

完全オフライン 知財AI環境の 構築戦略

2026 NVIDIA RTX Spark
(128GB) 導入による実務プロセス
変革とモデル選定ガイド

Confidential / 経営層・知財戦略担当者向けプレイブック

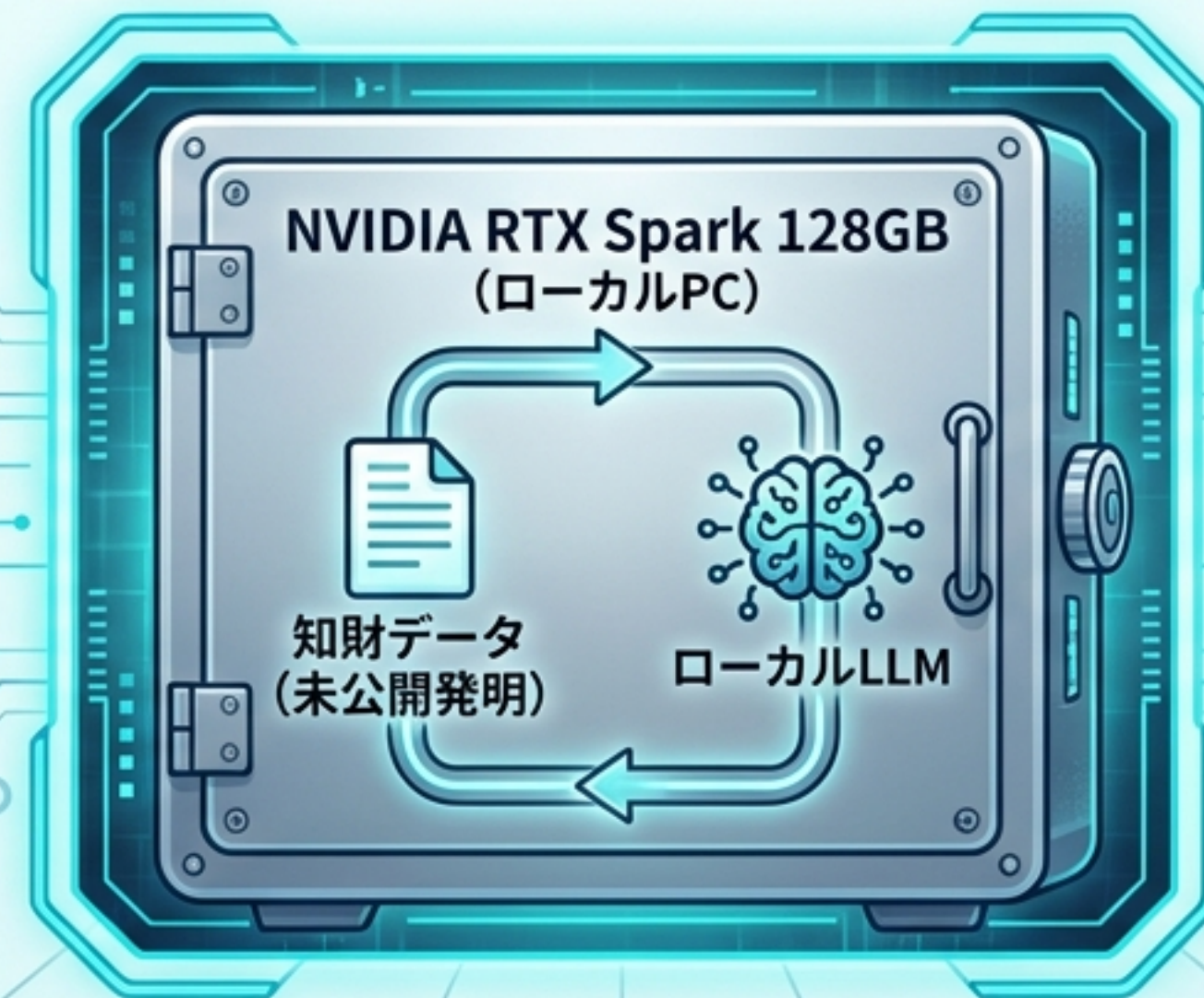


クラウドの制約を突破するハードウェアの進化

⚠ 従来環境 (Cloud LLM)



🔒 2026年ブレイクスルー (Local LLM)



従来の課題: 優れたAIが存在しても、未公開の特許情報や営業秘密を外部API・クラウドに送信することは、法務・セキュリティの観点から事実上不可能であった。

ブレイクスルー: メモリ128GBを備えたWindows PC向けスーパーチップの登場により、これまでデータセンター専用だった大規模言語モデルが、手元のPC上で完全オフライン稼働可能となった。

導入ターゲット：RTX Spark プラットフォームの技術要件

項目	RTX Spark (Target)	DGX Spark (Reference)
位置づけ	Windows PC向けスーパーチップ (OEM各社からノート/小型デスクとして発売)	NVIDIA製小型AI開発機
OS	Windows 11 (Arm)	DGX OS (Linuxベース)
CPU/GPU	20コア Grace CPU + 6,144 CUDAコア Blackwell RTX (FP4)	20コア Arm + Blackwell (FP4)
メモリ・帯域	128GB LPDDR5X, 300GB/s	128GB LPDDR5X, 273GB/s

結論：「RTX Spark」は実在するWindows PC向け製品であり、Linux前提の開発機（DGX Spark）とは異なる。一般的な社内PCとして調達・導入が可能でありながら、最大1PFLOPのAI性能をローカルで発揮する。

メモリ帯域 (300GB/s) が生成速度の絶対的上限を規定する

デコード速度上限 = [メモリ帯域 (300GB/s)] ÷ [1トークン生成に読み出す重みのバイト数]

ハードウェア帯域: 300GB/s

モデルのサイズ
(重みのバイト数)

33B Dense
(Q4: 約20GB)

上限 約15 トークン/秒
高速稼働・極めて実用的

33B Dense
(Q8: 約35GB)

上限 約8 トークン/秒
精度重視・実用的な速度

⚠ 100B Dense
(Q4: 約60GB)

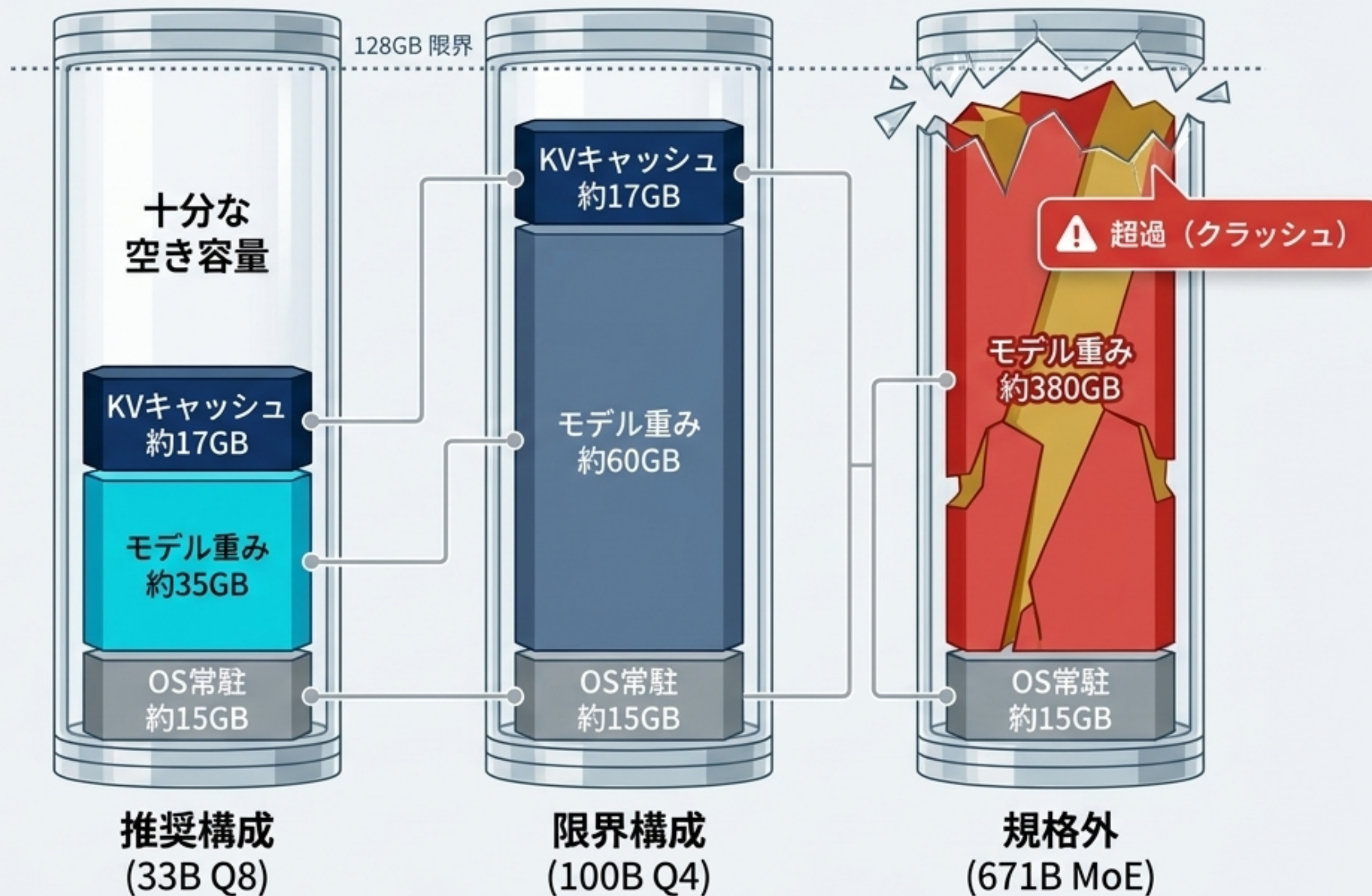
限界稼働 (数トークン/秒)
日常的な大量業務には不向き

単に「巨大なモデルが載るか」ではなく、
「実務に耐えうる速度で出力できるか」
が真の選定基準となる。

128GB メモリ・アロケーション・ゲージと規模の限界

長文処理（KVキャッシュ）の重要性

知財業務（長文明細書など）ではコンテキスト長が必須。LLM-jp-4 33Bで64Kトークンを処理する場合、約17GBの追加メモリを消費する。OSの常駐分と合わせ、余裕を持たせたモデル選びが不可欠である。

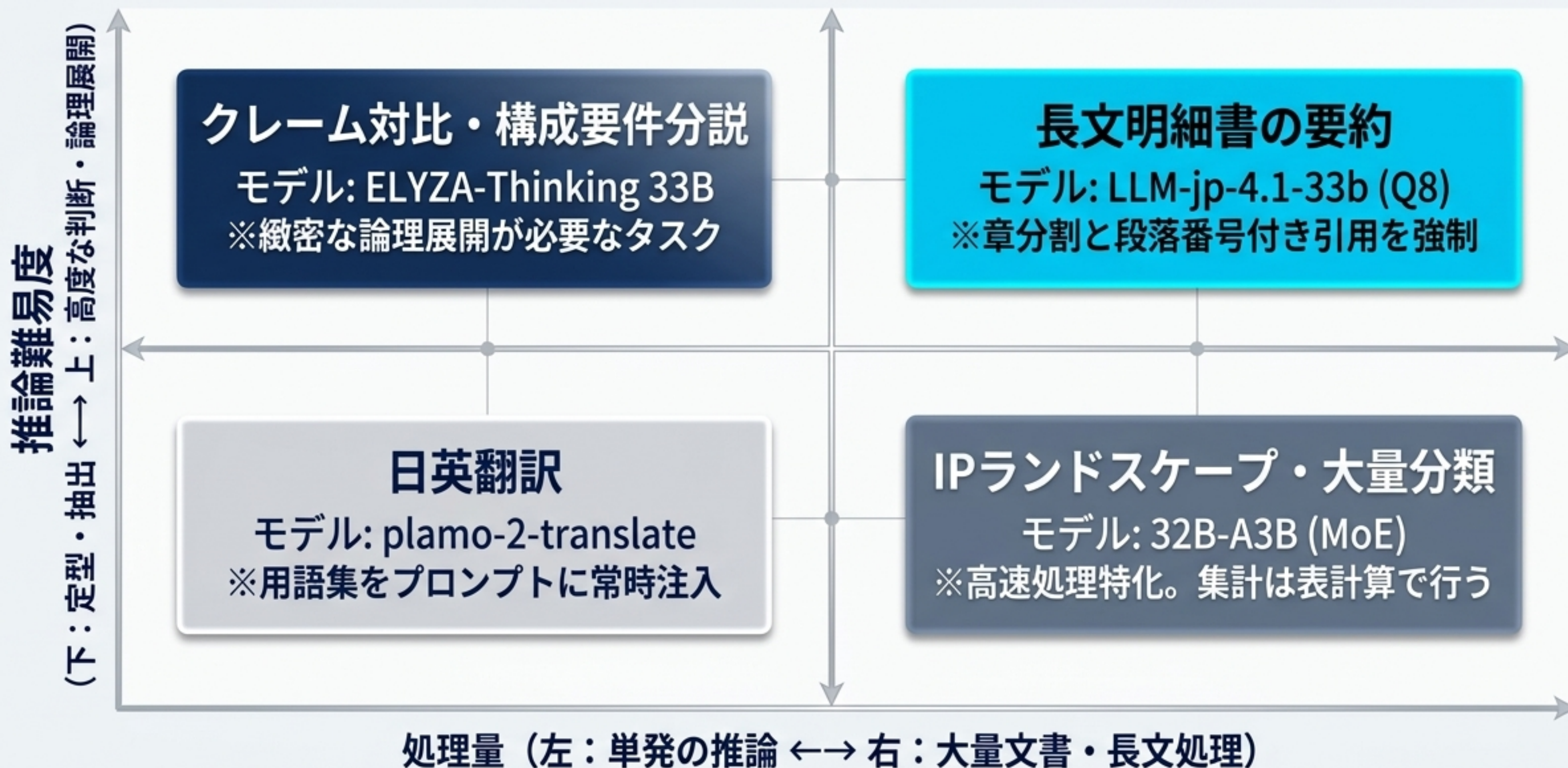


厳選：128GB環境下における国産LLM選定マトリクス

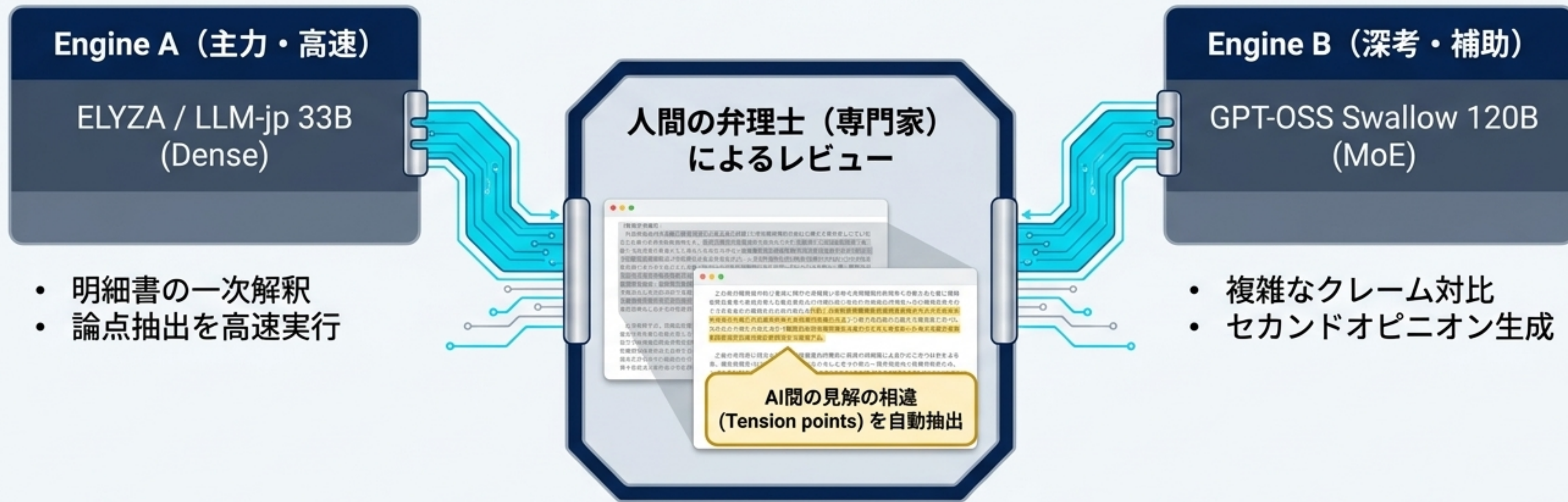
推奨度	モデル名	規模・構造	稼働環境 (128GB)	知財タスク適性とライセンス
◎ 主力	LLM-jp-4.1-33b-thinking	33B Dense	FP8/Q8推奨	学習データ公開により説明責任を満たす。主力候補。(Apache 2.0)
◎ 主力	ELYZA-Thinking 33B	33B Dense	FP8/Q8推奨	日本語の指示追従・知識想起に極めて優れる。(Apache 2.0)
○ 補助	GPT-OSS Swallow 120B	120B MoE	約60GBで実用可	難解な推論用のセカンドオピニオン。(Apache 2.0)
△ 注意	PLaMo 3 31B Base	31B Base	稼働可	指示調整なしのベースモデル。対話用途には不向き。(独自)
× 不可	Rakuten AI 3.0	671B MoE	メモリ超過	本機の128GBでは稼働不可。

意思決定ライン: 日本語性能とライセンスの明快さから、「LLM-jp-4.1系」と「ELYZA」の33Bクラスを主軸エンジンとして構築すべきである。

知財タスク別 難易度×処理量 マトリクス



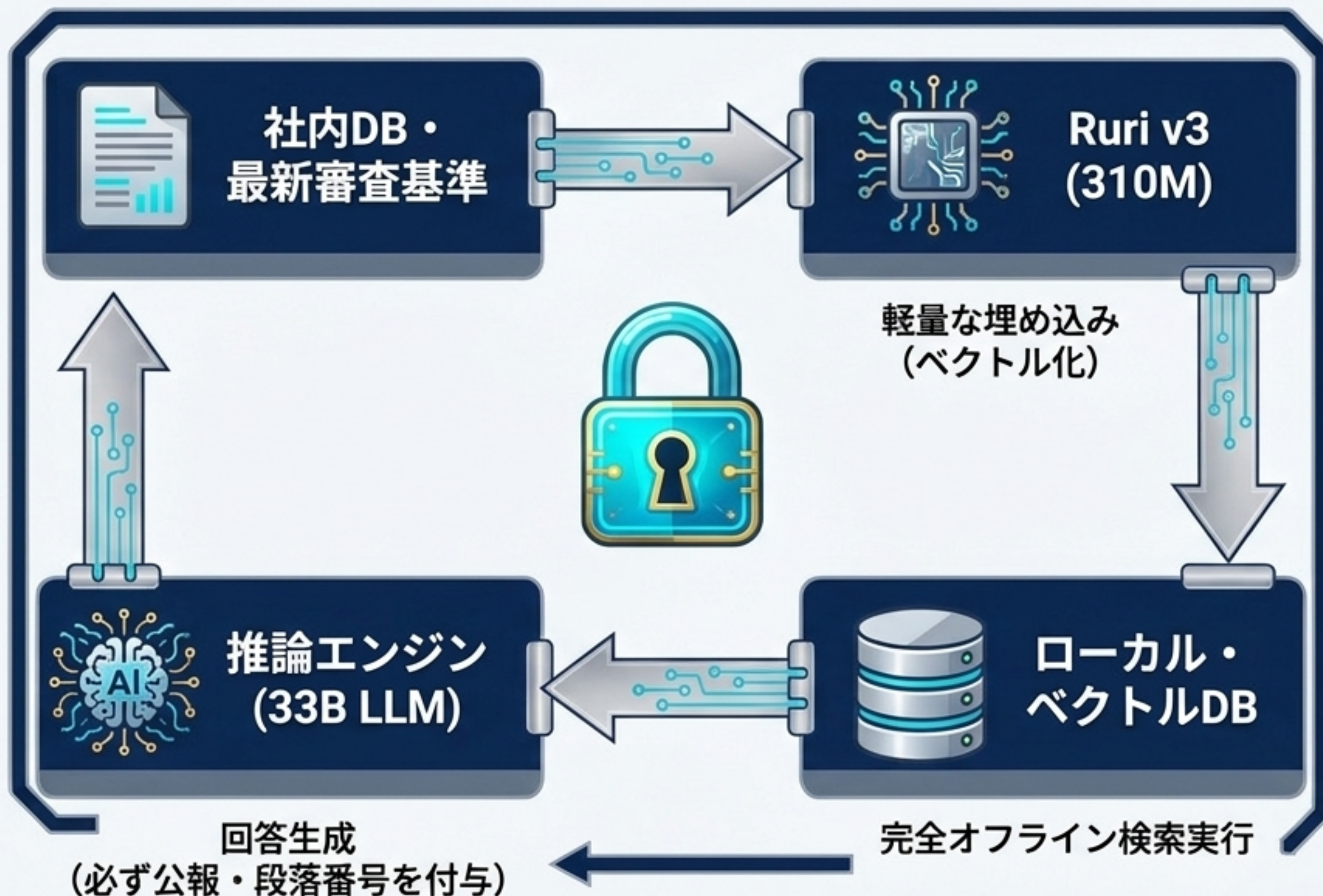
パラダイムシフト：単一モデル依存からの脱却と「ローカル・クロスチェック」



The Aha Moment

128GB環境の真の価値は、1つの巨大モデルを遅く動かすことではない。特定のタスクに最適化された30Bクラスの高速モデルを主力とし、難易度の高い判断には120Bクラスを併用する。AI同士の視点の差異を、人間の弁理士が重点的にレビューする体制こそが最も強靱なワークフローを生む。

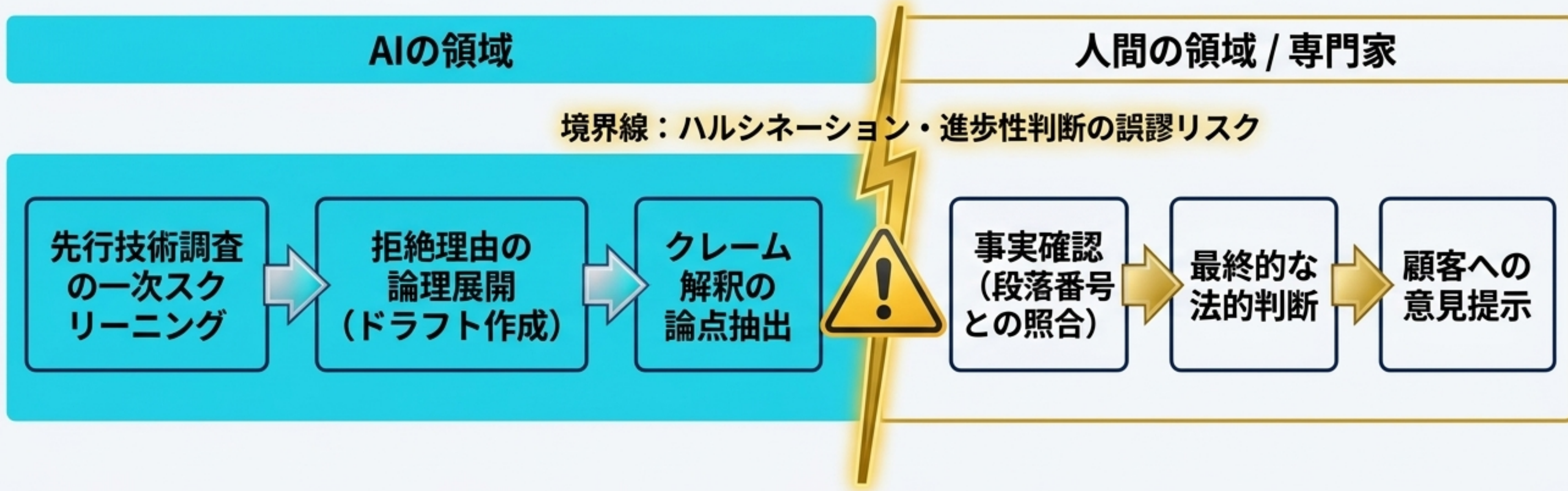
機密データを守り抜くローカルRAG（検索拡張生成）の構築



なぜRAGが必要か？

- LLM自体は最新の法改正や判例を知らない（学習時点で知識が止まっている）。
- CPUでも高速動作する国産産モデル「Ruri v3」をベクトル化の基盤とすることで、外部通信を一切行わない完全自律型の社内ナレッジベースがPC内に完成する。

ヒューマン・イン・ザ・ループ：AIは「超優秀な下書き」である



法的・実務的ガバナンス

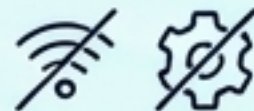
弁理士法75条（業務制限）の観点からも、AIが最終判断を下す設計は回避すべきである。LLMの出力はあくまで「壁打ち相手・論点の洗い出し」に留め、専門家（弁理士）による最終レビューを必須のワークフローとして組み込む。

ローカルLLM運用のための「4つのガードレール」

The Secure IP-AI Architecture

① 物理的・論理的遮断 (Air-Gapped)

Ollama等の外向き通信を完全拒否。
Windowsテレメトリの無効化を徹底。



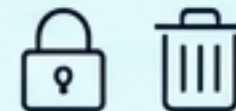
② 真正性と安全な形式 (Authenticity)

モデル取得元を公式組織に限定。
任意コード実行リスクを排除するため
GGUF/safetensors形式に限定。



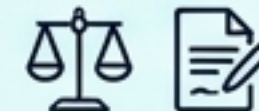
③ ログの暗号化と揮発化 (Log Management)

未公開發明を含む
プロンプト履歴の
保存場所を特定し、
BitLocker暗号化と
自動削除プロセス
を策定。

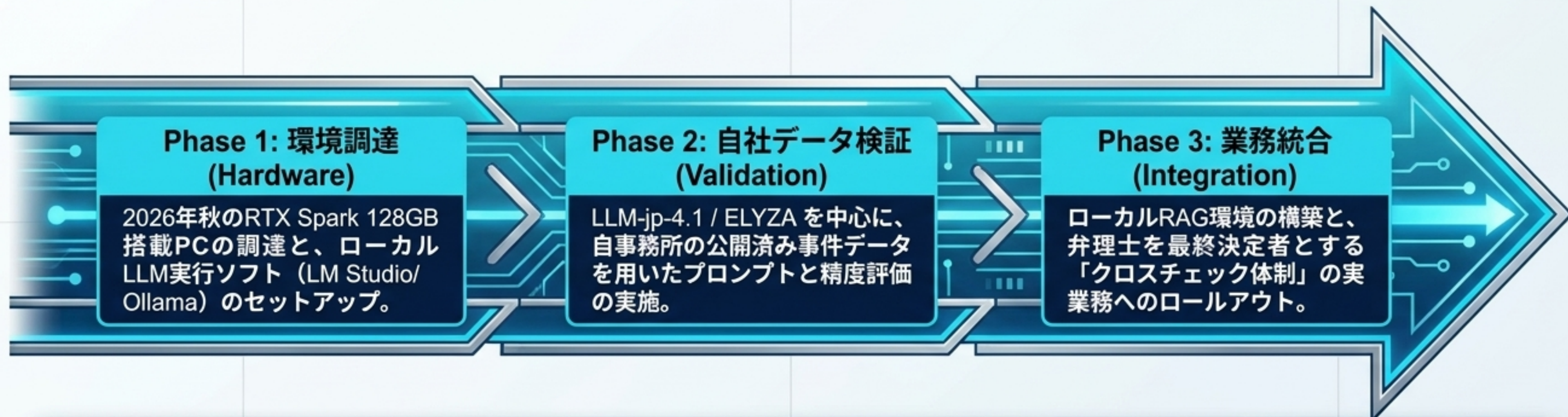


④ ライセンスと顧客同意 (Compliance)

Apache 2.0/MIT
の遵守。顧客との
NDAにおける
「第三者ツール
利用制限」条項の
事前確認。



結論とネクストアクション



「国産」であることの本質は、純粋な処理性能の優劣ではなく、学習データの透明性と『データ主権（手元から情報を絶対に出さないこと）』にある。RTX Sparkの登場は、知財業界における最も安全で強力なAI運用の扉を開いた。今すぐ検証フェーズへ移行すべきである。