Harmonic Society

WHITE PAPER ｜ 2026年版

ローカルLLM
企業導入ガイド

社外に出せないデータのための、セキュアな社内AI

― クラウドに送れない機密情報を、AIで活用する実践書 ―

発行：Harmonic Society株式会社
千葉市｜ホームページ制作・AI導入支援・IT顧問

本書でわかること

- 情報漏えいリスクを「構造的に」なくすという選択肢
- クラウドAPIとの費用比較と損益分岐点
- 予算別ハードウェア構成とモデルの選び方
- 5段階の導入ロードマップと運用チェックリスト

2026年7月発行

INTRODUCTION

はじめに ―「AIは使いたい。でも、このデータは外に出せない」

「図面や顧客リストを外部のAIに送るのは抵抗がある」（精密金属加工の経営者）
「個人情報を含む相談内容を外部に出さずに、文書の下書きを作りたい」（社会保険労務士事務所）

クラウド型AIは強力ですが、入力した内容は必ず外部サーバーに送信されます。**「そもそも社外に出してよいのか」という問いが残る業務は、どの会社にも必ずあります。**

契約書、財務データ、人事情報、設計図面、顧客の個人情報――。この問いへのひとつの答えがローカルLLMです。AIモデルを自社のPCやサーバーで動かすため、**データが社外に一切出ません。**

かつてはサーバー級の設備が必要でしたが、いまは20〜30万円のPC1台と無料のオープンソースソフトで、実務に使える環境が作れます。

本書の情報について：モデル名・GPU型番・価格・API料金は執筆時点（2026年7月、一部2026年3月時点の調査データ）の情報です。変化の速い分野のため、最終判断の際は最新情報をご確認ください。事例の数値はモデルケースを含みます。

本書のなりたち

28本
検証と実運用に基づくローカルLLM解説記事

5社
中小企業の導入事例ケーススタディ

1冊
企業の導入判断に必要な順序で再構成

ゴール：読み終えたとき「自社に必要か・いくらかかるか・何から始めるか」を自分の言葉で説明できること

「自社の中でAIを動かす」という選択肢

ローカルLLM ＝「大規模言語モデル（LLM）を自社のPC・サーバーで、完全に自社環境の中で動かす使い方」 クラウド型AIとの違いは、突き詰めればデータの通り道です。

図表1｜ローカルLLMとクラウドLLMの基本比較

比較項目	ローカルLLM	クラウドLLM（ChatGPT等）
データの扱い	**社内で完結**	外部サーバーに送信
コスト構造	初期投資＋電気代（固定費）	月額・従量課金（変動費）
ネットワーク	**不要（オフライン可）**	必須
カスタマイズ性	**高い（RAG・追加学習）**	制限あり
最高性能	最先端モデルには及ばない	**最高水準**
導入の手間	やや高い	**低い**

なぜ「今」なのか — 小型モデルの性能が実務水準に達した

1 オープンソースモデルの急速な進化

7B（70億パラメータ）クラスの小型モデルでも、**2024年の70Bクラスに匹敵する性能**を持つものが登場しています。

日本語に強いQwen3、GoogleのGemma 3、MetaのLlama 4など、無料で商用利用できる選択肢が揃いました（ライセンス確認は第5章）。

2 実行ツールの成熟

OllamaやLM Studioといった無料ツールにより、**モデルのダウンロードから対話開始まで数十分。**

ChatGPTのような画面（Open WebUI）を社内に用意すれば、エンジニアでない社員もブラウザから使えます。

かつてはサーバー級の設備が必要だった環境が、いまは **20〜30万円のPC1台＋無料のオープンソースソフト** で作れます。

仕組みは「2つの数字」だけ押さえれば十分

VRAM（GPUメモリ）容量 ＝ 動かせるモデルの最大サイズ

ローカルLLM選定で最も重要な数字は、GPUの演算性能ではなく**VRAM容量**です。機材選びは下表の3行を見れば決められます。

量子化 ＝ モデルの圧縮技術

標準的な「Q4_K_M」形式なら、**元モデルの約97%の精度を維持したままサイズを75%削減。**7Bモデルなら約3.5GB——家庭用ゲーミングPCでも動くサイズになります。

図表2｜VRAM容量と動かせるモデルの目安（4bit量子化時）

VRAM	動作するモデル	代表的なGPU	想定用途
8GB	7Bクラス	RTX 4060	FAQ応答・要約・下書き
16GB	13〜30Bクラス	RTX 4060 Ti 16GB	**実務レベルの文書作成**
24GB	70Bクラス	RTX 4090	高度な推論・複雑なタスク

この2つ（VRAMと量子化）を知っていれば、技術的な詳細に立ち入らなくても導入判断と機材選定はできます。

図表3｜中小企業のローカルLLM活用事例（当社ブログ掲載のケーススタディより）

企業	用途	構成	効果
製造業A社（45名）	品質検査レポート自動生成	Qwen3 14B＋社内サーバー	作成時間75%減（30〜45分→5〜10分）・月60時間削減。運用費は電気代込み月約3,000円（クラウド見積は月5万円）
会計事務所B社（12名）	顧問先からの定型質問対応	Gemma 3 12B＋RAG	初回応答 平均2時間→15分、対応工数 月40時間削減
IT企業C社（20名）	コードレビュー・文書生成	Mistral系＋VS Code連携	レビュー1件30分→10分、クラウドAPI月8万円が不要に
不動産D社（15名）	物件紹介文の自動生成	Qwen3 7B（営業部PC）	1件20〜30分→5分以下、掲載数1.5倍
物流E社（60名）	議事録・日報の自動要約	Gemma 3 4B＋Whisper	議事録45分→10分、既存PC流用で追加投資ゼロ

5社中4社が「データを外部に出したくない」を導入理由に。 クラウドに送れないデータこそが、ローカルLLMの主戦場です。

業種別の「刺さりどころ」と、成功企業の共通点

業種別の刺さりどころ

製造業
設計図面・特許関連文書の要約、品質レポート生成。オフラインの工場内でも動く

医療・介護
電子カルテの要約、医療文書の下書き。患者情報は法規制上、外部送信が困難

士業（会計・法律・社労士）
契約書レビュー、相談記録の整理。守秘義務との両立

受託開発・IT
顧客から預かったコードのレビュー・テスト生成。NDA違反リスクの回避

全業種共通
議事録の自動作成（音声認識Whisperもローカルで動く）、社内規程を参照して答えるFAQボット（第6章のRAG）

成功企業に共通する4つのパターン

1. **定型業務から始めている**
   レポート・FAQ・要約などのルーティンワークが入口
2. **人間によるチェックを組み込んでいる**
   AIの出力をそのまま使わない
3. **セキュリティを導入の決め手にしている**
   コスト削減は副次効果と位置づけ
4. **既存リソースを活用している**
   量子化モデル＋手持ちのゲーミングPCでまず検証

「リスク低減」ではなく「リスクの構造的排除」

クラウド型AIの対策は「信頼できる相手に預ける」こと。それでも残るリスクが4つあります。

1. **通信経路上のデータ傍受**
2. **提供者側の保管・管理**
3. **利用規約変更によるポリシー変化**
4. **海外サーバー保存の法的問題**

ローカルLLMはデータが物理的に社外へ出ないため、**この4つを構造的に排除します。**

ISMS（ISO27001）やPマークの監査でも「AIの処理は自社サーバー内で完結しており、外部送信はありません」と断言できます。**個人情報保護法・GDPR対応が求められる企業にとって、この説明のしやすさは大きな価値です。**

図表4｜ローカルLLMで安心して扱えるデータの例

- **顧客の個人情報（氏名・住所・購買履歴）**
- **契約書・法務文書（NDA含む）**
- **財務データ（売上・利益率）**
- **人事情報（評価・給与・相談記録）**
- **未公開の製品仕様・技術ノウハウ**
- **会議音声（文字起こしまでローカル処理）**

ただし「ローカルだから安全」は思い込み ― 対策の優先順位

ローカルLLMを社内ネットワークで公開することは、**社内にAIサーバーを1台立てること。**適切な設定を怠れば新たなリスクを生みます。

図表5｜セキュリティ対策の優先順位

優先度	対策	要点
最優先	**ネットワーク設定**	接続を社内LANに限定し、ファイアウォールで制御。インターネットへのポート公開は厳禁
重要	**認証の実装**	標準では認証機能がないため、Open WebUIのユーザー管理やリバースプロキシで認証を追加
重要	**モデルの安全な入手**	公式配布元（Hugging Face・Ollama公式）からのみ取得。非公式配布物は改ざんリスク
推奨	**ログと監査**	誰が何を入力したかの記録、保存期間の設定、四半期ごとのレビュー
推奨	**利用ポリシー**	利用範囲・アクセス権限・インシデント対応を1枚に明文化

あわせて：盗難リスクのあるノートPC運用はディスク暗号化を。対話ログには機密情報が残るため定期削除ルールを。この表はそのまま社内セキュリティ規程の下敷きに使えます。

費用構造は「初期投資＋月額固定」

図表6｜初期投資の3段階（執筆時点の目安）

構成	主要スペック	初期費用	動かせるモデル
エントリー	RTX 4060 Ti 16GB / RAM 32GB	**約25〜30万円**	7B〜13B
ミドル	RTX 4070 Ti Super / RAM 64GB	**約40〜50万円**	13B〜32B
ハイエンド	RTX 4090 / RAM 128GB	**約60〜80万円**	32B〜70B

意外に安い電気代

月約1,600円

エントリー構成・1日8時間×22日稼働の場合

ハイエンドでも月3,300円程度。ただし公平な比較には、減価償却費（4年）と保守工数（月2〜4時間）まで含めたTCO（総所有コスト）で見る必要があります。

図表7｜月額トータルコスト（TCOベース）

エントリー	**約12,300円**
ミドル	**約16,700円**
ハイエンド	**約27,900円**

減価償却＋電気代＋保守工数を含む月額換算

クラウドAPIとの損益分岐点 — 分かれ目は「利用量」と「必要な性能」

図表8｜利用規模別の月額コスト比較（試算）

利用シナリオ	高性能クラウドAPI	廉価クラウドAPI	ローカルLLM（エントリー）
ライト（3〜5人・月300万トークン）	約1,900円	**約110円**	約12,300円
スタンダード（10〜20人・月1,500万トークン）	約9,400円	約560円	約12,300円
ヘビー（20〜50人・月5,000万トークン）	約31,000円	約1,900円	**約16,700円（ミドル）**

試算の前提：2026年3月時点の料金に基づく試算です。API料金は改定が頻繁なため、具体的なサービス名は「高性能API／廉価API」と抽象化しています。高性能クラウドAPIで比較した場合の損益分岐の目安は「GPT-4oクラス相当の処理を1日530リクエスト以上」＝社内5〜10人が日常的に使う規模です。

ヘビー利用で高性能モデルが必要なら、ローカルが明確に優位。 月1.7万円の固定費で「使い放題」になります。結論の全体像は次ページへ。

損益分岐の結論と、見落としがちな「隠れコスト」

ライト利用

クラウドが圧倒的に安い

コスト目的だけのローカル導入は正当化できません

ヘビー利用×高性能が必要

ローカルが明確に優位

月1.7万円の固定費で使い放題に

廉価モデルで足りる業務

常にクラウドが安い

無理にローカル化する理由はありません

つまり純粋なコスト比較では決着しません。本当の分岐点は **「セキュリティ上、社外に出せないデータがあるか」** です。

見落としがちな「隠れコスト」— どちらもゼロではありません

クラウド側

API連携の開発費／レート制限対応／ドル建てゆえの為替リスク／仕様変更・値上げリスク

ローカル側

担当者の学習時間／ハードウェア故障時のダウンタイム／新モデル検証の手間

予算別・推奨構成

図表9｜予算別の始め方

予算	構成	できること
0〜5万円	既存PC（RAM16GB以上）でCPU推論、または Raspberry Pi 5（約2.7万円）	1〜4Bの小型モデルで検証。FAQボット程度なら常設も可
10〜30万円	RTX 4060 Ti 16GB搭載PC	**7B〜14Bで本格業務利用。中小企業に最もバランスが良い構成**
30〜80万円	RTX 4090等＋大容量RAM	27B以上・マルチモーダル・部署共有サーバー

GPU選びの鉄則は「演算性能よりVRAM容量」

VRAMが不足するとモデルの一部がメインメモリに退避（オフロード）し、**速度が2〜10倍遅くなります。**「動くが遅すぎて使われない」が最も多い失敗。実用の目安は**生成速度20トークン/秒以上**です。

MacBook Pro（128GBメモリ）という選択肢

Appleのユニファイドメモリは GPUとメモリを共有するため、通常は複数GPUが必要な**70Bクラスまで1台で動きます。**静音・低消費電力でオフィス向き。ただしメモリは後から増設できないため、購入時の容量選択が肝心です。

モデルの選び方 — 日本語業務ならまずQwen3

図表10｜用途別おすすめモデル（執筆時点）

用途	第一候補	補足
日本語の文書作成・チャット	Qwen3（8B/32B）	日本語性能トップクラス。Apache 2.0で商用も自由
軽量・省リソース	Gemma 3（4B/12B）	小型でも高性能。エッジ機器でも動作
コーディング支援	Qwen2.5-Coder / DeepSeek	コード特化。自社コードを外に出さない開発支援
長文処理・マルチモーダル	Llama 4 Scout	画像入力・超長文対応。要求スペック高め
議事録・要約	Qwen3 7B〜14B	Whisper（ローカル音声認識）と組み合わせ

選定の注意はライセンス

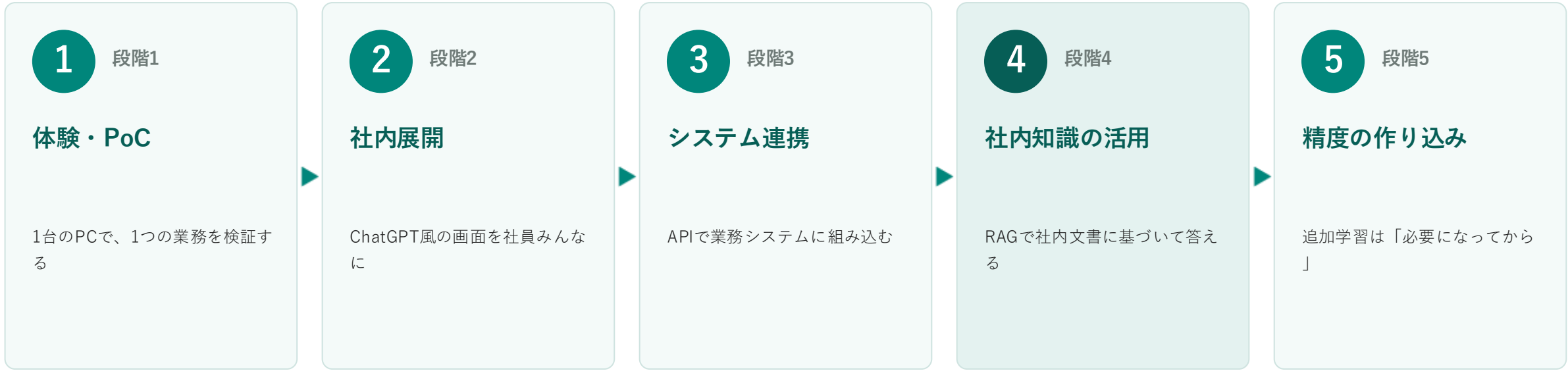
「オープンソース＝商用自由」ではありません。Apache 2.0やMITのモデルは自由に使えますが、非商用限定（CC-BY-NC）のモデルも存在します。導入前に必ず確認を。

量子化は迷ったら「Q4_K_M」

精度97%維持・サイズ75%減の標準形式です。日本語の品質を重視する業務では、一段上の「Q5_K_M」を選ぶとより安定します。

一足飛びに全社導入しない — 5段階の成熟度モデル

図表11｜ローカルLLM導入の5段階



実際に成果を出している企業は、この5段階を順に登っています。段階4のRAG（社内文書に基づいて答えるAI）が、ローカルLLM活用の本命です。

導入判断の材料は、段階1（所要：数十分〜半日）でほぼ揃います。 まず小さく試すのが最短ルートです。

PoCから社内展開、システム連携まで

段階1｜体験・PoC

無料ツールの**Ollama（コマンド操作・自動化向き）** または **LM Studio（画面操作・非エンジニア向き）** をインストールし、まず小さいモデルで「自社の実データで使い物になるか」を確かめます。

所要は数十分〜半日。判断材料はこの段階でほぼ揃います。

段階2｜社内展開

Open WebUI（無料） を立てると、社員はブラウザから**ChatGPTと同じ感覚で社内AIを使えます。**

ユーザー管理・アクセス制御・利用ログも備わっており、第3章のセキュリティ設定とセットで展開します。

段階3｜システム連携

ローカルLLMは**OpenAI互換のAPI** として公開できるため、既存のAI対応ツールの接続先を差し替えるだけで、社内システム・Slack/Teamsボット・夜間の一括要約バッチに組み込めます。

利用が増えたら高スループット構成（vLLM等）へ段階的に移行。

本命はRAG ― 社内文書に基づいて答えるAI

段階4｜社内知識の活用（RAG）

RAGは、社内規程・マニュアル・過去の報告書をAIに参照させて回答させる仕組みです。

ベテランの暗黙知を、組織の資産に変える――ローカルLLM活用の本命です。

例：「出張時の宿泊費の上限は？」
→ 規程の該当箇所を根拠付きで回答

Open WebUIの文書アップロード機能なら**プログラミング不要**で試せます。

段階5｜精度の作り込み

自社の用語・文体への最適化には追加学習（ファインチューニング）という手段も。QLoRAという手法で一般的なGPUでも実行可能になりましたが、難易度は高め。**多くの用途ではRAGで十分。「必要になってから」で構いません。**

紙書類が多い会社へ

画像対応モデル（マルチモーダル） も検討価値があります。スキャンした請求書から金額・日付を抽出して整理するといった処理を、**書類をクラウドに送らず実行できます。**

現実解はハイブリッド ―「全部ローカル」にはしない

最先端のクラウドモデルにしかできない仕事（高度な推論・高品質な創作・最新情報）は確実に存在します。**実務の最適解は、データの機密度で振り分けるハイブリッド運用です。**

図表12｜業務別の振り分け早見表

業務	推奨	理由
社内FAQ・チャットボット	**ローカル**	高頻度×機密性
顧客データ・財務データの分析	**ローカル**	個人情報・機密を含む
契約書・人事文書の処理	**ローカル**	守秘義務・社内規程
マーケティング文章の生成	**クラウド**	最高品質の文章力が必要
一般的な調べもの・壁打ち	**クラウド**	機密性が低く手軽さ優先
コード生成・レビュー	**ハイブリッド**	自社コードはローカル、一般質問はクラウド

実装のヒント：Open WebUIはローカルモデルとクラウドAPIを同じ画面で切り替えられるため、ハイブリッド運用の実装は見た目ほど難しくありません。

最終判断のフレームワーク

社外に出せないデータでAIを使う必要がある	▶ ローカル導入を検討
利用量が多く（目安：月1,500万トークン以上）、高性能が必要	▶ ローカルがコスト優位
少人数・低頻度・機密データなし	▶ クラウドで十分（無理にローカル化しない）

「全部ローカルにすべきか」の答えはノー。 機密はローカル、最先端の性能はクラウド——両方のいいとこ取りが現実解です。

巻末付録

ローカルLLM導入チェックリスト20 — 4段階で自社の状況を確認する

導入判断

- ☐ 1.「社外に出せないデータを使う業務」を具体的に洗い出した
- ☐ 2. その業務の月間作業時間と人件費を把握した
- ☐ 3. クラウドAI（学習オプトアウト＋法人プラン）で足りないかを先に検討した
- ☐ 4. 月間の想定利用量（人数×頻度）を見積もった
- ☐ 5. TCO（初期投資＋減価償却＋電気代＋保守工数）で費用を試算した
- ☐ 6. 社内に運用担当（または習得意欲のある人）がいるか確認した

検証（PoC）

- ☐ 7. まず1台のPC・1つの業務・小さいモデルで試した
- ☐ 8. 自社の実データで出力品質を確認した
- ☐ 9. 生成速度が実用水準（20トークン/秒目安）か確認した
- ☐ 10. 日本語品質が要件を満たすモデルを選んだ（迷ったらQwen3系）
- ☐ 11. モデルのライセンスが商用利用可能か確認した

構築・セキュリティ

- ☐ 12. VRAM容量を基準にハードウェアを選定した
- ☐ 13. 接続を社内LANに限定し、ポートを公開していない
- ☐ 14. ユーザー認証とアクセス権限を設定した
- ☐ 15. モデルは公式配布元からのみ取得している
- ☐ 16. 対話ログの保存期間と削除ルールを決めた

運用・定着

- ☐ 17. 利用ポリシー（使ってよい業務・データの範囲）を1枚にまとめた
- ☐ 18. AI出力への人間チェックをフローに組み込んだ
- ☐ 19. 定期的なセキュリティレビュー（四半期目安）を予定に入れた
- ☐ 20. 次の段階（RAG・API連携）の候補業務リストがある

判定の目安：6個以下 → まず第4章のコスト試算と第6章の段階1から。　16個以上 → RAG・システム連携（段階3〜4）に進む準備ができています。

発行元・ご相談窓口

Harmonic Society株式会社

千葉市を拠点に、中小企業のホームページ制作・AI導入支援・IT顧問を提供しています。本書のもとになったローカルLLM解説記事群は、代表エンジニアが自ら構築・検証した実践知をまとめたものです。ヒアリングから導入・定着まで、技術の分かる相手に直接相談できます。オンライン対応・全国OK。

サービスのご案内

IT顧問ライトプラン	AI・ITの継続相談を月1.5万円から
AI導入支援	ローカルLLM含む要件整理からPoC・社内展開まで伴走
AIトレーニング（マンツーマン）	貴社の実務を教材にした生成AI研修
ホームページ制作	コーポレートサイト 198,000円（税込）〜
SEO伴走プラン	SEO記事制作・分析・改善を月3万円（税別）で伴走

まずは30分の無料相談から

「うちの場合、ローカルとクラウドどちらが合うか」「PoCを最短で回すには」——本書を読んで浮かんだ疑問を、そのままお持ちください。売り込みはしません。

ご相談・お問い合わせ

harmonic-society.co.jp/contact/

お役立ち資料一覧

harmonic-society.co.jp/whitepaper/