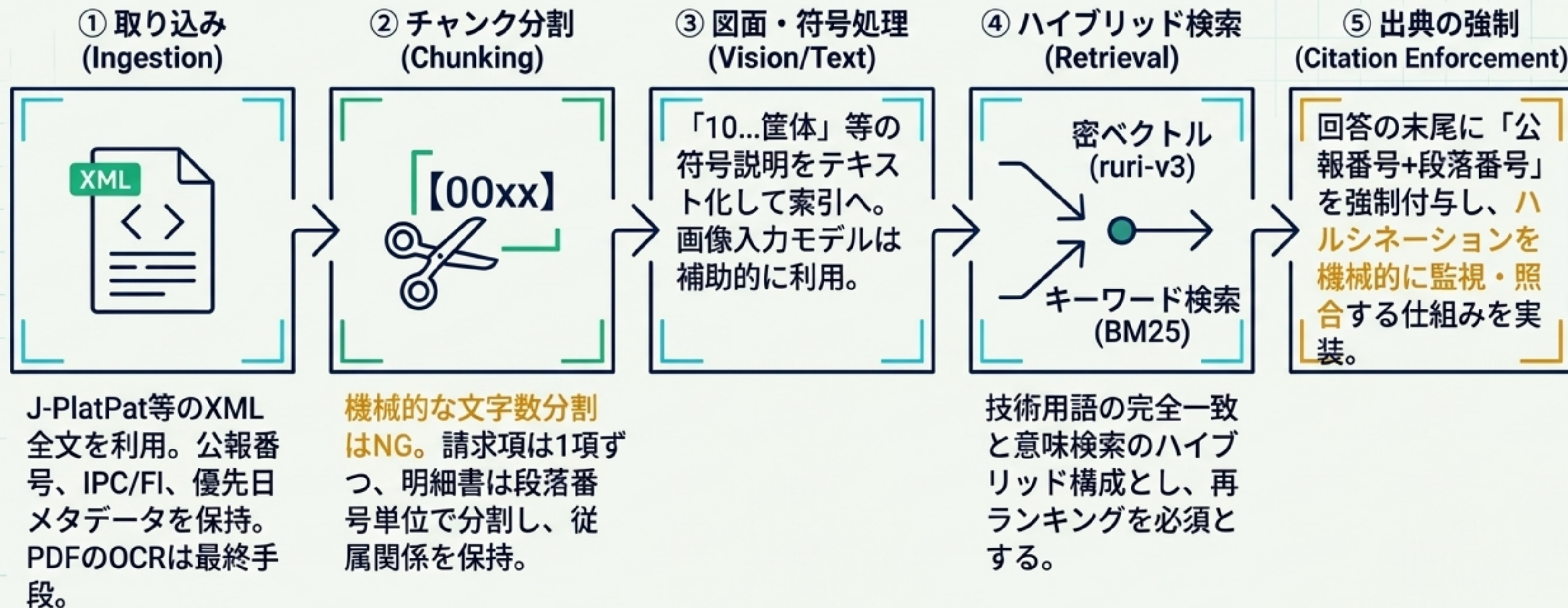
ベンチマーク動向 (JBE-QA, JAPAS等) を踏まえ、知財法務における高度な論理推論には120B級の能力が最低ラインとなる。

知財タスク別・推奨モデルマッピング



基盤レイヤー (RAG用 Embedding) : ruri-v3-310m (JMTEBスコア81.89) を全域に配置

特許文献特化型RAGの最適アーキテクチャ



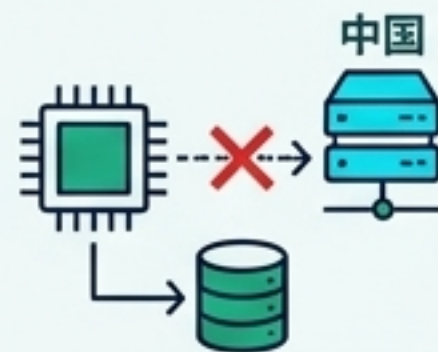
ディープダイブ：中国製モデルの論点と経済安全保障

✓ 事実と技術的現実

個情委・デジタル庁の警告
(2025年2月)は「クラウドサービス」が対象。



完全ローカル稼働（重みを動かすだけ）の環境下では、モデル自体が中国のサーバーへデータを送信する機能・通信能力は物理的に存在しない。

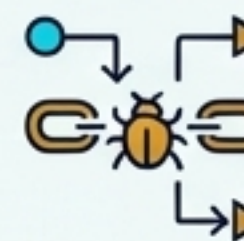


⚠ 残存する3つのリスク

1. 出力の偏り: 特定の政治的話題や記述におけるバイアス（知財実務への影響は限定的）。



2. サプライチェーンリスク: ツールチェーンやリモートコード経由の脆弱性。



3. クライアント方針: 防衛関連・重要技術分野では「中国製」という事実のみで契約違反となるリスク。



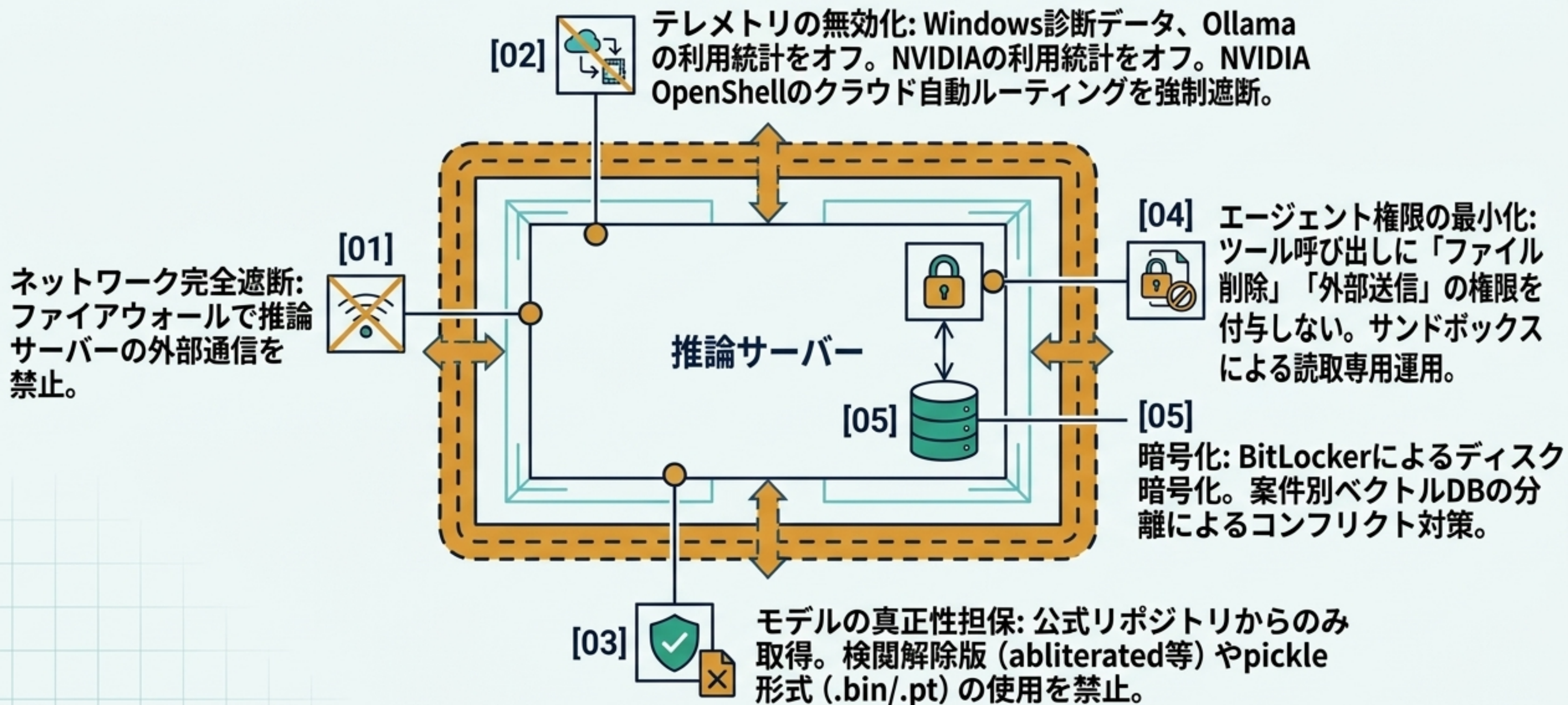
運用ルール: 防衛・経済安全保障案件においては、米国製または国産モデル（gpt-oss / LLM-jp等）へ即座に切り替えられるデュアル体制を敷く。



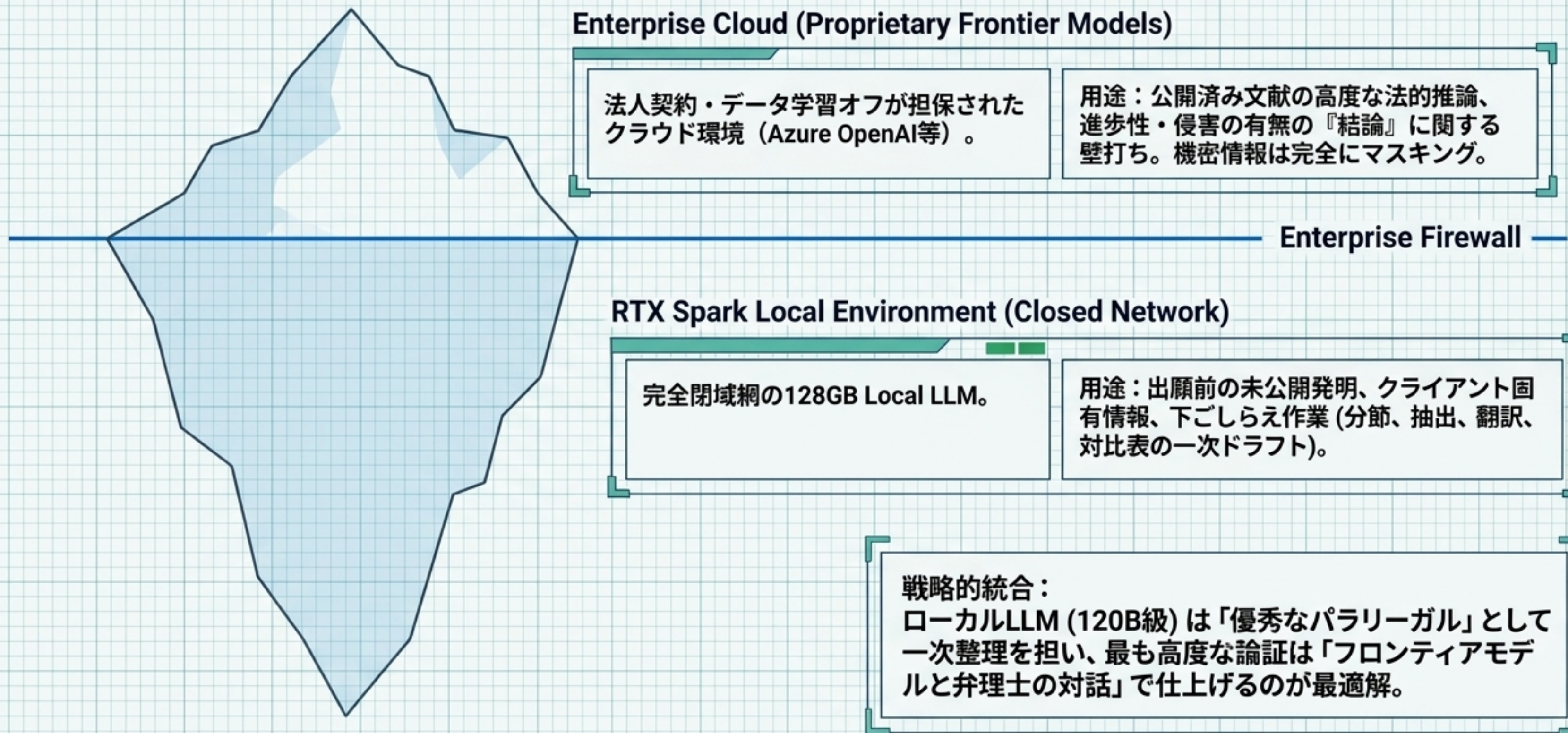
法務・倫理とガバナンス（日本弁理士会ガイドライン対応）



セキュリティとシステム設定の境界線

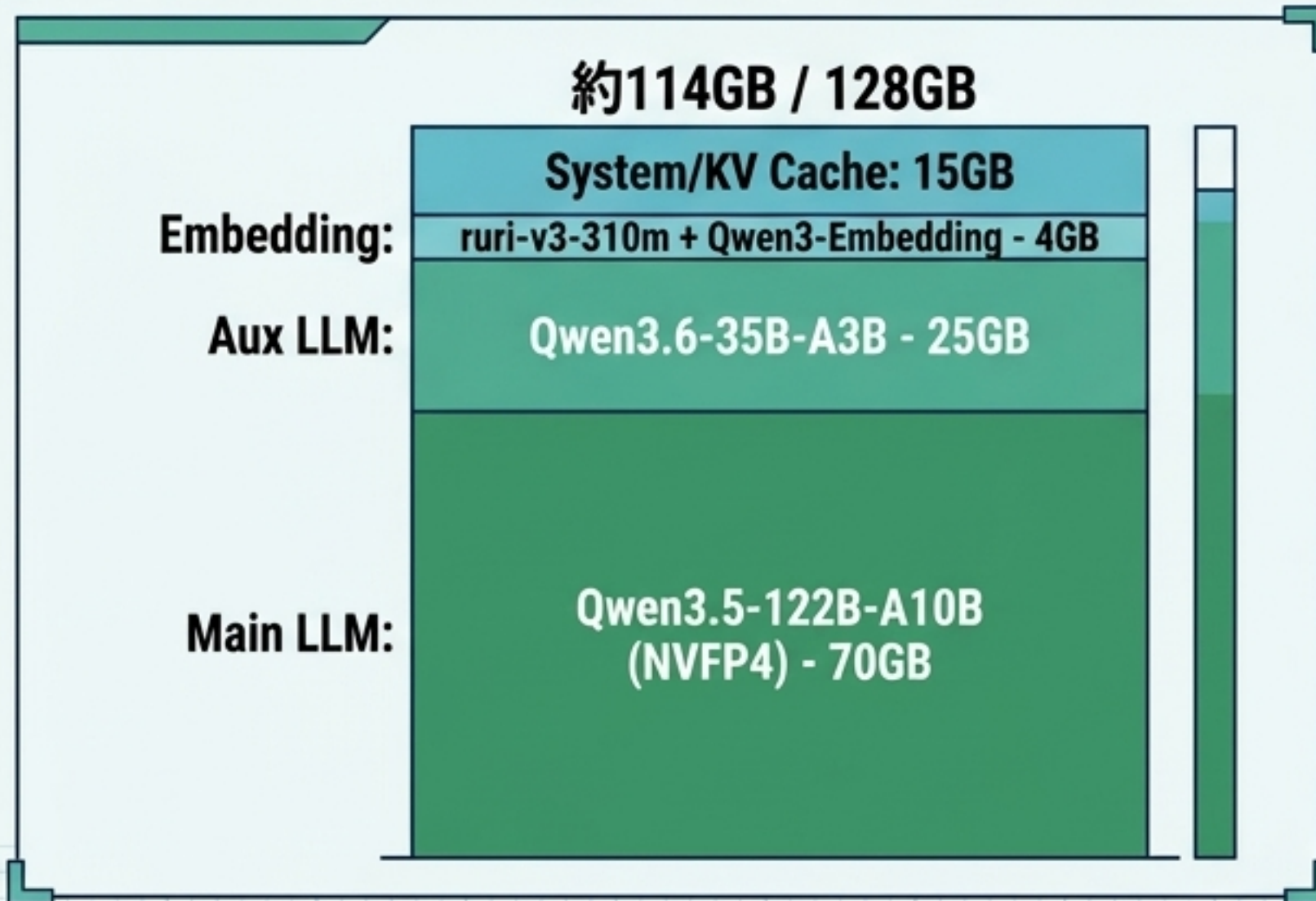


フロンティアモデル（クラウド）との使い分け戦略

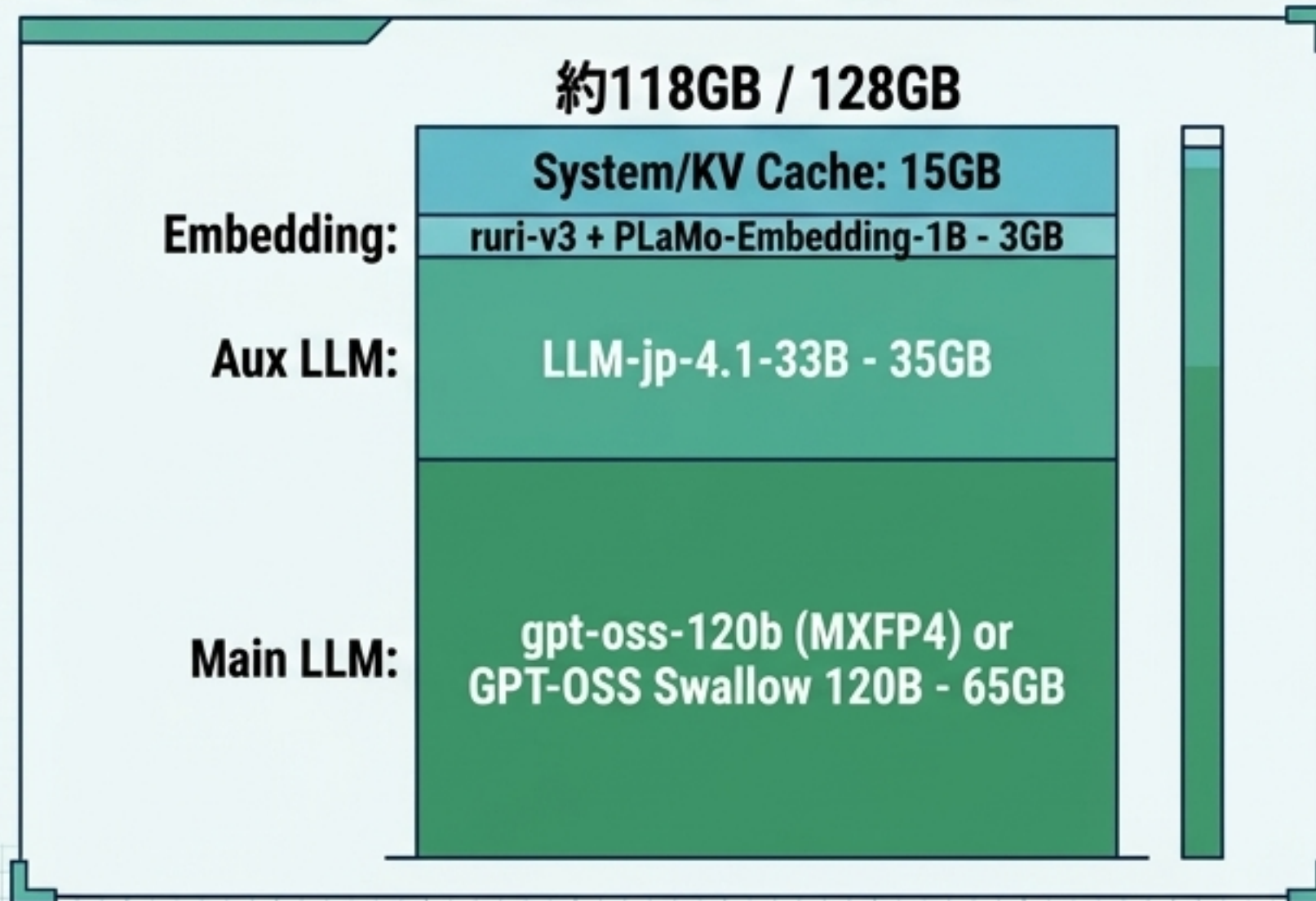


最終結論：RTX Spark向け推奨構成パッケージ

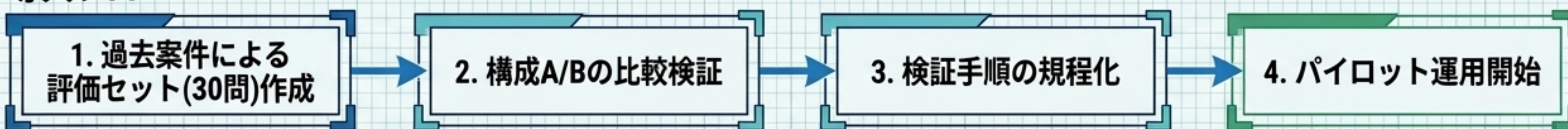
構成A：パフォーマンス最大化パッケージ



構成B：経済安全保障・非中国製パッケージ



導入フロー



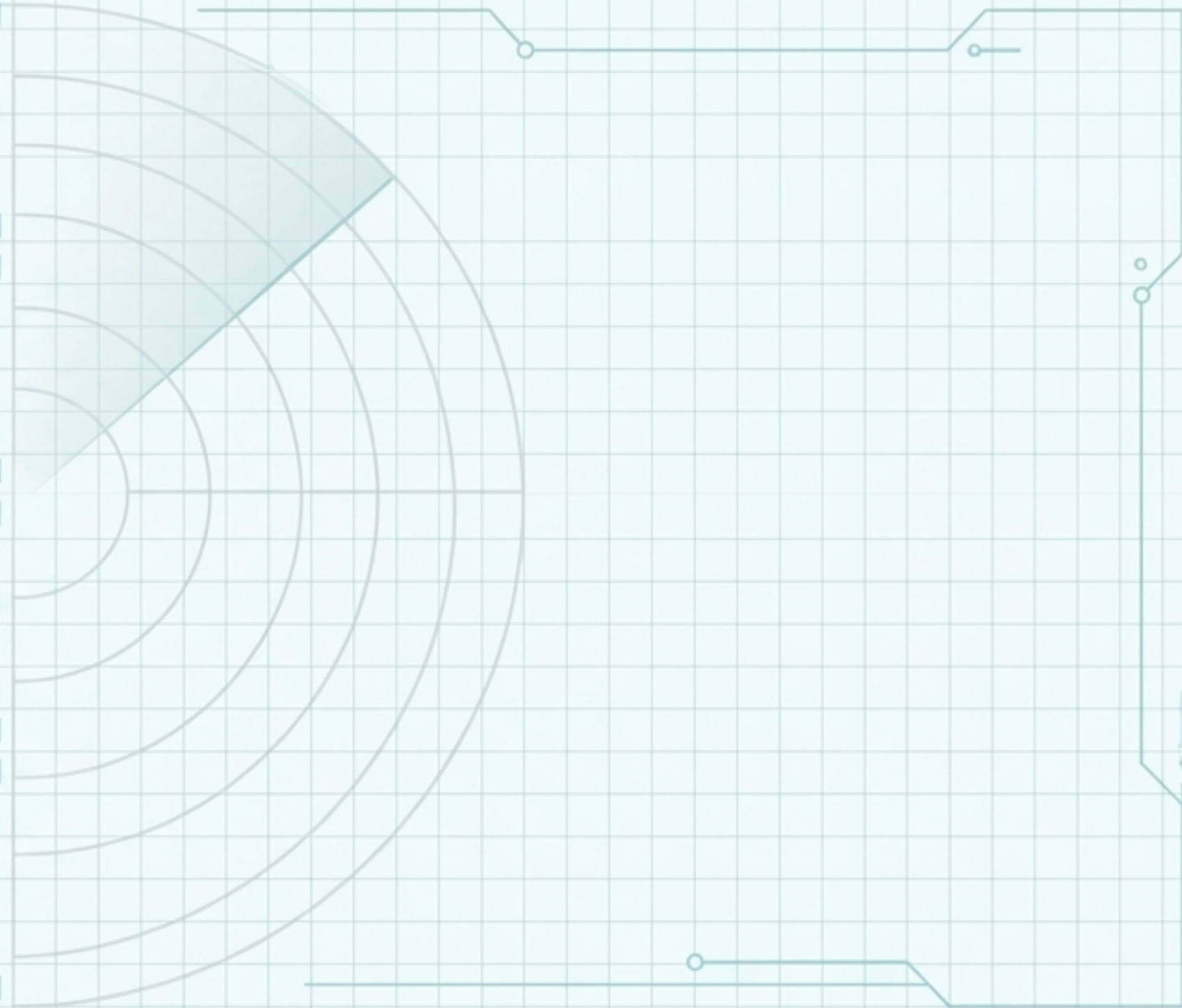
本レポートの限界と今後の監視事項 (Known Unknowns)

RTX Spark実機でのLLM生成速度の実測値（現在はDGX Sparkの数値をプロキシとして使用）。

ノートPC筐体における300GB/s帯域の持続性（サーマルスロットリングの影響）。

ネイティブLinuxの公式サポート有無、およびWindows on ArmでのvLLM / TensorRT-LLMの完全対応時期。

JPO（特許庁）またはJapioによる、日本語特許タスクにおける各モデルの直接比較ベンチマークの公開待ち。



【免責・推奨事項】

本ガイドラインは2026年10月10日時点の一次資料・検証結果に基づく。技術進化に伴うアーキテクチャの陳腐化リスクに備え、半期ごとの見直しを推奨する。