製造業におけるローカル AI 活用の総合調査

実装事例から技術選定 投資判断 研究開発と知財まで

調査基準日 2026 年 11 月 11 日 日本時間

GPT-6 Astra

経営判断の要点

製造業でローカル AI を導入する最大の価値は、機密性の高い現場データを扱い、通信環境に依存しすぎず、必要な時間内に業務を処理できることである。ただし、ローカルに置くだけで精度・安全・費用の問題が解消するわけではない。用途を絞った検査・予兆検知と、社内文書を扱う生成 AI を別々に評価し、必要に応じて組み合わせることが合理的である。

公開資料からは、エッジ画像検査や特定工程の AI 制御に本番稼働の実績が確認できる一方、工場向けローカル LLM では PoC、限定導入、製品化準備の事例が目立つ。機器の販売開始、AI モデルの公開、工場での実証、本番での業務効果は、それぞれ別の到達点として扱う必要がある。[1–4,7–12]

判断事項	本稿の推奨
最初の対象	機密マニュアル検索、保全記録の整理、既存検査の改善など、成果を測れる業務を選ぶ
基本構成	設備近傍の専用 AI と、工場・研究所内の生成 AI サーバを分けて配置する
モデル選定	日本語、図表、単位、型番、出典提示を自社データで評価する
投資判断	機器代だけでなく、文書整備・連携・監視・更新・人の確認費を含める
自動化の順序	読み取りと提案から開始し、書き込み・制御は個別の検証を経て拡張する

特に重要な三つの区別

第一に、ローカル AI はローカル LLM より広い。画像認識、時系列分析、最適化、音声認識も含まれる。第二に、推論がローカルでも、学習や保守、検索先が外部という構成はあり得る。第三に、国産モデル、国内データセンター、オンプレミス、完全オフラインは同じ意味ではない。

読み方

第 1〜5 章は定義・事例、第 6〜10 章は業務設計、第 11〜15 章はモデル・機器・運用・投資、第 16〜19 章は統制・評価・導入手順を扱う。末尾に出典を番号順で掲載する。実績は公表主体と段階を併記し、本稿による構成案・試算・推奨は実測結果と区別した。

1 ローカル AI の範囲と導入理由

本稿では、推論処理の全部または主要部分を、自社が管理する端末・工場・研究所・データセンター内で実行する AI をローカル AI と呼ぶ。配置場所と通信条件を明示するための作業上の定義であり、特定の製品区分や認証を意味しない。

形態	主な設置場所	適した処理	留意点
端末内 AI	PC、検査カメラ、組込機器	音声、画像、小型モデル	電力・温度・メモリ制限
設備エッジ AI	産業用 PC、ロボット周辺	画像検査、異常検知	周期、耐環境、交換性
工場内 AI	工場サーバ、構内エッジクラウド	複数設備分析、保全支援	構内ネットワークと冗長化
企業オンプレミス	研究所・本社・自社 DC	RAG、設計、知財、共通基盤	工場との回線断に備える
外部専有環境	事業者 DC、専用クラウド	共有基盤、広域分析	本稿ではローカルと区別

六つの導入理由

①図面・配合・製造条件・未公開発明を扱うための管理、②検査や搬送の応答時間、③回線断への耐性、④映像・波形を外部へ大量送信する負担の回避、⑤高頻度処理の費用管理、⑥モデルや更新時期を自社で選べることが主要な理由になる。優先順位は工場ごとに異なる。

データが外へ出ない範囲は個別に確認する。OCR、埋め込み、再順位付け、音声認識、推論、監視、バックアップの一部がクラウドなら、全処理がローカルとはいえない。ネットワーク切断試験だけでなく、通常運転中の通信先の確認も必要になる。

クラウドとの比較で確認する点

クラウドはモデル更新、需要変動への対応、大規模な学習、外部情報収集に使いやすい。ローカルは機器調達、障害復旧、更新、セキュリティ管理を利用企業側で負担する。機密データでも、契約・接続・保管条件によって許容できる企業向けクラウドはあるため、「機密なら必ずオンプレミス」と一律に決めず、業務要件から選択する。

また、ローカル 5G は通信方式であり、それ自体はローカル AI を意味しない。OKI の事例は、構内通信とエッジ映像解析を組み合わせた例である。個々の製品や工程によって、AI に加えてルールベースの判定も使われている。[16]

2 製造業の業務別活用マップ

以下は導入候補の整理であり、全項目が特定企業で実証済みという意味ではない。「優先」は、データを確保でき、現行業務と比較できる場合の本稿の判断である。

業務	ローカルで扱う情報	AI の役割	最初に測る指標
研究開発	実験記録、配合、失敗試験	検索、類似条件、仮説整理	探索時間、再試験率
材料探索	組成、プロセス、物性	予測モデル＋最適化	実験回数、物性達成率
製品設計	図面、仕様、変更履歴	整合確認、変更影響候補	確認工数、見落とし
生産技術	工程表、治具、PLC 仕様	手順案、コード作成支援	設計・検証時間
品質検査	画像、寸法、良否ラベル	検出、異常度算出	見逃し、過検出、周期
品質保証	不具合票、ロット、工程記録	原因候補、証跡整理	調査時間、再発率
設備保全	振動、電流、交換履歴	異常検知＋作業検索	停止時間、誤警報
操業管理	時系列、アラーム、運転日誌	状況要約、対応候補	復旧時間、判断精度
条件最適化	温度、圧力、投入量	予測・数理最適化・強化学習	歩留まり、原単位
搬送・ロボット	画像、位置、作業状態	認識、計画、動作補助	成功率、停止、接触
技能伝承	標準書、熟練者の説明	対話、動画索引、教育	独力での正答、習熟
生産計画	能力、納期、段取り制約	最適化器への条件整理	納期遵守、段取り
調達・営業技術	仕入先仕様、見積、顧客条件	比較、技術回答下書き	照合工数、回答品質
知財	発明提案、実験、社内先行例	発明整理、技術対比	確認工数、根拠一致

着手しやすい順序

機密文書検索と保全記録の要約は、設備へ直接影響しない構成で始められる。外観検査は撮像環境が安定し、既存の良否判定を比較基準にできる工程が適する。工程条件の自動変更やロボット動作生成は、誤りが品質・設備・人に与える影響を踏まえ、別の検証計画で進める。

3 現場 AI の実運用と製品提供

Audi 溶接スパッタの画像検査

Siemens の事例では、Audi はクラウドで学習した自社モデルを工場内の Industrial Edge へ展開し、車体の溶接スパッタを検出している。現場推論の目的は短い生産サイクルへの対応である。これは「学習は外部、推論は現場」という構成の明確な例で、全工程オフラインの事例ではない。[7]

OKI 本庄工場の外観異常判定

OKI は本庄工場の一部で稼働を開始した外観異常判定システムについて、エッジ領域で映像解析し、結果・画像・製品情報を管理サーバへ蓄積する構成を公表している。検査と証跡管理を結びつけた点が参考になる。ねじ有無などにはルールベースも用いており、すべてを深層学習に置き換える構成ではない。[16]

ENEOS マテリアルと横河電機 特定工程の自律制御

横河電機は 2023 年 3 月、蒸留工程を制御する強化学習 AI の正式採用を発表した。先行する 2022 年の実証では 35 日連続制御を確認し、プラントシミュレータ、操業データでの確認、既存のインターロック、DCS との統合を組み合わせている。生成 AI チャットを設備に直結した事例ではない。[9,10]

Siemens と FANUC 予兆検知の製品群

Siemens は 2026 年 3 月に DTA Onsite を発表し、産業用 PC 内で駆動系の状態監視・異常検知を実行する構成を示した。FANUC も FIELD system Basic Package で収集するサーボデータを使う AI サーボモニタを紹介している。これらは製品機能の確認であり、各導入先での停止損失削減額を証明するものではない。[12,18]

例	確認できる段階	解釈上の注意
Audi	工場へのモデル実装	学習と推論の配置が異なる
OKI	一部工程での稼働	AI とルール判定の組合せ
ENEOS マテリアル	正式採用	検証した対象工程での制御
DTA Onsite・FANUC	製品提供・機能公表	機能説明と導入効果は別
トヨタ向け Musashi AI	量産稼働・追加納入	公表資料だけで全データ経路は断定しない

Musashi AI はトヨタのギヤ検査装置について、2020 年 12 月からの量産稼働と 2022 年時点の計 8 台導入を公表した。現場 AI の実装例として有用だが、当該資料から完全オフライン運用まで推定しない。[17]

4 国内ローカル生成 AI の事例

大分キヤノンマテリアル ワークステーションでの検証

日本 HP の 2026 年 6 月の記事では、HP Z8 G5 と RTX PRO 6000 Blackwell Max-Q を用い、gpt-oss-120B を想定したローカル RAG を構築した。仮想的な現場データを用いた PoC であり、本番業務の定量効果とは区別する。機密業務にローカル、公開情報等に Copilot を使い分ける方向が示されている。[2]

荏原製作所 機密技術文書の検索

GBase の導入事例ページは、オンプレミスの RAG と 80B モデルを用いた PoC を記載し、技術情報検索が 30 分から 2 分になったと報告する。ただし提供会社の公表値で、対象件数や比較手順等の詳細は十分示されていない。全社の生産性が 93%向上したという意味ではない。[3]

三菱電機 小型の製造業特化モデル

2025 年 6 月の公式資料は、約 18 億パラメータのモデルで、FA 製品に関する特定の質問応答評価において 77.2%、比較対象の GPT-4o は 52.0%と報告した。学習データと業務領域を合わせる価値を示すが、一般能力で大型モデルを上回る証明ではない。2026 年度中の製品適用を目標とした発表である。[1]

別川製作所と金沢工業大学 限界も公開した共同研究

13 ページのマニュアルと 18 問で評価したオンプレミス構成の研究では、自然な文章でも質問に対応しない回答が生じ、表・図の解読や同時利用に課題が残った。公開時点では大学側のデータベースで稼働しており、工場での本番導入とは区別する。小規模な資料でも評価工程を省けないことが分かる。[4]

事例	段階	主な証拠の性格
大分キヤノンマテリアル	PoC	HP の取材記事・仮想データ
荏原製作所	PoC としての運用	提供会社の事例・効果数値
三菱電機	開発と製品適用計画	開発元の特定タスク評価
別川製作所	共同研究・実証	利用企業による課題も含む説明

本稿の判断として、ここから得られる共通点は「モデルの設置だけで完了しない」ことである。文書構造、検索設計、型番・単位の扱い、利用者数、承認手順まで含めて業務を設計する必要がある。

5 海外動向と事例の読み分け

Siemens の工程別 Copilot を混同しない

2024 年の発表では、Engineering Copilot は Azure OpenAI Service を利用し、Operations 向けはオンプレミス構成が計画されていた。2025 年の NVIDIA との発表では、現場運用向けのオンプレミス最適化と保全時間削減を説明している。製品名が似ていても、設計支援と現場支援で実行場所は異なる。[13,14]

NVIDIA の Siemens 事例ページは、文書・設備データと連携するオンプレミス Copilot と、初期パイロットでの生産性改善を紹介する。これは供給側が報告する初期成果である。特定工場の長期稼働、比較条件、保守費を伴う独立検証と同等には扱わない。[15]

企業専用 LLM とローカル LLM も別の軸

Panasonic-LLM-100b は、社内情報を追加事前学習する企業特化モデルとして発表された。経済産業省の事例は事業所・工場での試行を紹介する。ただし企業専用であることだけから、全拠点で完全オフライン稼働しているとは判断できない。モデルの専用化と設置場所は別に確認する。[5,6]

先進工場の AI がすべてローカルとは限らない

Schaeffler の Microsoft 事例は、Microsoft Fabric と Azure AI のエージェントによる工場情報活用である。Pegatron の NVIDIA 事例は、映像 AI やデジタルツインの広がりを示すが、工場内と外部の処理配分は用途ごとに確認する必要がある。製造業の成功例というだけで、ローカル導入の効果として引用すべきではない。[19,20]

公表表現	確認すべき内容
工場向け AI	推論の実行場所、設備との接続先
企業専用モデル	モデルの専有性、重みの保管・運用主体
オンプレミス対応	対応可能か、納入済みか、実業務稼働か
最大 30%改善	作業の範囲、母数、比較基準、実測か予測か
完全閉域	初期導入・認証・保守・更新・検索を含むか
情報漏洩リスクゼロ	端末持出し、権限、ログ、内部不正も検証したか

ローカル生成 AI の普及可能性は高いと考えられるが、本調査の事例抽出は市場全体の統計標本ではない。公開事例の数や宣伝上の改善率から、製造業全体の普及率や平均 ROI を算出することはできない。

6 研究開発 設計 知財での活用設計

研究所のローカル AI は、公開情報だけでは得られない実験条件・失敗原因・設計判断を横断して扱える点に価値がある。以下は実装案であり、特定企業で同一構成の効果を測定した結果ではない。

研究開発は文献と実験を結びつける

実験ノート、LIMS、分析装置の出力、配合表を実験 ID で結合し、「似た配合でうまくいかなかった例」「温度条件を変えたときの物性差」を検索する。LLM は記録の整理と説明、物性予測は専用の統計・機械学習モデル、次の実験条件の探索はベイズ最適化等の役割に分ける。生成文だけを実験結果として扱わない。

化学・材料では、濃度基準、湿潤・乾燥条件、測定法、ロット、検出限界が結果を左右する。欠測をゼロに置き換えたり、異なる試験法の値を同じ列へ統合したりすると誤った比較になる。データ整備の段階で単位・測定条件・不確かさを保持する。

設計では変更影響の確認を支援する

図面と BOM、仕様書、試験成績、設計変更通知を改訂番号で関連付ける。寸法、公差、材質、表面処理、適用規格を構造化して照合し、LLM は差分の説明や確認先の候補提示を担う。CAD や CAE の数値処理・干渉判定を言語モデルの推測で代用しない。

知財部門は未公開情報の整理から始める

発明提案書と実験記録から技術的特徴・効果・比較例・未検証事項を分け、技術者ヒアリングの質問案を作る。社内類似発明や既存の製造ノウハウとの関係を整理し、公開特許との対比表へつなぐ。公報検索の外部サービスへ送る検索語にも、未公開の組合せや用途が含まれ得るため、外部検索と社内評価の入力を分ける。

AI が推測した効果、実施していない試験、存在しない比較例を明細書へ混ぜない。特許庁の AI 関連事例は、進歩性・記載要件・発明該当性の判断を扱っており、AI を利用したこと自体が記載要件を軽くするわけではない。[42]

入力	AI に任せる処理	人が確定する事項
実験・設計記録	抽出、索引、対比、質問案	事実、単位、技術的因果
発明提案	構成と効果の整理、代替案	発明者の寄与、出願範囲
他社公報	候補抽出、クレーム要素整理	権利範囲、法的結論
共同研究資料	資料の分類、利用範囲の索引	契約上の利用権限

7 保全支援と技能伝承の実装

保全支援は検索と設備状態を結びつける

保全担当者が警報番号を入力すると、対象設備の型式・製造番号・ソフトウェア版を確認し、最新版のマニュアル、過去故障、交換履歴、現在のセンサ値を取得する。回答には対象機種、手順の出典、禁止事項、追加確認項目を示す。類似機種の手順をそのまま適用しない。

有効な出力は、原因を一つに断定する文章よりも、確認順序を伴う候補一覧である。「センサ故障」「実際の圧力異常」「配線不良」を分け、各候補を見分ける測定と、担当者が参照すべき標準書を示す。作業後の結果を記録し、次の検索へ反映する。

日報の自動化は記録の品質から評価する

音声入力で作業日報を作る場合、騒音・マスク・専門用語・多言語が認識を妨げる。設備番号、交換部品、時刻、単位は確認欄に分ける。「異常なし」と「異常あり」の取り違えを文体の流暢さで見逃さない。音声認識と要約の誤りを別に採点する。

技能伝承では暗黙知を検証可能な記録へ変える

熟練者へのヒアリングを録音し、判断の前提、通常と異なる兆候、やってはいけない操作、過去の失敗を抽出する。説明を標準書と突き合わせ、個人の慣習と正式手順を分ける。動画は工程・工具・条件・危険箇所にタグを付け、実際の場面へ戻れる索引として使う。

機能	最初の構成	主な評価
マニュアル検索	読み取り専用 RAG＋原文表示	正しい版・型式・根拠を返す率
故障対応候補	警報・履歴との検索連携	復旧時間、誤った手順の提示
音声日報	ローカル音声認識＋確認画面	型番・否定語・数値の認識
教育対話	出典付き説明＋小テスト	AI なしでの理解・作業能力
保全指示書	テンプレートへの下書き	修正時間、承認までの所要時間

初期導入では、AI に作業指示書の下書きを作らせ、保全責任者が承認する。権限を広げる場合は、作業区分ごとに許容範囲を定義する。慣れてきた利用者が出典確認を省略することも評価し、利用回数だけを成功指標にしない。

8 品質検査と不良原因解析

外観検査は撮像条件と判定基準が先にある

傷、欠け、異物、接合不良、印字、組付け状態では、必要な解像度、照明、撮像角度、搬送速度が異なる。AI の種類を選ぶ前に、最小欠陥寸法と視野、許容差、死角、検査タイミングを決める。見えない欠陥はモデルを大きくしても検出できない。

良品だけで学習する異常検知は、不良例を集めにくい工程で候補になる。ただし「いつもと違う」ことと「出荷不可」は一致しない。学習に含まれない正常なバリエーションによる過検出も評価する。Intel の製造業向けソフトウェア群は、画像・時系列のエッジ実装の選択肢を示している。[34]

見逃しと過検出を別々に測る

本稿の例示として、1 日 10 万個、実際の不良率 0.1%のラインを考える。不良 100 個のうち 99 個を検出しても 1 個は見逃す。良品 99,900 個に対する誤検出率が 1%なら、約 999 個が再検査に回る。全体の正解率だけでは、出荷リスクと再検査負荷の両方を評価できない。

指標	測定する内容
不良見逃し率	実不良のうち良品扱いした割合。欠陥種別・重大度別に集計
過検出率	実良品を不良扱いした割合。再検査の人件費に結びつける
処理時間	撮像から判定・排出まで。平均だけでなく遅い側も測る
条件別性能	品種、ロット、照明、汚れ、昼夜、設備更新前後
追跡可能性	画像、モデル版、判定値、再判定者を製品 ID へ関連付ける

原因解析は相関から因果へ進む

画像 AI が発見した不良を、ロット・材料・設備・作業者・時刻・条件と結合する。LLM は記録を整理し、統計モデルが関連を調べ、工程担当者が実験や再現試験で因果を確かめる。単に不良と同時に起きたイベントを原因として断定しない。

生成した欠陥画像は学習データの補助として検討できるが、評価用の実欠陥を置き換えない。学習・評価を画像単位で無作為に分けるだけでは、同じ製品や連続撮影が混ざり性能を過大評価する。ロットや時間帯、設備単位で分離した評価も行う。

9 工程最適化とロボットへの適用

予測 制約付き最適化 制御を区別する

歩留まり、エネルギー原単位、タクト、製品品質を改善するには、何を予測し、何を変更できるかを決める。製品の目標物性を予測するモデル、条件を探索する最適化器、実機へ指令を出す制御器は役割が異なる。LLM は制約の整理や結果説明を担えても、それだけで制御の安定性を保証しない。

モデル予測制御や強化学習が有効な場合もあるが、対象工程、外乱、操作制約、異常時の挙動を確認する。横河電機の事例が示すのは、シミュレーション、専門家の確認、既存安全機能、運用体制を組み合わせた導入である。[10]

ロボットでは動作計画と安全機能を分ける

画像・言語・動作を結びつけるモデルは、多品種作業や段取り変更への適応を支援し得る。一方、把持失敗、遮蔽、部品の変形、人の侵入、工具摩耗など、現場には学習外の状況がある。認識・計画の出力を許可された動作範囲へ制限し、停止や安全機能を別に確保する。

Jetson Thor 等は機器側の高度な推論を支える計算基盤である。大きな演算性能やメモリ容量を備えることと、ロボットとしての安全性・作業成功率が認められることは別である。[25]

導入段階	AI の権限	進む前に確認する事項
影運転	実設備へ出力しない	過去データと実時間での精度
判断支援	操作候補を示す	熟練者の評価、制約違反
限定自動化	許容範囲内のみ実行	監視、復帰手段、異常時停止
範囲拡張	検証済み条件を増やす	品種差、季節差、保守後の変化

数理最適化と生成 AI を組み合わせる

生産計画では、納期、設備能力、洗浄、段取り、材料制約を言語モデルで整理し、計画自体は制約を扱える最適化ソフトで計算する。結果を LLM が説明し、人が例外条件を追加する。最適化器の実行を伴わない自然文の「最適計画」を、そのまま生産指示へ転記しない。

本稿の推奨は、まずモデルの提案を実機へ反映しない影運転で誤りを集めることである。設備制御を伴う実証の期間は工程リスクに依存し、文書検索と同じ 90 日計画で完了するとは限らない。

10 推奨するシステム構成とデータ設計

二つの処理系と管理系を分ける

層	配置と処理	連携の原則
設備制御	PLC・DCS・安全機器	既存制御・安全機能を維持
現場推論	産業用 PC・カメラ・組込 GPU	画像・波形を必要な周期で処理
工場データ	履歴 DB・MES・品質管理	設備 ID・製品 ID・時刻を統一
社内生成 AI	RAG・音声・LLM・VLM	認証、文書権限、出典提示
モデル管理	検証環境・配布・監視	版固定、段階配布、切戻し
外部連携	公開情報検索・許可されたクラウド	データ区分と送信先を制限

初期構成では、生成 AI は履歴 DB や文書の読み取り専用の複製へアクセスする。設備データを取得する接続口と、設備へ書き込む接続口を分離する。読み取りでも過大な問い合わせが設備やネットワークに負荷を与え得るため、取得頻度とキャッシュを設定する。

RAG に必要なのは文書の切り分けだけではない

文書を取り込み、テキスト・図表を抽出し、型番・版・適用日・機密区分・アクセス権を付与する。全文検索と意味検索を組み合わせ、候補を絞ってモデルに渡す。型番や規格番号のように一文字の差が重要な情報は、意味が似ているだけでは代替できない。

検索段階で利用者の閲覧権限を適用する。回答後に隠す方法では、権限のない文書が生成処理へ混入する可能性が残る。文書更新時には索引、キャッシュ、引用先を更新し、旧版と現行版を区別する。

学習と検索の使い分け

頻繁に変わる仕様や現行マニュアルは、出典を提示できる検索連携が適する。書式・用語・分類・回答の振る舞いをそろえる目的では、追加学習が候補になる。独自の事前学習は大量の良質なデータ、学習資源、評価・保守体制が必要であり、最初の手段にしない。

ベクトル検索の索引を作ることは、通常、LLM の重みを更新する学習とは異なる。RAG でも、検索漏れ、誤った図表抽出、古い文書、誤引用は起こり得る。生成モデルを交換する前に、どの工程が誤りを生んだかを分解して測定する。

11 モデル候補と選び方

以下は用途別の検証候補であり、性能順位ではない。配布条件、追加学習、再配布、商用利用、対応ランタイムは、採用するモデルの版ごとに確定する。API 提供と重みの配布、企業オンプレミス契約は別の提供形態である。

候補	確認できる特徴	製造業での検証項目
Qwen3.8-27B	27B の画像・動画対応モデル。公式モデルカードは Apache 2.0。[27]	図面・表、日本語、数値、量子化後の精度
gpt-oss-20b／120b	ローカル実行可能な公開重み。公式に 16GB／80GB メモリでの動作を説明。[26]	文章推論、ツール利用、負荷時の速度。画像は別処理を設計
ELYZA Thinking 1.0	llm-jp-4 由来の 33B／32B-A3B を 2026 年 10 月公開。[28]	社内用語、日本語指示、長い推論の待ち時間
Ministral 3	3B／8B／14B、画像理解、Apache 2.0。[29]	小型構成での抽出・分類、日本語実務精度
tsuzumi 2	NTT の企業向けモデル。PRIMERGY でオンプレミス動作検証完了の発表。[30,32]	契約、専用語、導入・保守条件
PLaMo 3.1 Prime	2026 年 10 月、オンプレミスの Deep Research にも対応と発表。[31]	社内検索、根拠提示、閉域時の取得対象
cotomi	NEC の生成 AI アプライアンスに搭載。[33]	日本語、RAG、提供機能と更新条件

モデル選定の順序

まず 10〜30 種類の実タスクを選び、答えと根拠を専門家が確認する。次に実データの難例を含む評価セットを作る。その後、複数モデルを同じ文書、同じ検索条件、同じ出力制限で比較する。量子化率や推論モードを変えた際は、型番・単位・否定・境界条件への影響を再確認する。

巨大な MoE モデルでは、一度に計算へ使うパラメータ数が小さくても、モデル全体の重みの保管・転送が不要になるわけではない。「Active 3B」を「3B モデルと同じメモリ」と読み替えない。国産であることも、日本語性能、完全閉域、学習データの権利処理を一括して保証するものではない。

汎用ベンチマークの上位モデルが、工場の型番検索や少数分類で最も経済的とは限らない。逆に、複雑な設計推論や特許の技術対比では小型モデルが不足する場合がある。難しい案件だけ大きなモデルへ回す構成も比較対象にする。

12 ハードウェアの選択肢と公表価格

仕様は 2026 年 11 月 11 日までに参照したメーカー・販売者の公表情報による。下表の用途は本稿の提案であり、同時利用人数や速度を保証するものではない。

構成	確認した主な仕様	用途と注意
一般 PC／AI PC	CPU・GPU・NPU 構成による	小型モデル、音声、文書前処理。NPU の TOPS だけで LLM 対応を判断しない
DGX Spark 128GB 構成	統合メモリ 128GB、273GB/s、Arm CPU、DGX OS。[21]	部門 PoC、推論・開発。128GB すべてを重みには使えない
RTX PRO 6000 Blackwell Max-Q	GPU メモリ 96GB、最大 300W、1,792GB/s。[23]	大きめのモデルや VLM。電源・冷却・ホスト側メモリも設計
HP ZGX Fury G1n	GB300、コヒーレントメモリ 748GB。CPU 側 496GB＋GPU 側 252GB。[24,45]	研究・大型モデル。容量とメモリ階層の速度差を考慮
Jetson Thor T5000	メモリ 128GB、電力 40〜130W、FP4 疎演算 2,070TFLOPS。[25]	ロボット・現場推論。開発キットと産業用完成機を区別
複数 GPU サーバ	選択する GPU・接続・冗長化による	共有サービス、同時利用。構内設備・運用体制が必要

価格は時点と販売条件を付けて扱う

NTTPC の DGX Spark ページには税込 893,200 円の表示がある一方、新規注文受付停止も明記されていた。これは参照価格であり、その価格で調達可能と判断しない。[22]

日本 HP が 2026 年 10 月 26 日に発表した ZGX Fury G1n の希望販売価格は税込 28,490,000 円からで、販売開始日は 10 月 27 日とされる。Windows 対応は同発表で 2026 年 12 月以降の予定とされており、調査時点の対応済み仕様として扱わない。[24]

調達前に確認する実務条件

デスクサイド機を製造ラインへ置く場合は、粉塵、油煙、温度、振動、保全作業、電源停止、盗難を考える。産業用筐体や別室設置が必要になることもある。Arm 版と x86 版で動くソフトやドライバは異なり、Windows 利用者がブラウザ経由で Linux サーバを使う構成も選べる。

装置の最大電力、電源ユニット容量、実測の消費電力は別の値である。GPU の TDP だけから施設全体の電気代を計算せず、ホスト、ストレージ、ネットワーク、空調、待機時間を含めて測定する。

13 メモリ容量と同時利用の見積もり

重みが収まる容量と実用時の容量は異なる

重みだけの概算は「パラメータ数×ビット数÷8」で計算できる。実際には量子化の補助情報、非量子化部分、推論用バッファ、入力処理、KV キャッシュ、OS などの余裕が必要になる。以下は十進 GB による算術値であり、機器の必要容量表ではない。

パラメータ数	4 ビット	8 ビット	16 ビット
8B	4GB	8GB	16GB
27B	13.5GB	27GB	54GB
32B	16GB	32GB	64GB
70B	35GB	70GB	140GB
120B	60GB	120GB	240GB

例えば 27B を 4 ビットにすれば理論上の重みは 13.5GB だが、「16GB の GPU で長文・画像・多人数が快適」とは結論できない。量子化形式や対応カーネルによっても必要容量・速度・精度が変わる。

長い入力と同時利用が容量を増やす

通常の Attention を仮定した例では、KV キャッシュの容量は、おおむね「層数×KV ヘッド数×ヘッド次元×K と V の 2 個×保存バイト数×トークン数×同時系列数」で増える。仮に 32 層・8KV ヘッド・128 次元・16 ビット保存なら、32,768 トークンで 1 系列あたり約 4GiB になる。これは説明用の仮定であり、ハイブリッド方式等のモデルにはそのまま当てはまらない。

モデルが 100 万トークンを仕様として受け付けても、その長さを手元の機器で実用速度で処理できるとは限らない。長文全投入と RAG による絞り込みを比較する。社内資料の版・権限管理の点でも、すべてを長い入力へ詰める方式が常に優れるわけではない。

購入前に測定する性能

試験項目	測定方法
応答	最初の表示までの時間、回答完了時間をそれぞれ測る
同時利用	1・5・10・20 要求など、業務に合う負荷で比較する
長文・画像	実際のページ数・解像度・回答長で測る
持続運転	連続処理、発熱、待機列、メモリ不足を確認する
業務処理能力	1 時間あたりに人が承認できた案件数と修正時間を測る

理論 TOPS や tokens/s だけで比較しない。回答の短さ、正確さ、再試行、確認負荷を含む「使える結果が何件得られたか」を最終指標にする。

14 ソフトウェアと継続運用

試用と共有サービスで必要な機能が違う

個人検証には Ollama やローカル実行アプリが利用しやすい。部署内の共有では、推論エンジンに加え、認証、利用枠、文書権限、監査、監視、バックアップが必要になる。多数の同時要求や複数 GPU を扱う場合は、vLLM 等の配信基盤や商用製品を検討する。[35,36]

Ollama にはクラウド機能もあり、公式 FAQ はローカル専用モードの設定を説明している。「アプリ名がローカル向けだから外部通信しない」とは判断しない。vLLM の公式文書も、内部通信の分離や公開ポートの制限、API キーだけに依存しない保護を求めている。[35,36]

運用で管理する対象

対象	管理する情報	運用上の対応
モデル	出所、ハッシュ、ライセンス、量子化版	審査済み版のみ配布
ソフトウェア	OS、ドライバ、依存ライブラリ	互換性検証、更新計画
文書と索引	版、適用日、権限、削除履歴	更新と削除を検索に反映
プロンプト	テンプレート、ツール権限	変更管理、回帰評価
実行環境	GPU 負荷、待ち時間、温度	監視、容量追加、復旧
ログ	入力、出典、出力、操作履歴	保存期間、閲覧制限
現場機器	設置先、モデル版、最終通信	遠隔管理、切戻し、交換

モデル更新を現場へ一斉配布しない

新モデルはまず検証用の過去問題で評価し、限定部署・限定設備で試す。品質が下がった場合に以前のモデルへ戻せるよう、重み、推論設定、検索索引、文書版を関連付けて保存する。更新で回答が変わる場合は、利用者へ変更内容を説明する。

完全オフラインでは、モデル取得、脆弱性修正、ライセンス確認、時刻同期、証明書更新、障害ログ回収をどう行うかが課題になる。保守担当者が USB や一時ネット接続で対応するなら、その手順も含めて閉域運用を定義する。

内製と外部支援の分担

内製すべき中心は、業務の正解、データ利用権、評価セット、受入基準である。GPU 構築や監視を委託しても、業務の正誤判定までは丸投げできない。逆に、継続運用人材が不足する企業が推論基盤をすべて自作する必要もない。商用アプライアンスの保守範囲と出口条件を比較する。[32,33,43,44]

15 費用対効果をどう判断するか

初期費用と運用費用を分けて積み上げる

初期費用には、端末・サーバ・ストレージ、ネットワーク、ソフトウェア、文書整備、既存システム連携、評価、安全設計、教育を含める。運用費用には、電力、冷却、保守、ライセンス、モデル更新、資料更新、監視、利用者対応、専門家による確認を含める。

API 料金が発生しなくても計算資源と運用は無料にならない。小規模・低頻度の利用では共有クラウドの方が経済的な場合があり、機密性・応答時間の要件によっては費用差より業務実行の可否が重要になる。

部門用文書 AI の計算例

以下は相場でも見積書でもなく、投資判断のための仮定である。利用者 100 人、月 20 日、時間単価 6,000 円、初期投資 600 万円、月額運用費 30 万円とする。削減した時間のうち、実際に追加業務や費用削減へ結びつく割合を「実現率」とする。

仮定	保守的	標準	高利用
1 人 1 日短縮時間	5 分	15 分	25 分
月の総短縮時間	約 167 時間	500 時間	約 833 時間
実現率	20%	30%	50%
月の便益換算	20 万円	90 万円	250 万円
運用費差引後	マイナス 10 万円	60 万円	220 万円
単純回収期間	回収できない	10 か月	約 2.7 か月

標準ケースでは、100 人×20 日×15 分÷60×30%×6,000 円＝月 90 万円。ここから運用費 30 万円を引く。3 年間の費用は 600 万円＋30 万円×36＝1,680 万円となる。税・割引率・追加開発・入替費用はこの簡易例に含めない。

時間の余裕が生じただけなら、便益は能力余力であり、そのまま現金支出の削減にはならない。残業削減、外注削減、案件数増加、停止損失回避など、実際の成果へ結びついた分を別に確認する。

現場 AI の経済性

検査では、再検査・不良流出・手直しの費用を合わせる。予兆保全では、警報が増えたことで交換や点検が過剰になっていないかを見る。設備停止の回避額は、実際のボトルネック能力や受注状況に依存する。停止時間に売上単価を単純に掛けて便益を過大評価しない。

クラウドとの比較も、同じ回答品質、利用者数、ピーク負荷、保守水準、停止時対応をそろえる。年間の承認済み処理件数あたりの総費用を算出すると、トークン単価だけの比較より判断しやすい。

16 セキュリティと知的財産の管理

ローカル設置だけでは解決しない問題

主な残存リスクは、内部の権限逸脱、端末盗難、ログやバックアップの持出し、外部連携への誤送信、不正なモデルや依存ソフト、検索資料に埋め込まれた指示による操作誘導である。OWASP は、情報漏洩、供給網、プロンプトインジェクション、過剰なエージェント権限等を別々のリスクとして整理している。[40]

工場のネットワークでは、業務・重要度に応じたゾーン分割と責任分界を設計する。経済産業省の工場セキュリティ資料と NIST の OT ガイダンスは、その検討の基礎になる。生成 AI サーバの便利さを理由に、制御ネットワークへ無制限の接続を与えない。[37–39]

管理対象	具体的に確認すること
外向き通信	OCR、検索、埋め込み、推論、監視、ライセンス認証の接続先
社内権限	部署・案件・共同研究ごとの検索権限と回答への反映
エージェント	読み取り、書き込み、設備操作、外部送信の権限分離
ログ・重み	機密を含むログ、追加学習モデル、索引、バックアップの保護
外部委託	管理者アクセス、遠隔保守、再委託、障害時のデータ回収
終了時	文書・索引・重み・評価結果を返却できる契約と削除確認

知財部門が関与すべき事項

未公開発明や営業秘密を扱うには、アクセス制限、秘密であることの認識、取扱いルールを整える。経済産業省の営業秘密管理指針も参照し、ローカル AI という技術構成だけを秘密管理の代替にしない。[41]

自社が保有する資料でも、共同研究先、顧客、装置メーカー、論文データベース等の契約で利用が制限される場合がある。検索利用、追加学習、委託先への開示、重みの共有、派生モデルの再配布を分けて確認する。公開重みのモデルも、版ごとのライセンスや付随する条件を調達審査に含める。

AI の提案と技術者の発案・選択・実験・修正を区別して記録すると、発明の成立経緯や出願内容の確認に役立つ。外部 AI へ入力したという事実だけで一律に新規性喪失と決めず、実際の公開状態、契約、アクセス可能性を個別に確認する。

技術的な防御と、契約・権利・品質の確認は別の管理である。たとえばデータが社外に出なくても、誤った対比表や他者の秘密を権限外で利用する問題は残る。

17 評価方法と 90 日間の導入計画

評価セットを先に作る

文書 AI では、代表質問に加え、答えがない質問、旧版と新版が食い違う質問、機種違い、表の脚注、否定条件、数値・単位、権限外資料を含める。初期は 100〜300 問程度を実務者と整備する案が考えられるが、必要件数は用途の種類と許容する誤りによって変わる。

検索で根拠が得られたか、根拠に沿って回答したか、引用が正しいか、不明時に回答を控えたか、最終的に業務が完了したかを分けて評価する。LLM による自動採点だけで完結せず、重要案件は業務専門家が採点する。

期間の目安	作業	次へ進む条件
1〜2 週	対象業務・情報区分・現状工数を確認	業務責任者と基準値が決まる
3〜4 週	データ整理、評価セット、構成比較	権利・版・アクセス権が整理できる
5〜8 週	複数モデル比較、利用者試験	品質と速度が業務基準を満たす
9〜10 週	実業務で限定運用、費用を測定	確認込みの便益が認められる
11〜12 週	障害・復旧・更新・権限試験	担当者、保守、停止条件が確定する
13 週	継続・修正・中止を判断	本番予算と運用責任を承認する

この 90 日案は、文書検索等の限定業務を想定する。設備制御、危険工程、規制対象の品質判定などは、必要な工程検証や安全評価を別途計画する。

受入基準の例と限界

例として、重要質問では出典のない断定を許可しない、型式と文書版を必ず表示する、権限外資料を返さない、復旧手順を実機で確認する、といった条件を設定する。合格値は事業と品質の責任者が決めるもので、本稿の例は共通規格ではない。

300 件で重大誤りがゼロでも、未知の状況で誤りがない証明にはならない。独立した同条件の試行を仮定すれば、ゼロ件観測時の誤り率の 95%片側上限は、おおむね 3÷件数であり、300 件なら約 1%である。稀な危険事象や環境変化は別に検証する必要がある。

中止や再設計を選ぶ条件

根拠資料が整わない、確認時間を含めると短縮しない、既存検索で同程度の成果が出る、障害対応の責任者がいない場合は、モデルを大型化する前に業務やデータを見直す。PoC の成功条件は「動いた」ことではなく、運用できる品質と経済性を確認できたことである。

18 業種別の導入優先順位

同じ製造業でも、工程の連続性、品種数、情報の機密性、不良の影響が異なる。以下は一般的な設計案であり、各社の具体的な工程や規制に応じて変更する。

業種	着手候補	ローカル化の理由	主要な検証点
自動車・部品	外観、組付け、保全、設計変更	短周期、設備情報	品種差、誤検出、追跡性
半導体・電子	欠陥分類、装置ログ、工程解析	画像量、レシピ秘匿	ロット差、微小欠陥、時刻整合
化学・素材	配合検索、実験解析、条件最適化	ノウハウ、連続操業	単位、物性条件、外乱
機械・装置	図面・仕様、保守、PLC 支援	顧客仕様、技術情報	改訂管理、機種ごとの適用
食品	異物・包装、衛生記録、保全	現場応答、工程記録	照明、品種・季節、衛生手順
医薬・医療機器	記録検索、照合、逸脱整理	機密情報、証跡	手順変更、検証、監査
金属・鉄鋼	表面検査、設備異常、原単位	高速処理、連続運転	熱・粉塵、設備変動
紙・繊維	連続面検査、品種切替、条件検索	映像量、操業継続	速度、模様差、材料ロット
多品種の中小企業	見積仕様、図面検索、作業記録	顧客秘密、担当者不足	資料整備、保守負荷、利用頻度

企業規模による構成の違い

小規模企業は、対象業務を一つに絞り、保守付きの端末・アプライアンスや既存の検査パッケージから始める選択肢がある。担当者が一人しかいない場合は、内製の自由度より復旧手段やベンダーの引継ぎを重視する。

中堅企業は、研究・設計・知財の共通文書基盤と、工場の現場 AI を段階的に整える。大企業は、共通の認証・評価・配布基盤を持ちながら、工場ごとの設備・品質条件を維持する。モデルを統一することと、すべての工場で同じ判定基準を使うことは別である。

現場への定着

画面は担当者が使う設備番号や作業番号を中心に構成し、出典の該当箇所へすぐ戻れるようにする。回答の良否を一操作で報告できるようにし、現場担当者が負担を感じる長文説明や大量の確認操作を避ける。品質を守る確認を残しながら、日常業務へ自然に組み込む。

19 調達条件と今後の見通し

提案依頼に含める項目

領域	ベンダーへ求める回答
実行場所	各処理とデータの保存先を示す構成図
オフライン	外部通信を切ったときに利用できる全機能
性能	自社データでの精度、同時要求、長文・画像の測定結果
利用権	モデル・文書・追加学習・再配布の条件
権限	既存 ID、文書 ACL、部署・案件間の分離方法
安全と操作	読み取り・書き込み・設備制御の権限と承認
更新と保守	対応 OS、モデル更新期限、脆弱性対応、復旧時間
費用	3 年間の全費用、増設、追加利用、終了・移行費
退出可能性	文書、索引、評価結果、設定のエクスポート
責任分担	誤回答、停止、情報事故、品質不良時の対応範囲

今後に関する本稿の見通し

第一に、文書、画像、音声、設備時系列を組み合わせる現場支援が増えると考えられる。第二に、全処理を一つの巨大モデルで行うより、専用モデルと小型生成 AI を組み合わせる設計の重要性が高まる。第三に、複数の工場へ安全に更新を配布し、性能の変化を検出する運用能力が競争力を左右する。これらは確認した技術・製品動向からの見通しであり、導入率の予測値ではない。

PLaMo の閉域での社内調査機能や、オンプレミスで利用できる国産モデルの提供は、社内情報を扱う選択肢を広げている。ただし Web 検索を行えば外部接続が発生するため、社内限定調査とインターネット調査を同じ「オフライン機能」と説明してはならない。[31,32]

最初に着手する三案件

推奨案は、①研究・設計・知財の機密文書検索、②一つの設備群での保全記録検索、③既存検査工程での画像 AI 比較である。同じモデルを強制せず、それぞれに業務責任者、正解データ、現状工数、受入条件を置く。設備変更を伴わない二案件から学び、品質・制御へ段階的につなぐ。

調査方法と解釈の限界

企業・製品開発元の発表、当事者の事例、官公庁資料、公式技術文書を中心に調べた。公表資料は発表当時の状況を示し、現在の全社展開や個別契約、非公開の故障率・保守費まで明らかにするものではない。引用先の改善値を他社へそのまま適用せず、機密情報の具体的な扱いと投資採否は自社条件で確かめる。

参考文献 1 から 8

全資料の参照日 2026 年 11 月 11 日。日付は公表日または対象資料の版を示す。随時更新ページは内容・価格・提供条件が変わる場合がある。

[1] 三菱電機「エッジデバイスで動作する製造業向け言語モデルを開発 リリース全文」。2025 年 6 月 18 日。

https://www.mitsubishielectric.co.jp/ja/pr/2025/pdf/0618-a.pdf

[2] 日本 HP「検証事例 大分キヤノンマテリアル株式会社 HP ワークステーションによる高精度なローカル LLM 運用を体感」。2026 年 6 月 12 日。

https://jp.ext.hp.com/techdevice/technology/oita-canon-material-local-llm-poc/

[3] Sparticle GBase「荏原製作所 GBase OnPrem 導入事例」。日付記載なし。

https://gbase.ai/onprem/showcase/ebara/

[4] 別川製作所「スマートファクトリーに適応した生成 AI による自由対話システムを開発」。対象年度 2023 年度。

https://bcreer.betsukawa.co.jp/case/industry_academia_collaboration/202404_01.php

[5] パナソニック HD「パナソニック HD とストックマーク Panasonic-LLM-100b 開発で協業」。2024 年 7 月 2 日。

https://news.panasonic.com/jp/press/jn240702-3

[6] 経済産業省 GENIAC「Use Case vol.1 パナソニック HD とストックマーク 国内最大規模の企業特化 LLM 開発」。日付記載なし。

https://www.meti.go.jp/policy/mono_info_service/geniac/geniac_magazine/usecase_01.html

[7] Siemens「Audi ensures quality with Siemens AI and Edge」。日付記載なし。

https://www.siemens.com/en-gb/company/insights/audi-ai-industrial-edge-quality-assurance/

[8] Audi Japan「アウディ 生産における人工知能の展開を拡大」。2026 年 2 月 5 日。

https://www.audi-press.jp/press-releases/2026/s5n52g0000005fsr.html

参考文献 9 から 16

全資料の参照日 2026 年 11 月 11 日。日付は公表日または対象資料の版を示す。随時更新ページは内容・価格・提供条件が変わる場合がある。

[9] 横河電機 ENEOS Materials「Yokogawa's Autonomous Control AI Is Officially Adopted for Use at an ENEOS Materials Chemical Plant」。2023 年 3 月 30 日。

https://www.yokogawa.com/news/press-releases/2023/2023-03-30/

[10] 横河電機 JSR「世界初 AI による自律制御で化学プラントを 35 日間連続制御」。2022 年 3 月 22 日。

https://www.yokogawa.co.jp/news/press-releases/2022/2022-03-22-ja/

[11] 横河電機「A world's first is now routine ENEOS Materials Successfully Achieves Autonomous Control at a Plant Using AI」。日付記載なし。

https://www.yokogawa.com/mea/library/resources/references/successstory-eneos-materials/

[12] Siemens「Drivetrain Analyzer Onsite Siemens introduces new AI-powered on-premises analytics for industrial drives」。2026 年 3 月 26 日。

https://press.siemens.com/global/en/pressrelease/drivetrain-analyzer-onsite-siemens-introduces-new-ai-powered-premises-analytics

[13] Siemens「Siemens Industrial Copilot expanded adopted by thyssenkrupp」。2024 年 11 月 12 日。

https://press.siemens.com/global/ro/node/7042

[14] NVIDIA Siemens「Siemens and NVIDIA Expand Partnership to Accelerate AI Capabilities in Manufacturing」。2025 年 6 月 11 日。

https://investor.nvidia.com/news/press-release-details/2025/Siemens-and-NVIDIA-Expand-Partnership-to-Accelerate-AI-Capabilities-in-Manufacturing/default.aspx

[15] NVIDIA「Siemens Accelerates Product Development and Innovation With Industrial AI」。日付記載なし。

https://www.nvidia.com/en-eu/case-studies/siemens-accelerates-product-development-and-innovation-with-industrial-ai/

[16] OKI「ローカル 5G とスマート工場ソリューションの外部化 OKI テクニカルレビュー No.240」。2022 年。

https://www.oki.com/global/ja/technologies/otr/2022/n240/r04/

参考文献 17 から 24

全資料の参照日 2026 年 11 月 11 日。日付は公表日または対象資料の版を示す。随時更新ページは内容・価格・提供条件が変わる場合がある。

[17] 武蔵精密工業「トヨタ自動車向けに AI 外観検査装置を追加導入 検査の効率化で電動化需要に対応」。2022 年 12 月 14 日。

https://www.musashi.co.jp/newsrelease/ai_3.html

[18] ファナック「IoT 新機能 FIELD system Basic Package の特長と新機能について」。2025 年 7 月。

https://www.fanuc.co.jp/ja/product/new_product/2025/202507_fsbp.html

[19] NVIDIA「Pegatron Scales Factory Operations With Visual AI Agents and Digital Twins」。日付記載なし。

https://www.nvidia.com/en-us/case-studies/pegatron-scales-factory-operations-with-visual-ai-digital-twins/

[20] Microsoft「Schaeffler unlocks factory insights with Microsoft Cloud for Manufacturing」。日付記載なし。

https://www.microsoft.com/en/customers/story/20385-schaeffler-microsoft-fabric

[21] NVIDIA「DGX Spark Hardware Overview User Guide」。随時更新。

https://docs.nvidia.com/dgx/dgx-spark/hardware.html

[22] NTTPC コミュニケーションズ「NVIDIA DGX Spark 製品情報と価格」。随時更新 受付停止表示を含む。

https://www.nttpc.co.jp/gpu/lineup/project_digits.html

[23] NVIDIA「RTX PRO 6000 Blackwell Max-Q Workstation Edition Specifications」。随時更新。

https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-6000-max-q/

[24] 日本 HP「高度な AI 開発に特化した HP ZGX Fury G1n AI Station を発表」。2026 年 10 月 26 日。

https://jp.ext.hp.com/info/newsroom/2026/20261026/

参考文献 25 から 32

全資料の参照日 2026 年 11 月 11 日。日付は公表日または対象資料の版を示す。随時更新ページは内容・価格・提供条件が変わる場合がある。

[25] NVIDIA「Jetson Thor Series Modules Quick Specs」。随時更新。

https://www.nvidia.com/en-in/autonomous-machines/embedded-systems/jetson-thor/

[26] OpenAI「Introducing gpt-oss」。2025 年 8 月 5 日。

https://openai.com/index/introducing-gpt-oss/

[27] Qwen「Qwen3.8-27B Model Card」。公開日 2026 年 8 月 14 日 モデルカード随時更新。

https://huggingface.co/Qwen/Qwen3.8-27B

[28] ELYZA「ELYZA RSI Research 設立 基盤モデル 評価 自律改善に関する研究成果を公開」。2026 年 10 月 2 日。

https://elyza.ai/news/2026/10/02/rsiresearch

[29] Mistral AI「Introducing Mistral 3」。2025 年 12 月 2 日。

https://mistral.ai/news/mistral-3/

[30] NTT「更なる進化を遂げた NTT 版 LLM tsuzumi 2 の提供開始」。2025 年 10 月 20 日。

https://group.ntt/jp/newsrelease/2025/10/20/251020a.html

[31] Preferred Networks「国産生成 AI 基盤モデル PLaMo Prime に Web 検索 Deep Research 搭載」。2026 年 10 月 27 日。

https://www.preferred.jp/ja/news/pr20261027

[32] NTT 西日本 エフサステクノロジーズ「Private AI Platform on PRIMERGY で純国産 LLM tsuzumi 2 が利用可能に」。2026 年 8 月 26 日。

https://www.ntt-west.co.jp/news/2608/260826a.html

参考文献 33 から 40

全資料の参照日 2026 年 11 月 11 日。日付は公表日または対象資料の版を示す。随時更新ページは内容・価格・提供条件が変わる場合がある。

[33] NEC「NEC Generative AI Appliance Server」。随時更新。

https://jpn.nec.com/gpu/cotomi/index.html

[34] Intel「Manufacturing AI Suite Open Edge Platform Documentation」。2026.2 版。

https://docs.openedgeplatform.intel.com/2026.2/ai-suite-manufacturing.html

[35] Ollama「FAQ ローカル専用設定とネットワーク設定」。随時更新。

https://docs.ollama.com/faq

[36] vLLM「Security 公式ドキュメント」。随時更新。

https://docs.vllm.ai/en/latest/usage/security/

[37] 経済産業省「中小規模の製造事業者向けに工場のセキュリティを確保するための解説書を策定」。2025 年 4 月 11 日。

https://www.meti.go.jp/press/2025/04/20250411005/20250411005.html

[38] 経済産業省「工場システムのセキュリティ対策ガイドライン 別冊 スマート化を進める上でのポイント」。2024 年 4 月 4 日。

https://www.meti.go.jp/press/2024/04/20240404002/20240404002.html

[39] NIST「SP 800-82 Revision 3 Guide to Operational Technology Security」。2023 年 9 月。

https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-82r3.pdf

[40] OWASP「2025 Top 10 Risk and Mitigations for LLMs and Gen AI Apps」。2025 年版。

https://genai.owasp.org/llm-top-10/

参考文献 41 から 45

全資料の参照日 2026 年 11 月 11 日。日付は公表日または対象資料の版を示す。随時更新ページは内容・価格・提供条件が変わる場合がある。

[41] 経済産業省「営業秘密を守り活用する 営業秘密管理指針等」。管理指針 2025 年 3 月改訂。

https://www.meti.go.jp/policy/economy/chizai/chiteki/trade-secret.html

[42] 特許庁「AI 関連技術に関する特許審査の事例について」。ページ更新 2026 年 5 月 25 日。

https://www.jpo.go.jp/system/laws/rule/guideline/patent/ai_jirei.html

[43] 日立製作所「生成 AI の本格的な業務活用に向け 業務特化型 LLM 構築 運用サービスを提供開始」。2024 年 8 月 29 日。

https://www.hitachi.com/ja-jp/press/articles/2024/08/0829a/

[44] 富士通「Fujitsu Kozuchi Enterprise AI Factory」。随時更新。

https://global.fujitsu/ja-jp/offering/kozuchi-enterprise-ai-factory

[45] 日本 HP「HP ZGX Fury G1n AI Station 製品詳細 スペック」。随時更新。

https://jp.ext.hp.com/workstations/zgx_fury_ai_station/