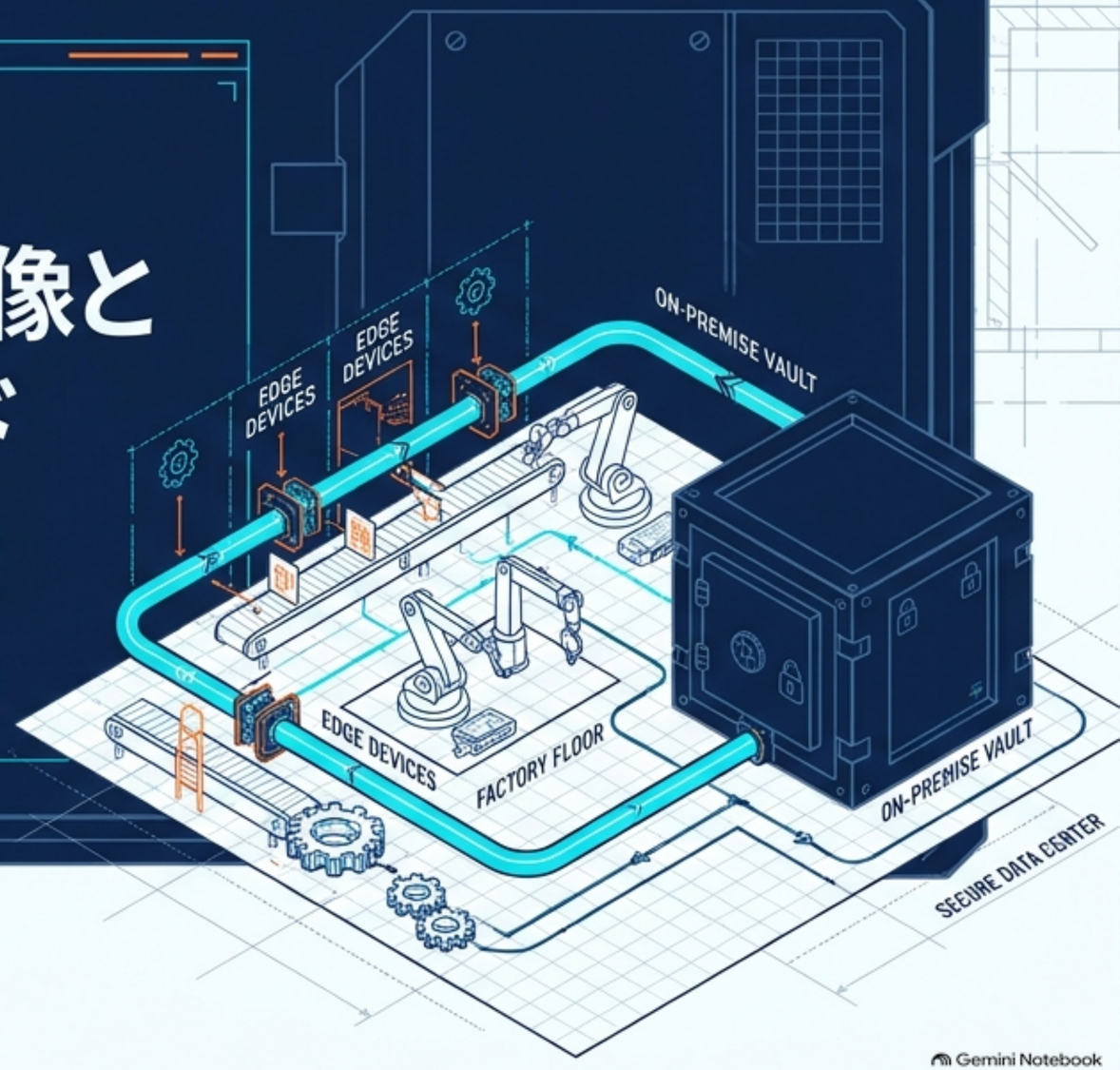
TECHNICAL & IP GOVERNANCE STRATEGY REPORT | 2026

製造業における「ローカルAI」活用の全体像とオンプレRAG実装ガイド

未公開発明・エッジ制御から、
100万円ワークステーション構築手順まで



Gemini Notebook

The Strategic Playbook：本レポートの3つの核心

パブリッククラウドLLMの試行期は終わり、製造業は**「完全閉域でのローカルAI本番実装」**へ移行した。



Trend：クラウドからローカルへの回帰

推論ワークロードの**過半（55%）**が**オンプレミス・エッジ**へ。

数千万円のサーバー室から、**100万円のAIワークステーション**への価格破壊。



Architecture：知財を守り抜く5層構造

顧客NDA図面や未公開発明を外部に出さない「**One-Way Sync**（片方向同期）と「**ニューロ・シンボリック連携**」によるハルシネーション排除。

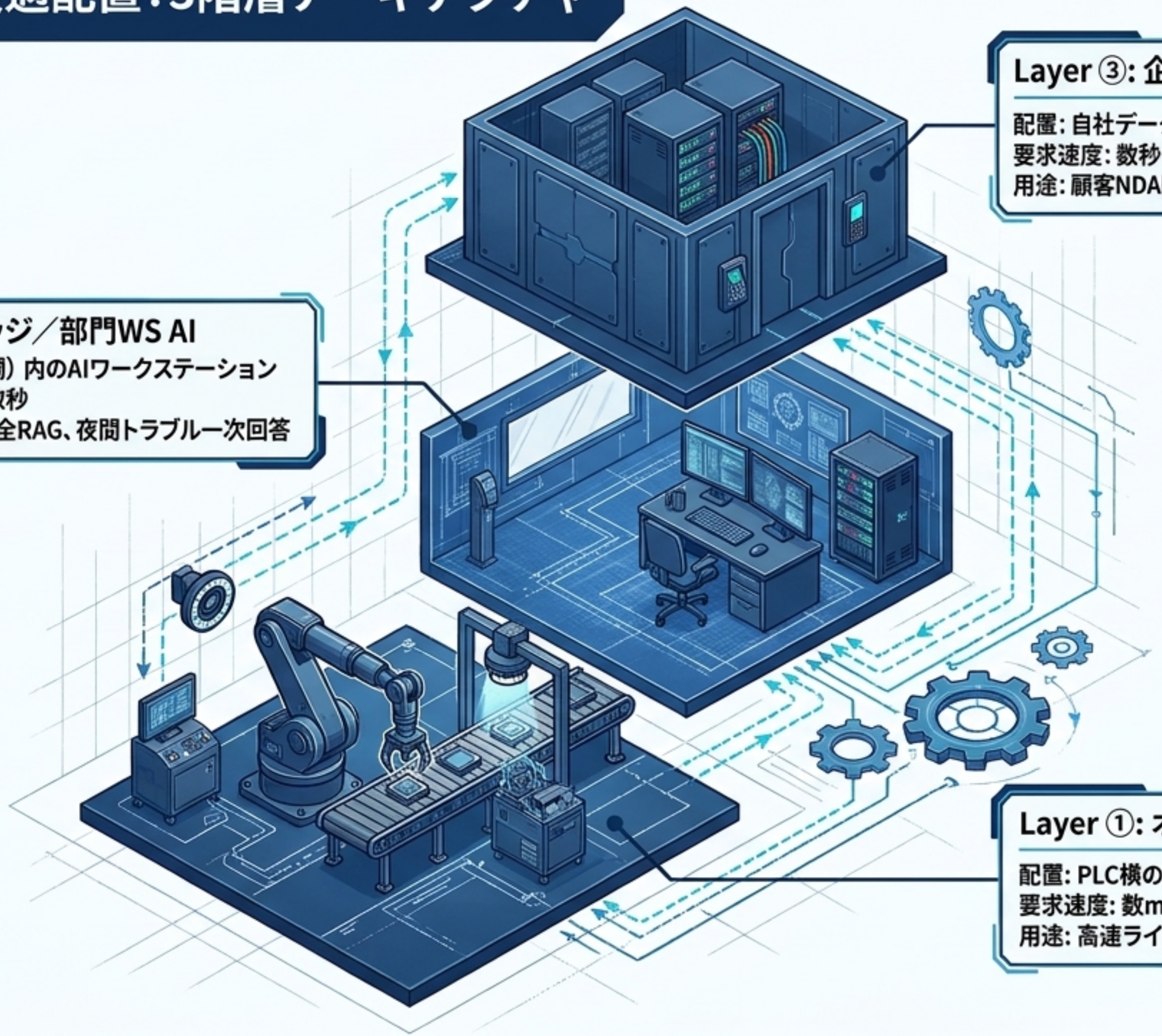


Execution：6週間・100万円でのオンプレ実装

中堅・中小企業でも実現可能な、**商用OSS**（vLLM / Dify等）を活用した**ハードウェア選定と立ち上げロードマップ**。

Gemini Notebook

製造業AIの最適配置：3階層アーキテクチャ



Layer ③: 企業内オンプレミスAI基盤

配置: 自社データセンター、R&D拠点専用GPUクラスタ
要求速度: 数秒〜数分（高度解析）
用途: 顧客NDA図面解析、製造レシピ最適化、未公開発明探索

Layer ②: 工場エッジ／部門WS AI

配置: 工場閉域網（OT網）内のAIワークステーション
要求速度: 数100ms〜数秒
用途: 設備故障診断・保全RAG、夜間トラブル一次回答

Layer ①: オンデバイス／組込みエッジAI

配置: PLC横の産業用PC、スマートカメラ
要求速度: 数ms〜40ms未満（リアルタイム制御）
用途: 高速ライン外観検査、自律ピッキング（フィジカルAI）

Gemini Notebook

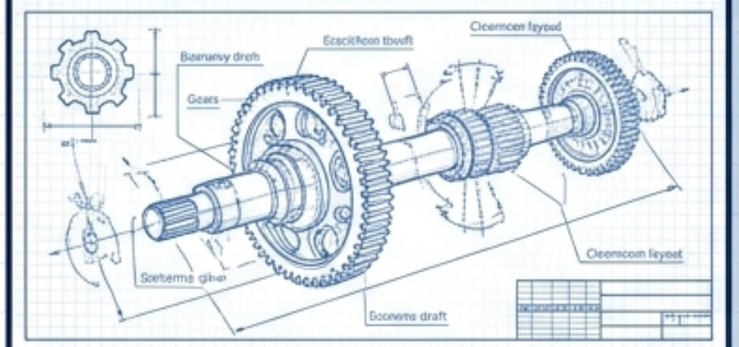
【Diagnostic Matrix】 製造業が「ローカルAI」を選ぶ5つの決定的理由

評価軸	パブリッククラウドLLM	ローカルAI（オンプレ/エッジ）
① データ主権・知財保護	✗ NDA・未公開データの送信自体が内規・契約違反リスク	✓ 外部送信ゼロ。営業秘密・レシピを完全保護
② レイテンシ（遅延）	✗ クラウド往復に約1.5秒。リアルタイム制御不可	✓ エッジ推論で40ミリ秒未満（97%短縮）
③ TCO・コスト最適化	✗ 24時間監視ではAPIトークン従量課金が無限に膨張	✓ 100万トークンあたり最大18倍安価（償却資産化）
④ BCP耐性・OT網分離	✗ ネット遮断時・災害時に現場AIが停止	✓ インターネットから遮断されたOT網内で完結稼働
⑤ 品質・再現性の担保	✗ 勝手なモデル更新で出力傾向・フォーマットが変化	✓ モデルウェイトを固定し、自社の品質管理サイクルで運用

Gemini Notebook

業務領域別のローカルAI実装事例（最前線のランドスケープ）

① 設計・開発・技術伝承



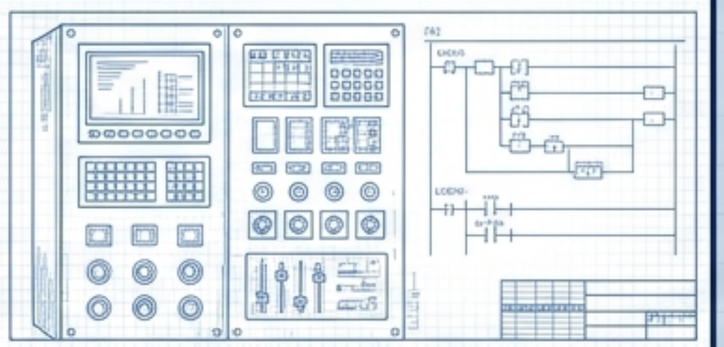
① 設計・開発・技術伝承

（オンプレRAG × SLM）

ダイフク：顧客機密が極めて高い半導体搬送システム開発に、閉域稼働型のオンプレLLM＋RAGを導入。

IHI / パナソニック：自社専用LLM（SLM）を活用し、機密技術文書を社内に閉じたまま解析・活用。

② 設備保全・PLC制御



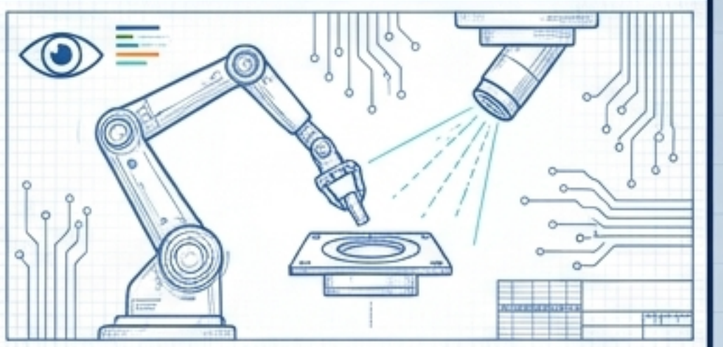
② 設備保全・PLC制御

（OT閉域網 × 保全エージェント）

ダイキン工業：設備図面と保全記録から「故障診断AI」を構築。10秒以内・90%の精度で原因と対策を回答。

キーエンス：PLC専用AI「KV Intelligence」で、ラダー生成と波形時系列比較（ディープデバッグ）を閉域保護下で実現。

③ 外観検査・フィジカルAI



③ 外観検査・フィジカルAI

（VLM × 自律制御）

シーメンス：エッジAI実行を25倍高速化。生産計画コスト最大20%削減、仮想立ち上げ工数85%削減。

日本精工（NSK）：生成AIで不良品画像を合成し、少数サンプルからエッジ検査AIを短期立ち上げ。

Gemini Notebook

製造業特有の「3つの壁」と導入成功の要諦

壁① データ構造化の壁
（三面図、結合Excel、現場隠語）

VLM/専用OCRによる前処理と、熟練者へのAIインタビューによる「暗黙知の言語化」に投資する。

壁② ハイブリッド線引きの壁
（どこまで閉域にするか）

NDA図面・未出願発明・保全ログは「完全ローカル」。
公開特許やSC情報は「閉域クラウド」と明確に使い分ける。

壁③ 耐環境管理と知財ガバナンス
（粉塵環境・著作権抵触）

高温・粉塵対策としての産業用PC選定。
納入装置のPLCプログラムをAIに読ませる際の、著作権・営業秘密条項の事前クリアランス。

Gemini Notebook

R&D・知財におけるクラウドAIのリスク（オプトアウトの限界）

「学習利用オフ（オプトアウト）」であっても、外部サーバーへの送信自体が以下の致命的な法的リスクを引き起こす。



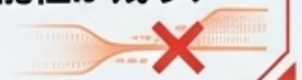
漏洩するクラウド（データ流出）



セキュアな閉域環境（完全保持）

1. 特許法第29条（新規性喪失）のグレーゾーン

不正利用監視等のためプロバイダ側にプロンプトが一定期間保持される規約では、外部エンジニアの閲覧可能性が残り、未公開発明の新規性を喪失するリスクがある。



2. 不正競争防止法上の「秘密管理性」の揺らぎ

ノウハウ秘匿（先使用権確保）領域において、外部クラウド送信自体が秘密管理体制の不備として争われる原因となる。



3. 共同研究・顧客NDAの「第三者送信禁止」

大学との共同研究データや預かり図面は、第三者（クラウドベンダー）サーバーへのアップロード自体が一律禁止されている。



Gemini Notebook

【Architectural Deep-Dive】知財・R&D向け 5層リファレンス・アーキテクチャ

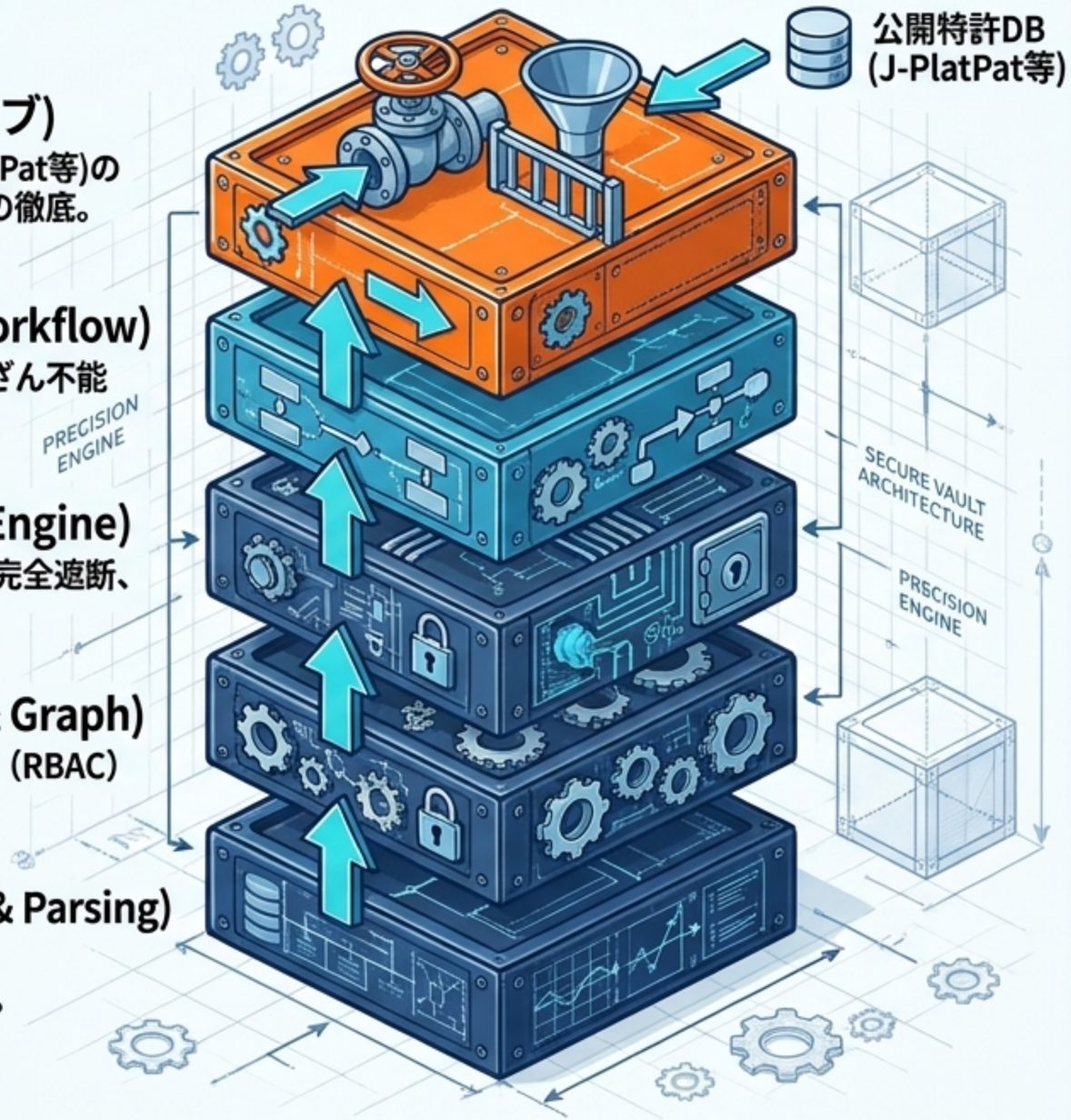
⑤ 外界連携層 (One-Way Sync バルブ)
生データは絶対に出さず、公開特許DB(J-PlatPat等)の情報を内へ引き込む「片方向データフロー」の徹底。

④ エージェント実行層 (IP / R&D Workflow)
発明発掘、先行技術対比。全入出力ログを改ざん不能ストレージへ保存（発明者証跡）。

③ ローカル推論基盤層 (Inference Engine)
閉域GPUでの高速並列推論。外部通信ポート完全遮断、モデルウェイトのハッシュ検証・固定。

② 知識格納・検索層 (Hybrid RAG & Graph)
プロジェクト単位の属性ベースアクセス制御（RBAC）を強制付与したベクトル・グラフ検索。

① データ収集・前処理層 (Ingestion & Parsing)
外部API不使用。ローカルVLMによる図表・化学構造式OCRとMarkdown/JSON構造化。



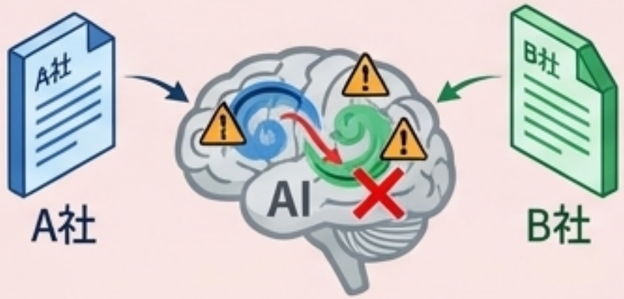
Gemini Notebook

内部ガバナンスの落とし穴： 見落とされがちな「チャイニーズ・ウォール」

致命的な落とし穴

RAG汚染（コンタミネーション）の危険性

A社由来の共同研究データを無防備な共有RAGに入れると、B社担当者のAI回答にA社の機密が混入し、契約違反に発展する。



自然人発明者の要件

AIの出力と人間の創作寄与が混ざり、特許法上の「発明者」が特定できなくなる。

必須の内部統制アクション

- **RBAC (属性ベースアクセス制御) の徹底:** NDAデータはファインチューニングには絶対に使わず、RAGの検索段階で所属外のデータを完全遮断する。
- **自然人の寄与記録:** 入出力履歴をユーザーID・タイムスタンプ付きで発明届出書に紐付け保存する。
- **オープンモデルの出所検証:** 商用・出力制限のないクリーンなモデル（Apache 2.0等）を選定し、閉域搬入前に脆弱性検査を実施。

Gemini Notebook

【Synthesis】ハルシネーションを排除する「ニューロ・シンボリック連携」

特許明細書の実施例生成において、ハルシネーションをゼロにする最大の鉄則は「ローカルLLMに生データを直接読ませて、計算や選別をさせない」ことである。



Gemini Notebook

【Explanatory Visual】ELN実測データの決定論的仕分けと監査

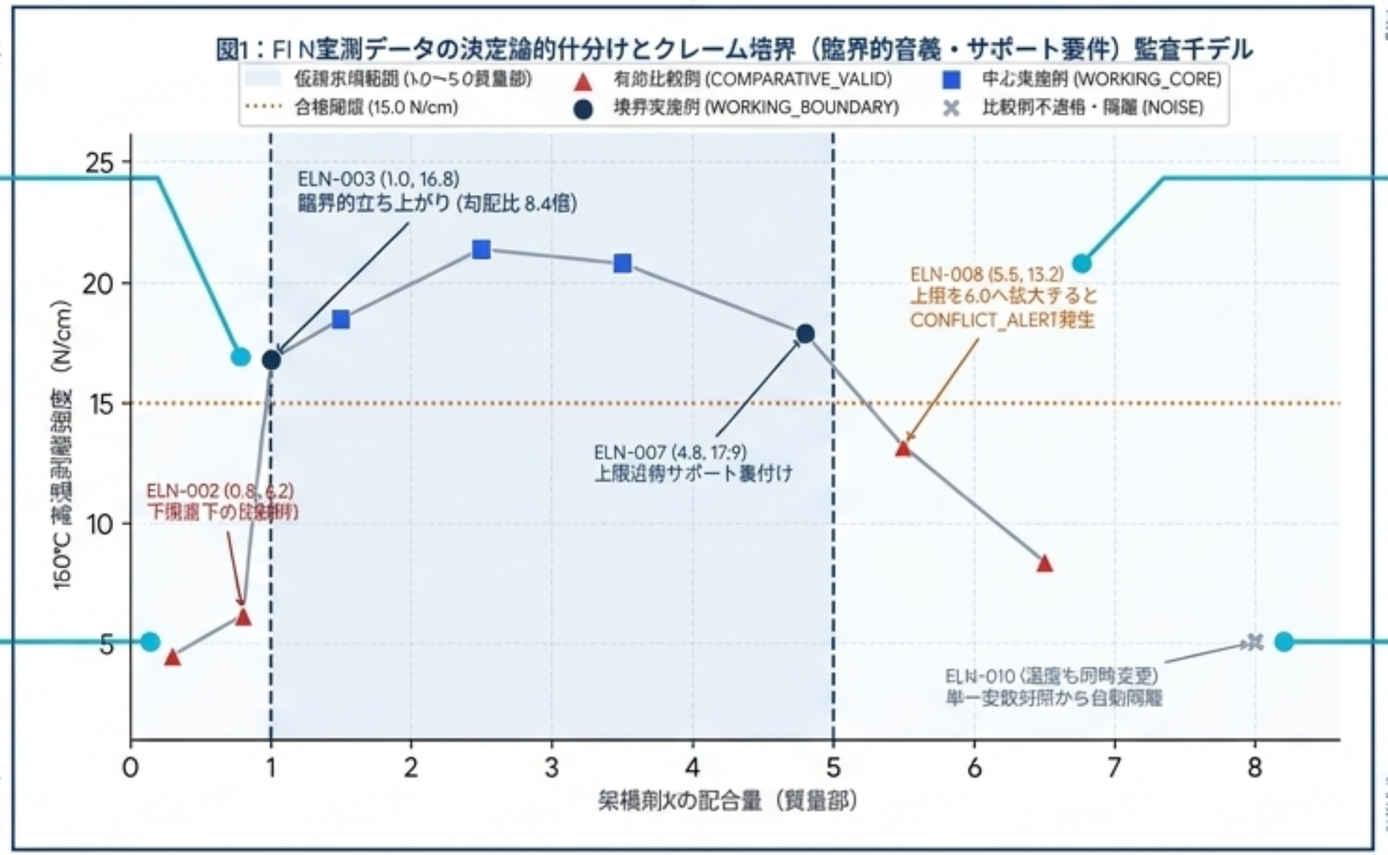
ローカルLLMへ渡す前に、Pythonが仮請求項と実測データを照合し、法的な「ロール（役割）」を自動付与する。

境界実施例

クレーム上限・下限近傍。サポート要件の裏付けとなる最重要データ。

有効比較例

クレーム範囲外かつ特定パラメータ以外は実施例と同一。
数値限定の「臨界的意義」を立証する。



重大警告

クレーム範囲内なのに効果未達。実施可能要件違反を防ぐため減縮を検討。

自動隔離

複数条件が同時に変更されたデータは、比較例として法的に無効なため排除。

Gemini Notebook

LLMの暴走を防ぐガードレール：3段階プロンプトチェーン

単一の巨大プロンプトは条件取り違えを起こす。必ず以下の3段階に分割し、文法レベルで出力を拘束する。

Stage 1: 監査 (Audit)

【絶対遵守事項】推測・補完の厳禁。上下限サポートの空白領域や、比較例の妥当性を「JSON内の事実のみ」から検証し指摘する。

Stage 2: 執筆 (Drafting)

[絶対遵守事項] 完全転写原則。数値を1文字も変更・四捨五入しない。欠損値(null)を勝手に補完せず、[要確認]プレースホルダをそのまま出力する。

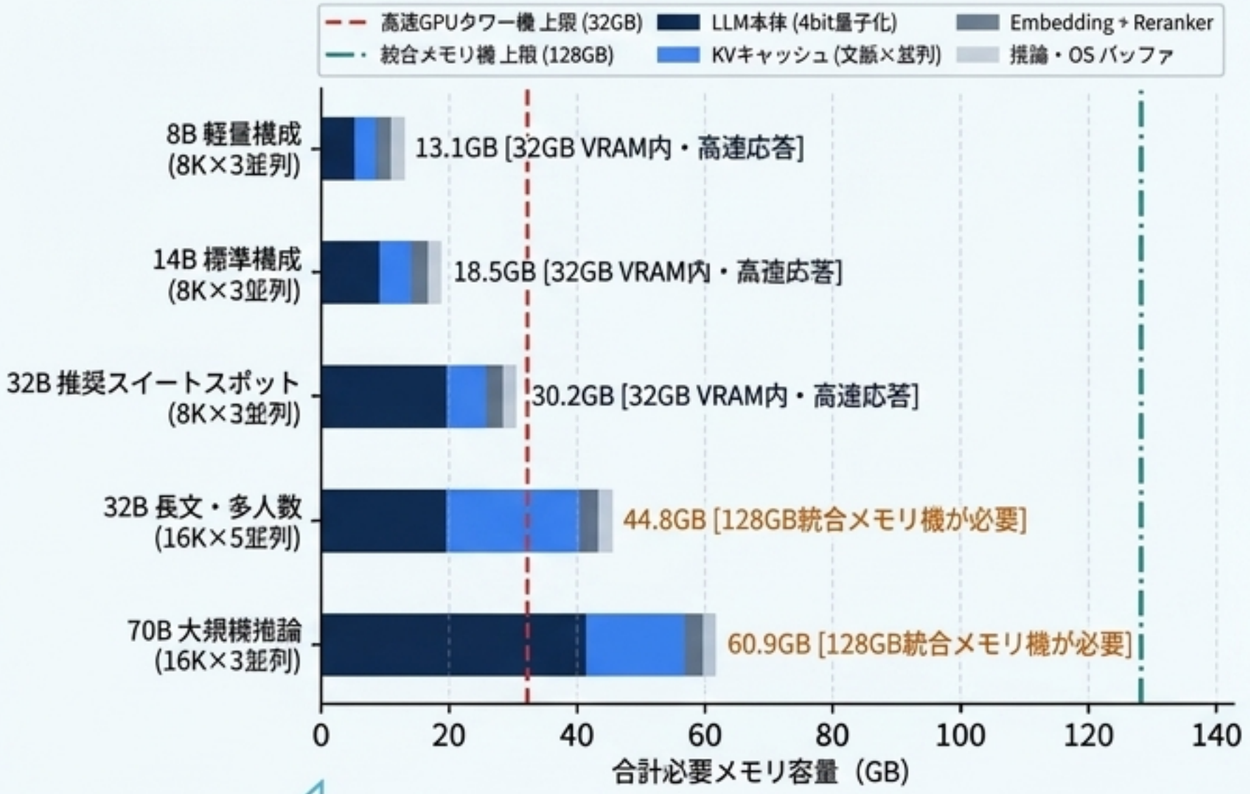
Stage 3: 考察 (Analysis)

[絶対遵守事項] 断定表現の禁止。作用機序には必ず「〜と推察される」とし、権利解釈を限定しない免責定型文を一字一句違えず挿入する。

Gemini Notebook

予算100万円で実現するハードウェアとOSS技術スタック

数千万円のサーバー室は不要。32GB VRAM搭載の高速GPUタワー機（約85〜120万円）とライセンス費用ゼロの標準スタックで構築可能。





[VRAM最適配分 (Total 32GB)]
- 回答生成LLM (14B-32B 4bit量子化)：約12〜19.5GB
- 検索用Embedding + Reranker: 約2.5GB

※推論精度を劇的に上げるため、Rerankerの併用が必須。

Gemini Notebook

オンプレRAG 本番稼働への「6週間ロードマップ」

Week 1: 対象選定・環境手配

【WARNING】
最初から全社文書を入れない。過去トラブル対策書など「1テーマ」に絞る。

Week 2: 基盤構築・閉域搬入

【WARNING】
ネット遮断前に全コンテナとモデルを取得し、自動更新をオフにしてから固定IP接続する。

Week 3: 製造文書クレンジング

【WARNING】
精度を分ける最重要工程。Excelのセル結合を解除し、チャンク先頭に「型番・改訂日」を強制付与。

Week 4: ハイブリッド検索実装

【WARNING】
型番の1文字違いを拾うためハイブリッド検索有効化。「記載がない事項は該当なしと答えよ」とプロンプト固定。

Week 5: 通信遮断・権限分離

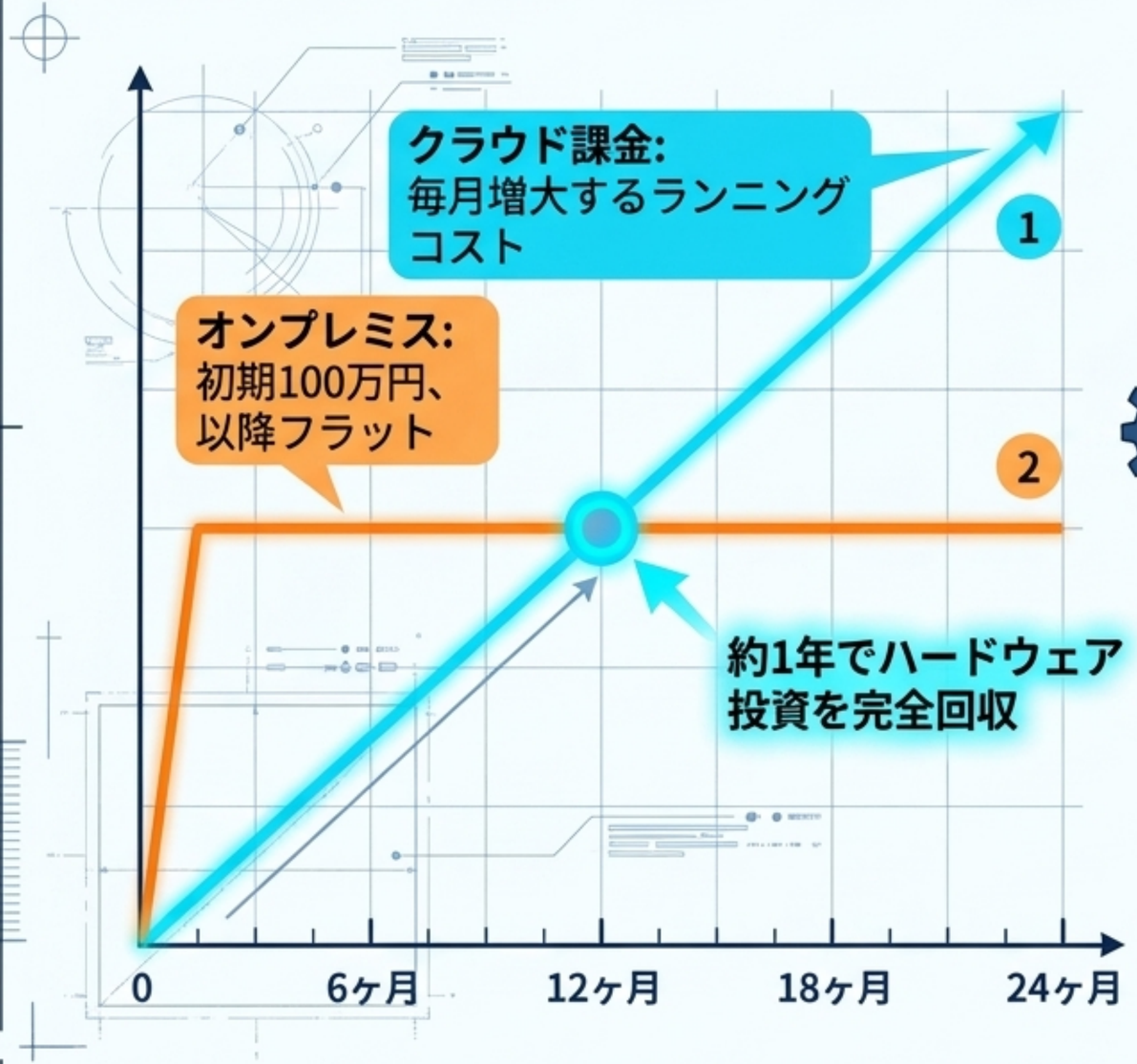
【WARNING】
パケット監視で外部通信試行ゼロを確認。部門別・NDA別にRBAC（閲覧権限）を分離。

Week 6: 50問実機検証・定着

【WARNING】
誤答の大半は元手順書の古さ・暗黙知の未記載。Q&Aシートで追補し正答率85%超へ。

Gemini Notebook

ROI（投資対効果）と総括：製造業ローカルAI成功の要件



ローカルAI実装の本質は「クラウドの単なる置き換え」ではない。

1. **最適配置:** 現場の100万円WSからR&D閉域GPUまで、機密性と遅延に応じた物理的隔離。
2. **データ構造化:** 結合ExcelやELN生データを、AIが誤読しないFact-Locked JSONへ前処理。
3. **決定論的ロック:** 数値判定やアクセス権限はPythonでロックし、LLMには言語化のみを担わせる。

【CONCLUSION】
このアーキテクチャにより、営業秘密保護・知財品質向上・現場技術伝承の3つが同時に実現する。

Gemini Notebook