TECHNICAL & IP GOVERNANCE STRATEGY REPORT ｜ 2026 年 11 月

製造業における「ローカル AI」活用の全体像と知財・R&D データ連携およびオンプレ RAG 実装ガイド

― 製造現場エッジから未公開発明・ELN 実測データの明細書ドラフティング、100 万円ワークステーション構築手順まで ―

Gemini 4 Carbon

【実務の要点】エグゼクティブ・サマリー（本レポートの構成と核心）

2025〜2026 年にかけ、製造業の AI 活用は「パブリッククラウド LLM の試行」から「営業秘密・顧客 NDA 図面・未公開発明を一切社外へ出さず、現場や開発拠点の閉域網で運用する『ローカル AI』の本番実装」へと移行している。

本レポートでは、これまでの検討結果を 4 つの章に体系化して収録した。

・第 1 章：製造業におけるローカル AI 急拡大の背景、3 階層アーキテクチャ、国内外の最新事例と導入の壁

・第 2 章：R&D・知財部門で未公開発明・実験データを安全に扱う 5 層リファレンス・アーキテクチャと内部統制

・第 3 章：電子実験ノート（ELN）実測データからハルシネーションを完全排除して特許明細書の実施例・比較例・臨界的意義を生成するニューロ・シンボリック連携と 3 段階プロンプト設計

・第 4 章：中堅・中小製造業が 100 万円クラスの AI ワークステーションで 6 週間で立ち上げるオンプレ RAG 構築手順

第 1 章　製造業における「ローカル AI」活用の全体像とマクロ潮流

製造業における「ローカル AI」とは、単一のソフトウェアを指すものではなく、①オンプレミス LLM（自社サーバーや部門ワークステーション上の生成 AI）、②工場エッジ AI（産業用 PC やゲートウェイでのリアルタイム推論・視覚言語モデル VLM）、③オンデバイス／フィジカル AI（ロボット・検査機・PLC 直結の自律制御）の 3 領域を包含する実践的な実装概念である。

■ 1.1 2026 年にローカル AI が急拡大している 3 つの構造的要因

- **推論ワークロードの過半がオンプレミス／エッジへ移行：**企業の AI 推論に占めるオンプレミスおよびエッジ実行の比率は、2023〜2024 年頃の 12〜30%水準から 2026 年には約 55%へと急伸しており、とりわけ製造業が最も速いペースでローカル移行を牽引している。
- **ハードウェアの価格破壊（数千万円のサーバー室から 100 万円前後の AI ワークステーションへ）：**従来は実用的な社内生成 AI 基盤に数千万円規模の GPU サーバーが必要だったが、NVIDIA DGX Spark や大容量メモリ搭載ワークステーション等の登場により、従来の約 1/20 の初期費用で 8B〜32B（さらには 70B）級のモデルを現場・部門単位で運用可能となった。
- **小規模言語モデル（SLM）とオープンウェイトモデルの成熟：**用途を設計文書照会、保全記録の検索・要約、特定フォーマット変換に絞れば、量子化・蒸留技術を用いた 7B〜32B クラスの小型モデル＋RAG 構成で、クラウド上の巨大フロンティアモデルに匹敵・凌駕する実務精度を達成できるようになっている。

■ 1.2 製造業におけるローカル AI の 3 階層アーキテクチャ

データの発生源、許容される応答遅延（レイテンシ）、および機密レベルに応じ、製造業のローカル AI は以下の 3 階層で最適配置される。

階層区分	配置場所・ハードウェア	主なモデル・技術	要求応答速度	代表的なユースケース
① オンデバイス／組込みエッジ AI	ロボット制御盤、スマートカメラ、PLC 横の産業用 PC（IPC）	軽量 CNN、エッジ VLM、VLA（視覚・言語・行動）モデル	数 ms〜40ms 未満（リアルタイム制御）	高速ライン外観検査、自律ピッキング（フィジカル AI）、異常検知・即時停止
② 工場エッジ／部門 WS AI	工場閉域網（OT 網）内の AI ワークステーション、エッジサーバー	8B〜32B 級ローカル LLM、マルチモーダル VLM＋ローカル RAG	数 100ms〜数秒（対話・常時監視）	設備故障診断・保全 RAG、夜間トラブル一次回答、カメラ映像の VLM 状況把握
③ 企業内オンプレミス AI 基盤	自社データセンター、R&D 拠点内の専用 GPU クラスタ	32B〜100B 超の企業専用 LLM（継続事前学習・FT 対応）	数秒〜数分（高度解析・バッチ）	顧客 NDA 図面・仕様書解析、製造レシピ最適化、未公開発明探索・技術伝承

■ 1.3 製造業がクラウドではなくローカル AI を選ぶ 5 つの決定的理由

- **1. 営業秘密・製造レシピ・顧客 NDA 図面の完全保護（データ主権）：**稼働パラメータ（製造レシピ）、未公開実験データ、不良解析データ、さらに自動車・半導体・防衛産業等で取引先と締結する厳格な NDA 対象の預かり図面は、学習オプトアウト設定であっても外部クラウド送信自体が契約・内規で禁じられているケースが多い。欧州 NIS2 指令やサイバーレジリエンス法（CRA）等の規制強化もローカル実装を後押ししている。
- **2. 超低遅延（レイテンシ解消）と通信帯域の限界突破：**クラウド往復で約 1.5 秒を要する推論も、現場エッジ推論なら 40 ミリ秒未満へと約 97%短縮できる。高精細カメラの常時監視映像や振動・電流波形をすべてクラウドへ送る帯域負荷を回避し、現場で VLM 処理して結果のみを抽出する構成が不可欠となっている。
- **3. トークン従量課金の回避と TCO 最適化：**24 時間稼働ラインの常時監視や自律エージェントの裏側処理では、クラウド API の従量課金が膨張する。高負荷・定常稼働ワークロードでは自社ハードウェア上でオープンモデルを稼働させる方が 100 万トークンあたり最大 18 倍安価との試算もあり、減価償却資産としてコストを固定化できる。
- **4. OT ネットワーク分離（エアギャップ環境）と通信途絶時の BCP 耐性：**ランサムウェア対策としてインターネットから遮断された工場制御網（OT 網）内でも完結稼働し、災害・通信障害で外部回線が途絶した緊急時にも現場の復旧支援 AI が止まらない。
- **5. モデルのバージョン固定による「品質・再現性」の担保：**クラウド側の勝手なモデル更新で回答傾向や出力フォーマットが変化する事態を防ぎ、検証済みのモデルウェイトを固定して自社の品質管理サイクルに合わせて更新できる。

■ 1.4 業務領域別の最新導入事例（2025〜2026 年）

◆ （1）設計・開発・技術伝承（オンプレミス RAG × 小規模言語モデル／専用 LLM）

- **ダイフク（クリーンルーム事業部 × リコー）：**顧客機密が極めて高い半導体工場向け搬送システム開発において、2026 年 2 月に閉域稼働型の「RICOH オンプレ LLM スターターキット」（27B クラスのオンプレ LLM＋RAG 基盤）を部門導入し、膨大な設計・作業資料の検索と提案業務を効率化。
- **IHI：**少ない計算機資源でも高精度に稼働する量子化済み小規模言語モデル（SLM）とオンプレミス RAG を組み合わせた設計支援環境を構築し、機密技術文書を社内に閉じたまま活用。
- **パナソニック ホールディングス × ストックマーク：**1,000 億パラメータ規模の企業専用 LLM「Stockmark-LLM-100b」にグループ内の社内技術データを追加学習させ、汎用パブリックモデルでは回答できない固有技術・専門知識の活用を推進。

◆ （2）設備保全・異常対応・PLC 制御エンジニアリング（OT 閉域網 × 保全 AI エージェント）

- **半導体検査装置メーカー工場（オンプレ GPU 運用）：**クラウド利用不可の機密環境下で社内 GPU サーバーを構築し、夜間・休日の異常連絡票に対して過去類似事例からローカル LLM が一次回答を自動提示する体制を稼働。
- **ダイキン工業 × 日立製作所：**工場設備図面をナレッジグラフ化し、保全記録と熟練者の故障分析プロセスを学習させた「設備故障診断 AI エージェント」を運用。故障原因と対策を 10 秒以内・90%以上の精度で回答。
- **PLC 制御設計・デバッグの AI エージェント化：**キーエンスの PLC 専用 AI「KV Intelligence」（2026 年 11 月発表）は、自然言語からのラダー／ST 言語生成に加え、停止時のドライブレコーダ波形を正常時と時系列比較する「ディープデバッグ」を搭載。オムロン（「CX-One Agent」）やシーメンスも制御設計 AI を展開しており、制御プログラム資産の閉域保護と合わせた実装が進んでいる。

◆ （3）外観検査・VLM 工程監視・フィジカル AI（自律工場化）

- **VLM による工程文脈理解と合成データ活用：**良品・不良品の二値判定にとどまらず、長時間カメラ映像をローカル VLM で解析して作業動線のボトルネックや不安全行動（ヒヤリハット）をテキスト警告する運用が拡大。日本精工（NSK）×データグリッドのように少数サンプルから不良品画像を生成 AI で合成し、エッジ検査 AI を短期立ち上げする手法も定着している。
- **シーメンス × NVIDIA・国内ロボット各社のフィジカル AI：**NVIDIA GPU 搭載の産業用 PC（IPC）によりエッジ AI 実行速度を 25 倍に高速化。独エアランゲン工場では 100 以上の AI アルゴリズムが稼働し、生産計画コスト最大 20%削減・仮想立ち上げ工数最大 85%削減を達成。国内でも安川電機、ファナック、デンソー、Mujin 等による低遅延エッジ自律ロボット実装が本格化している。

■ 1.5 製造業特有の「3 つの壁」と導入成功の要諦

- **壁① 図面・結合 Excel・暗黙知というデータ構造化の壁：**製造業の文書は三面図、公差記号、セル結合だらけの Excel 表、現場略語で占められる。モデルの巨大さを追うより、VLM／専用 OCR による表・図面の構造化前処理と社内用語辞書の整備、さらにはコニカミノルタや川崎重工業のように熟練者への「AI インタビュー」で暗黙知を言語化してから RAG へ格納する工程が精度を決める。
- **壁② 完全ローカルと閉域クラウドのハイブリッド線引き：**顧客 NDA 図面・未出願発明・製造レシピ・OT 保全ログは「ローカル AI 必須（社外送信ゼロ）」とし、公開特許・学術論文の網羅的ランドスケープ分析や公開サプライチェーン情報突合は「閉域クラウド AI（オプトアウト契約）」を使い分けるハイブリッド設計が最適解となる。
- **壁③ 現場機器の耐環境管理と知財・保守契約ガバナンス：**高温・粉塵環境でのサーマルスロットリング対策として産業用 PC を選定するとともに、納入装置の PLC プログラムや図面を自社ローカル AI へ読み

込ませて解析・改変する際は、装置メーカーとの保守保証範囲や著作権・営業秘密条項に抵触しないよう法務・知財部門で事前ルールを策定する。

第 2 章　研究開発・知財部門における未公開発明／実験データのローカル LLM 活用アーキテクチャ

研究開発（R&D）および知財部門で「出願前の発明提案書」や「実験生データ（成功・失敗データ）」を扱う AI 基盤は、一般的な社内規程 Q&A とは法的・技術的要件の次元が異なる。特許法上の新規性喪失リスク遮断、不正競争防止法上の営業秘密要件の立証、共同研究 NDA の遵守、そして架空データの明細書混入防止をシステム構造自体で担保しなければならない。

■ 2.1 なぜ「オプトアウト契約のクラウド AI」だけでは不十分なのか

- **1. 特許法第 29 条第 1 項（新規性喪失）のグレーゾーン排除：**外部クラウド API では学習利用オフであっても、不正利用監視（Abuse Monitoring）や障害解析のためプロンプトが一定期間（例：30 日間）プロバイダ側サーバーに保持され、権限を持つ外部エンジニアが閲覧し得る規約が残る場合がある（ゼロデータリテンション特約がない限り完全消去されないケースも多い）。日本弁理士会のガイドライン等でも指摘される通り、出願前発明は物理的に社外ネットワークへ出さない閉域構成が最も確実な防御となる。
- **2. 不正競争防止法上の「秘密管理性・非公知性」の厳格維持：**配合比率、焼成温度プロファイル、触媒条件など、あえて出願せずノウハウ秘匿（先使用権確保＋営業秘密保護）とする領域では、外部クラウド送信自体が「秘密管理性の揺らぎ」と争われるリスクを断つため、アクセス制御と監査ログを自社閉域サーバー内で完結させる意義が大きい。
- **3. 共同研究契約・顧客 NDA の「第三者サーバー送信禁止条項」：**大学・公的機関との共同研究データや B2B 顧客預かり仕様書の解析では、クラウドベンダーを含む第三者サーバーへのアップロード自体が一律禁止される契約形態が一般的である。
- **4.「失敗実験データ（ネガティブデータ）」の知財資産化：**電子実験ノート（ELN）に眠る膨大な失敗条件は社外秘の塊である一方、特許明細書の「比較例」設計や「数値限定発明の臨界的意義」を立証する決定的な武器となる。

■ 2.2 R&D・知財ローカル LLM の 5 層リファレンス・アーキテクチャ

未公開発明と公開特許公報を安全かつ高精度に突合するため、「完全閉域セキュアゾーン」と「DMZ（公開情報同期層）」を分離し、データの流れを「外から内への片方向引き込み」に限定する 5 層アーキテクチャを採用する。

レイヤー	主な役割・処理内容	主要コンポーネント・技術例	セキュリティ・知財統制の要点

レイヤー	主な役割・処理内容	主要コンポーネント・技術例	セキュリティ・知財統制の要点
① データ収集・前処理層 (Ingestion & Parsing)	ELN 生データ、試験表、CAD 図面、発明提案書、公開特許公報の構造化	ローカル VLM（Qwen-VL 等）による図表・化学構造式 OCR、SMILES 変換	外部 OCR API 不使用。表組・実験条件を Markdown 表や JSON へ構造化し実験 ID を保持
② 知識格納・検索層 (Hybrid RAG & Graph)	ベクトル検索＋全文検索＋知識グラフによる高精度検索	Qdrant / Milvus / Elasticsearch（オンプレ）、GraphRAG（課題・手段・実験 ID 紐付）	文書・プロジェクト単位の RBAC（属性ベースアクセス制御）をメタデータに強制付与
③ ローカル推論基盤層 (Inference Engine)	閉域 GPU 上での高速並列推論・構造化 JSON 出力制御	vLLM / TensorRT-LLM / SGLang、32B 量子化モデル（部門 WS）～70B+モデル	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・バージョン固定
④ エージェント実行層 (IP / R&D Workflow)	発明発掘、先行技術対比、実施例・比較例マッピング、FTO 一次評価	Dify On-Prem / LangGraph 等、社内 MI（マテリアルズ・インフォマティクス）連携	全入出力ログ・タイムスタンプを改ざん不能ストレージへ保存（先使用権・発明者証跡）
⑤ 外界連携（片方向ゲート） (One-Way Sync)	公開特許・論文 DB の定期バルク同期、または検索クエリの抽象化出力	J-PlatPat / 商用特許 DB からの公報同期バッチ、抽象化検索式ゲートウェイ	「生データは絶対に外へ出さず、公開公報を内へ引き込む」片方向データフローの徹底

■ 2.3 R&D・知財連携における 4 大コア・ワークフロー

- **① ELN／MI データからの「発明発掘・数値限定ポイント抽出」**：実験条件マトリクスと評価結果から「特定数値範囲で急激に特性が向上する変曲点」や「従来トレードオフとされた 2 特性の両立ロット」をローカル LLM が自動ハイライトし、研究者への対話インタビューを通じて製造工程上の暗黙知を引き出し発明提案書ドラフトへ変換する。
- **② 2 段階ハイブリッド方式による「外部漏洩ゼロの先行技術調査・FTO スクリーニング」**：【Step 1】未公開発明提案書をローカル LLM に入力し、具体的化合物名や固有数値を上位概念化して IPC/FI/CPC 分類と検索論理式のみを生成する。【Step 2】その検索式で商用特許 DB から母集団公報（数百～数千件）をオンプレミス環境へ一括ダウンロードし、閉域内のローカル LLM 上で「未公開発明 vs. 各先行文献」の

クレームチャートを一括生成する。これにより自社の注目領域を外部に一切悟られず精密な対比が可能となる。

- **③ 事実データ固定型ドラフティング支援：**ELN の実験 ID を厳格にバインディングした上で実施例・比較例を構成し、サポート要件・実施可能要件の脆弱箇所（実験裏付けの空白領域）を事前に炙り出す（詳細は第 3 章にて詳述）。
- **④ ノウハウ秘匿（先使用権確保）と特許出願の戦略的仕分け：**他社製品からのリバースエンジニアリング容易性（侵害立証可能性）と公知文献からの容易想到性をローカル LLM で壁打ちし、「出願権利化すべき技術」と「工場閉域内に秘匿しタイムスタンプ付きで先使用権証拠化すべきプロセス技術」の仕分けを支援する。

■ 2.4 見落とされがちな「内部ガバナンス」3 つの落とし穴と対策

- **1. チャイニーズ・ウォール（社内情報隔壁）の RAG 汚染防止：**A 社との共同研究データと B 社との共同研究データを単一の共有 RAG インデックスに無防備に混在させると、A 社由来の秘密知見が B 社担当者の AI 回答に混入する（コンタミネーション／契約違反）。対策として、共同研究・NDA データはモデルの重みに記憶させるファインチューニングには絶対に使用せず、RAG のメタデータ権限フィルタ（ACL/RBAC）で所属プロジェクト外のチャンクを検索段階で完全遮断する。
- **2. 自然人発明者の寄与記録とプロンプト監査ログ：**現行特許法では自然人のみが発明者となり得る。AI との対話で着想が深まった際、人間の創作寄与と AI の整理出力を切り分けられるよう、プロンプト・入力実験 ID・出力履歴をユーザーID およびタイムスタンプ付きで発明届出書に紐付け保存する。
- **3. オープンウェイトモデルのライセンスと出所検証：**商用利用制限や出力利用制限（他モデル学習への転用制限等）のない Apache 2.0 / MIT ライセンス等のクリーンなモデルを法務・知財部門で選定し、閉域搬入前にウェイトファイルの改ざん・脆弱性検査を実施する。

第 3 章　ELN 実測データからの特許明細書（実施例・比較例・臨界的意義）生成パイプラインとプロンプト設計

電子実験ノート（ELN）の実測データから特許明細書の「実施例・比較例」および「臨界的意義」を生成する際、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。日本特許庁におけるサポート要件・実施可能要件の厳格審査や、中国 CNIPA 等における AI 生成の架空実験データに対する拒絶・無効リスクを踏まえ、数値計算・クレーム照合は決定論的プログラム（Python）が担い、ローカル LLM は『構造化 JSON にロックされた事実の言語化と論理監査』のみを担うニューロ・シンボリック構成を採用する。

■ 3.1 ELN → 明細書ドラフトの 4 段階データ連携パイプライン

◆ Step 1：正規化・対照実験ペアリング層（Python ETL 処理）

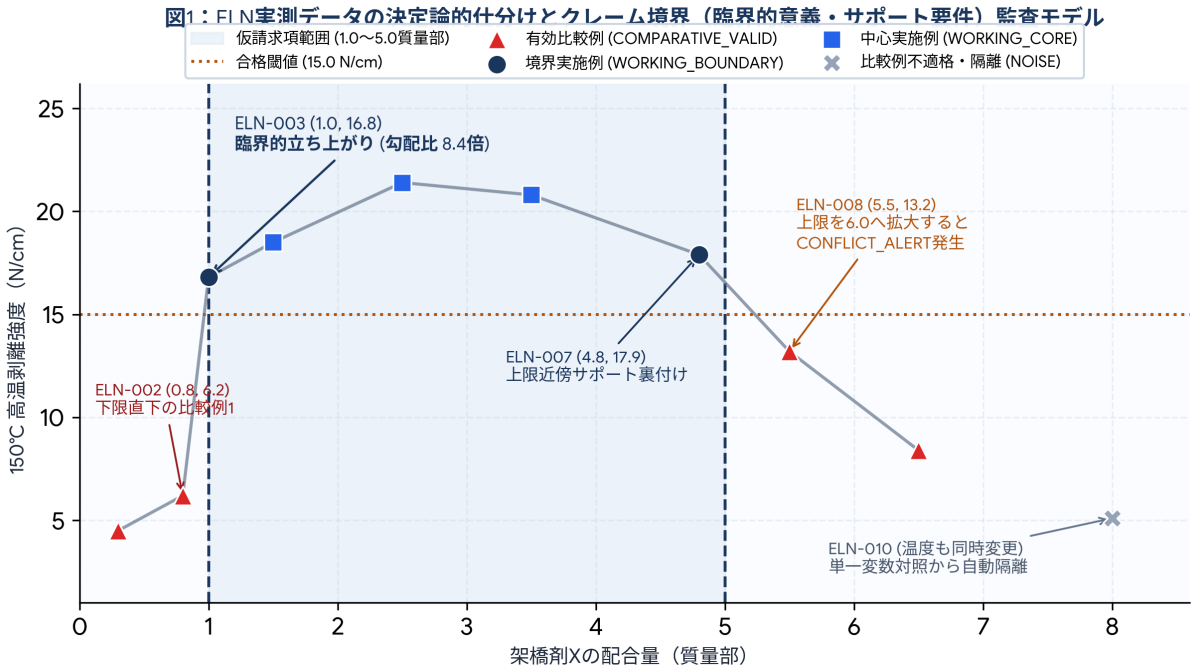
- **単位・環境条件の正規化：**「wt%」「重量%」を「質量%」へ統一し、「室温（RT）」等の曖昧表記には ELN の環境ログ（例：23°C、50%RH）を紐付ける。未記録の必須条件には推測値ではなく明示的に null（欠損フラグ）をセットする。
- **単一変数対照（OFAT: One-Factor-at-a-Time）のグラフ化：**比較例として法的に有効な対比を行うには「クレーム特定事項以外の条件が実施例と実質同一」でなければならない。複数条件を同時変更した探索初期の失敗データは NOISE_MULTI_VAR として比較例候補から自動隔離する。

◆ Step 2：仮クレーム条件との決定論的マッピングと臨界的意義算出

仮請求項の数値範囲と各実験レコードを Python スクリプトで照合し、以下のロールラベルと変化率（差分商 ΔY/ΔX）を機械的に確定させる。

自動付与ラベル	Python 側の決定論的判定ロジック	明細書上の役割と知財リスク判定
WORKING_BOUNDARY	クレーム範囲内、かつ上限・下限から所定幅（5〜10%）以内	【境界実施例】サポート要件（特許法 36 条 6 項 1 号）の上下限裏付けの要
WORKING_CORE	クレーム範囲内の中央領域、かつ目標効果の合格閾値を達成	【中心実施例】発明のベストモード（最良の形態）候補
COMPARATIVE_VALID	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条件	【有効比較例】数値限定の臨界的意義を立証する直接対照データ
CONFLICT_ALERT	クレーム範囲内であるにもかかわら	【重大警告】放置すると実施可能要件・サ

自動付与ラベル	Python 側の決定論的判定ロジック	明細書上の役割と知財リスク判定
	ず、目標効果が合格閾値未満	ポート要件違反（減縮・除くクレーム検討）



◆ Step 3：「Fact-Locked JSON」データスキーマ仕様（LLM 入力ペイロード）

ローカル LLM には自由記述の実験ノートを渡さず、以下のように役割・数値・差分基準 ID が完全にロックされた JSON のみを入力する。

Fact-Locked JSON ペイロード構造例

```
{
  "invention_context": {
    "target_claim_1": "樹脂 A 100 質量部に対し、架橋剤 X を 1.0 質量部以上 5.0 質量部以下含有し、硬化温度 120℃以上 160℃以下で熱処理してなる耐熱接着剤組成物。",
    "evaluation_standard": "JIS K 6850 準拠、150℃環境下での高温剥離強度を測定。合格基準は 15.0 N/cm 以上。"
  },
  "criticality_metrics": {
    "lower_bound_transition": {
      "comp_id": "ELN-2026-002", "work_id": "ELN-2026-003",
      "x_change": "0.8 質量部 → 1.0 質量部", "y_change": "6.2 N/cm → 16.8 N/cm",
      "slope_ratio_vs_internal": 8.4
    }
  },
  "experiments": [
    {
```

```
      "eln_id": "ELN-2026-003", "assigned_label": "実施例 1", "role": "WORKING_BOUNDARY_LOWER",
      "locked_conditions": {"樹脂 A": "100 質量部", "架橋剤 X": "1.0 質量部", "硬化温度": "140℃", "硬化時間": "30 分", "撹拌回転数": null},
      "locked_results": {"高温剥離強度": "16.8 N/cm", "外観判定": "良好（気泡なし）"},
      "delta_base_id": null
    },
    {
      "eln_id": "ELN-2026-002", "assigned_label": "比較例 1", "role": "COMPARATIVE_VALID_LOWER",
      "locked_conditions": {"樹脂 A": "100 質量部", "架橋剤 X": "0.8 質量部", "硬化温度": "140℃", "硬化時間": "30 分", "撹拌回転数": null},
      "locked_results": {"高温剥離強度": "6.2 N/cm", "外観判定": "凝集破壊発生"},
      "delta_base_id": "ELN-2026-003"
    }
  ]
}
```


◆ Step 4：決定論的ベリファイア（出力後の自動逆照合ゲート）

ローカル LLM が生成した明細書テキストから正規表現・構文解析スクリプトで「すべての数値＋単位ペア」を逆抽出し、入力 JSON の locked_conditions / locked_results に存在しない数値（勝手な四捨五入・比率計算・架空の撹拌時間など）が 1 つでも検出された場合は赤字ハイライト警告または自動再生成を実行する。

■ 3.2 ハルシネーションを排除する 3 段階プロンプトチェーン（実装テンプレート）

単一の巨大プロンプトで「監査・執筆・考察」を同時に行わせると、32B〜70B クラスのローカル LLM でも注意機構（Attention）が分散し条件取り違えを起こすため、必ず以下の 3 段階に分割して実行する。

【Stage 1】サポート要件監査＆意識的除外リスク検証プロンプト

```
# Role
あなたは製造業の知財部門に所属する熟練の特許実務家です。与えられた JSON 形式の実験データと仮請求項のみに基づき、明細書執筆前の「サポート要件・実施可能要件・意識的除外リスク」の監査レポートを出力してください。

# Strict Rules（絶対遵守事項）
1. JSON 内に存在しない実験条件や数値を推測・補完することは厳禁とする。
2. 以下の 4 点を必ず検証し、該当する `eln_id` を明示して指摘すること。
   - 【上下限サポート監査】：仮請求項の各数値範囲（上限・下限）の近傍を裏付ける `WORKING_BOUNDARY` データが存在するか。空白領域がある場合、クレーム減縮案または追加実験条件を提示せよ。
   - 【クレーム内不整合監査】：`CONFLICT_ALERT`（クレーム範囲内だが効果未達のデータ）が存在する場合、従属項や除くクレームで切り分け可能なパラメータ差が JSON 内にあるか確認せよ。
   - 【対照実験の妥当性監査】：各比較例（`COMPARATIVE_VALID`）が、対応する実施例（`delta_base_id`）に対
```

して検証対象の特定事項のみが異なる対照実験になっているか確認せよ。
- 【意識的除外（均等論第 5 要件）リスク監査】：本来であれば本発明の変形例・均等物として包含し得る準良好データを、安易に「比較例（不良）」として排除していないか警告せよ。

```
# Input JSON
{{FACT_LOCKED_JSON}}
```

【Stage 2】事実データ固定型・実施例／比較例ドラフティングプロンプト

```
# Role
あなたは正確性を最優先する特許明細書ドラフターです。入力 JSON の `experiments` 配列に基づき、日本特許明細書の【実施例】および【比較例】の段落テキストと、【表 1】の Markdown テーブルを作成してください。

# Zero-Hallucination Constraints（ハルシネーション排除制約）
1. 【完全転写原則】：数値、単位、成分名、評価結果は、入力 JSON の文字列を 1 文字たりとも変更・四捨五入・単位換算せずにそのまま転写せよ。
2. 【欠損値の推測禁止】：JSON 内で値が `null` となっている項目（例：撹拌回転数など）について、一般的な実験手順を勝手に補完して記述することは絶対に行うな。文脈上言及が必要な箇所には必ず `[要確認: ELN 未記載パラメータ（項目名）]` というプレースホルダをそのまま出力せよ。
3. 【差分記述（Delta 記述）の厳格制御】：`delta_base_id` が指定されている実施例 2 以降および比較例については、「実施例 1 において、［変更項目］を［変更前数値］から［変更後数値］に変更したこと以外は、実施例 1 と同様にして…」の定型構文を用いよ。JSON 上で同一でない項目を「同様」とまとめるミスを厳に禁ずる。
4. 【トレーサビリティタグ保持】：各段落の末尾に社内監査用タグ `（参照 ID: {{eln_id}}）` を必ず付記せよ。

# Input JSON
{{VERIFIED_EXPERIMENTS_JSON}}
```

【Stage 3】臨界的意義・作用機序ドラフティングプロンプト（限定解釈ガードレール付）

```
# Role
あなたは特許明細書の「発明の効果・考察」パートを執筆する知財専門家です。入力 JSON の `criticality_metrics` および実施例・比較例の対比結果に基づき、数値範囲の臨界的意義と技術的考察の段落を作成してください。

# Drafting Guardrails（知財実務ガードレール）
1. 【事実ベースの対比論証】：実施例群と比較例群の具体的な実験番号と測定値（JSON 内の確定値のみ）を引用し、クレーム境界を境として評価特性が非連続的・顕著に変化している事実を客観的に記述せよ。改善倍率は暗算せず `criticality_metrics` 内の計算値のみを使用すること。
2. 【断定表現の禁止と権利解釈セーフティネット】：作用機序に言及する場合は決して断定せず、必ず「〜と推察される。」と記述した上で、段落末尾に以下の免責定型文を一字一句違えず挿入せよ。
   「なお、上記の作用機序は現時点における推察を含むものであり、本発明は当該作用機序や特定の理論によって何ら限定的に解釈されるものではない。」
```

```
3. 【比較例の過度な貶め禁止】：比較例については客観的数値事実のみを淡々と述べ、「全く使用に耐えない」「致命的な欠陥がある」といった過剰な否定表現（将来の改良出願や均等主張の妨げとなる表現）は一切使用しないこと。

# Input JSON
{{VERIFIED_EXPERIMENTS_JSON}}
```

■ 3.3 ローカル LLM 推論パラメータと審査対応実務の急所

- **repetition_penalty（繰り返しペナルティ）を 1.0〜1.02 に抑える：**一般チャット用の初期値（1.1〜1.2）のままでは、LLM が同一表現の反復を嫌い「質量部」を「重量部」、「高温剥離強度」を「熱間接着力」と勝手に言い換えて記載不備を招く。特許明細書ジョブではペナルティなし（1.0 近傍）への固定が必須である。
- **temperature ＝ 0.0 ＋ Guided Decoding の併用：**vLLM 等の guided_json 機能で監査出力や表構造を文法レベルで拘束し、再現性を最大化する。
- **「出願明細書 ⇔ ELN 対照台帳」の自動秘匿アーカイブ：**出願前の最終クレンジングで文中の（参照 ID: ELN-2026-003）タグを除去する際、自動的に「明細書の実施例番号 ⇔ ELN 実験 ID・測定生データパス・実験日・実験者名」の対応台帳を社内秘匿ストレージへ保存し、外国審査での真正性照会や先使用権立証に備える。

第 4 章　中堅・中小製造業が 100 万円クラスのワークステーションで始めるオンプレ RAG 構築手順

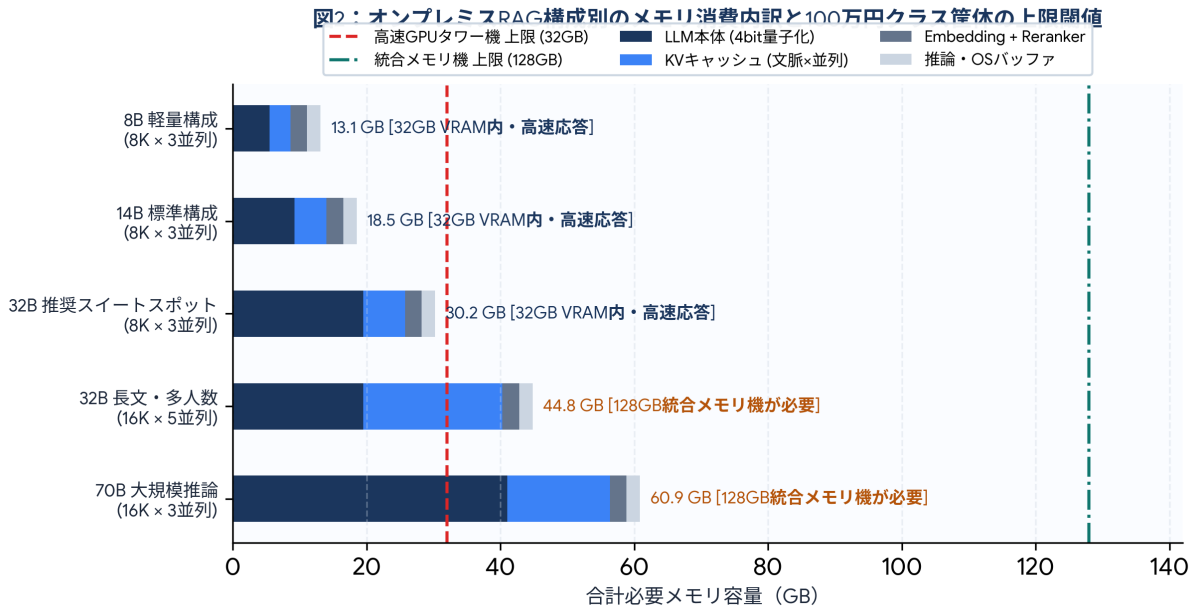
従業員 30〜300 名規模の中堅・中小製造業において、社内専用のオンプレミス RAG（検索拡張生成）を立ち上げるのに数千万円のサーバー室も専任 AI エンジニアも不要である。100 万円前後の AI ワークステーション 1 台、事務所の 100V コンセント、そして商用利用可能な OSS を組み合わせることで、図面・過去トラブル集・作業手順書を社外に出さない実用基盤を約 6 週間で構築できる。

■ 4.1 予算 100 万円クラスで選べる 3 つのハードウェア構成（2026 年基準）

オンプレミス RAG の快適さは「メモリ容量（搭載できるモデルの大きさ）」と「メモリ帯域幅（回答の書き出し速度）」のバランスで決まる。

構成タイプ	代表構成例	利用可能メモリ	実勢価格目安	特徴と最適な現場用途
高速 GPU タワー機（応答速度最優先）	RTX 5090 搭載 BTO WS ＋本体 RAM 128GB	32GB 高速 VRAM	85 万〜120 万円	毎秒 40〜80 トークン超の高速応答。10〜30 名の現場 Q&A とスキャン PDF の高速 OCR に最適
統合メモリ小型 AI 機（大容量モデル対応）	NVIDIA DGX Spark	128GB 統合メモリ	75 万〜95 万円	省電力・小型筐体で 70B 級大型モデルや長文コンテキストも格納可能。検証・技術部門向け
静音省電力 Mac 機（事務所内・少人数）	Mac Studio 上位構成（M4 Max / M5 Max）	128GB 統合メモリ	75 万〜95 万円	極めて静音・低消費電力（150〜250W）。設計・知財・品質保証等の少数精鋭壁打ちに最適

- **周辺機器を含めた必須スペック：**元文書 PDF とベクトル DB の肥大化に備え、内蔵ストレージは NVMe SSD 2TB〜4TB（ミラーリング推奨）を確保する。また工場・事務所の瞬時電圧低下や停電から DB 破損を防ぐため、正弦波出力の UPS（無停電電源装置：実売 3 万〜6 万円）を必ず組み込む。



■ 4.2 ライセンス費用ゼロで組める標準 OSS 技術スタック

- **OS・コンテナ実行基盤：**Ubuntu 24.04 LTS ＋ Docker Compose。全コンポーネントをコンテナ化して一括起動・丸ごとバックアップ可能にする。
- **推論エンジン（vLLM または Ollama）：**同時利用 2〜3 名の検証なら導入容易な Ollama、部門全体（同時 5 名以上）への本番公開では並列バッチ処理に優れた vLLM を採用する。
- **32GB VRAM へのモデル最適配分（推奨スイートスポット構成）：**①回答生成 LLM（約 12〜19.5GB）：Qwen 系・Gemma 系・Llama 系日本語強化モデルやリコー・NTT 等の国産オープンモデル（14B〜32B・4bit 量子化、Apache 2.0 等の商用可ライセンス）。②検索用 Embedding＋Reranker（約 2.5GB）：検索結果を再採点して上位を絞り込む Reranker の併用が RAG 精度を劇的に引き上げる。③図表 OCR 用軽量 VLM（夜間バッチ切替起動）：7B〜11B 級の視覚言語モデル。
- **RAG・画面管理基盤：**Dify Community Edition（オンプレミス版）＋ Qdrant（日本語ハイブリッド検索対応ベクトル DB）。ノーコードでチャット画面と部門別アクセス権限を管理できる。

■ 4.3 6 週間で本番稼働させる段階的構築ロードマップ

フェーズ	主要タスク	実務上の重要チェックポイント（失敗回避の勘所）
Week 1 対象選定・環境手配	対象 1 業務（100〜300 文書）の選定、電源・排熱・固定 IP 手配	最初から全社文書を入れない。「過去トラブル対策書」「作業標準書」「検査基準書」のいずれか 1 テーマに絞る（三面図 CAD の幾何解析は第 2 段階へ回す）。
Week 2	Ubuntu＋Docker 導入、モデ	ネット遮断前に全コンテナとモデルウェイトを取得し、オフ

フェーズ	主要タスク	実務上の重要チェックポイント（失敗回避の勘所）
基盤構築・閉域搬入	ル取得・ハッシュ検証、閉域 LAN 設置	ライン起動設定（自動更新通信オフ）を完了させてから工場・社内 LAN へ固定 IP 接続する。
Week 3 製造文書クレンジング	セル結合 Excel 解除、スキャン PDF の Markdown OCR、用語辞書整備	RAG 精度の 8 割を決める最重要工程。Excel のセル結合を解除して 1 行 1 レコード化し、チャンク先頭に「設備名・型番・改訂日」を強制付与。現場隠語辞書を登録する。
Week 4 ハイブリッド検索実装	ベクトル＋キーワード検索＋Reranker 設定、根拠強制プロンプト固定	型番・エラー番号（E-402 と E-403）の 1 文字違いを拾うためハイブリッド検索を有効化。「社内文書に記載がない事項は推測せず該当なしと答え、参照ファイル名・頁を明記せよ」と固定する。
Week 5 通信遮断・権限分離	ファイアウォール外向き遮断実証、プライベート SSL、部門 RBAC 設定	パケット監視で外部インターネットへの通信試行ゼロを確認。「製造課共通」「品証限定」「特定顧客 NDA」でナレッジベースの閲覧権限を分離する。
Week 6 50 問実機検証・定着	現場キーマン 5 名による 50 問テスト、不足知見の Q&A 追補、バックアップ設定	誤答の大半は AI ではなく「元手順書の古さ・暗黙知の未記載」が原因。不足知識を Q&A シート 1 枚で追補して正答率 85%超へ引き上げ、NAS 夜間自動バックアップを開始する。

■ 4.4 ランニングコスト・投資回収（ROI）と「ひとり情シス」を守る保守設計

- **電気代は月額 3,000〜7,000 円前後、約 1 年で初期投資回収：**32GB GPU 搭載機や 128GB 統合メモリ機の待機電力は 50〜100W、推論ピーク時でも 350〜600W 程度に収まる。社員 30 名に外部の商用 RAG クラウド（月額数千円/人＋容量課金）を契約すれば年間 120 万〜200 万円以上の継続費用が発生するため、オンプレミス機なら約 1 年でハードウェア投資を回収でき、以降はトークン課金ゼロで使い放題となる。
- **ハード故障を抱え込まない保守体制：**BTO ワークステーション導入時にベンダーの「3〜5 年オンサイト翌営業日修理保守（十数万円程度）」を必ず付帯させる。OS と Docker 環境・Qdrant データを社内 NAS へ夜間イメージバックアップしておけば、万一のハード障害時も専任エンジニアなしで数時間以内に復旧できる。

【実務の要点】全章を通じた総括：製造業ローカル AI 成功のロードマップ

製造業におけるローカル AI 実装の本質は、単なる「クラウドからオンプレへの置き換え」ではない。

① 現場の 100 万円ワークステーションから R&D の閉域 GPU 基盤まで、機密性と応答速度に応じて最適

配置し、

② 図面・結合 Excel・ELN 実験生データを「AI が誤読しない構造化データ（Markdown / Fact-Locked JSON）」へ整え、

③ 数値判定やアクセス権限（チャイニーズ・ウォール）は決定論的プログラムで厳格にロックした上でローカル LLM の言語化・推論能力を引き出す。

このアーキテクチャを確立することで、営業秘密保護・知財品質向上・現場技術伝承の 3 つを同時に高次元で実現できる。