製造業ローカルAI活用・オンプレRAG構築ガイド

製造業におけるローカルAIの3階層アーキテクチャ

③企業内オンプレミスAI基盤
拠点内GPUで32B〜100B超モデルを用い未公開発明や顧客図面を解析。




②工場エッジ／部門WS AI
OT閉域網内のワークステーションで8B〜32Bモデルを稼働し保全支援。





①オンデバイス／組込みエッジAI
ロボットやPLC直結で40ms未満の超低遅延制御と高速外観検査を実現。




予算100万円クラスハードウェア構成比較



高速GPUタワー機

利用可能メモリ：
32GB VRAM + 128GB RAM

最適な現場用途：
応答速度優先（10〜30名の現場対話・高速OCR）



統合メモリ小型AI機

利用可能メモリ：
128GB 統合メモリ

最適な現場用途：
省電力・小型筐体で70B級大型モデルの検証対応



静音省電力Mac機

利用可能メモリ：
128GB 統合メモリ

最適な現場用途：
事務所内での設計・知財部門による少数精鋭利用

6週間で本番稼働させるオンプレRAG構築手順

W1: 対象選定・手配



トラブル集など1テーマに絞り機器手配。

W2: 基盤構築・閉域化



オフライン設定完了後に社内LAN接続。

W3: 文書整形・前処理



Excelセル結合解除と現場辞書登録。

W4: 検索機能実装



ベクトル・キーワードの複合検索設定。

W5: 遮断検証・権限分離



外部通信遮断の確認と部門別閲覧制限。

W6: 実機検証・運用開始



50問テストで正答率85%超へ引き上げ。

Gemini Notebook