COLUMN | AI活用　株式会社DnD

ローカルLLMとは何か

社内情報をクラウドに出さずに
AIを活用する仕組みと始め方

2026年7月12日　株式会社DnD（dnd-inc.jp）

ローカルLLMとは何か——クラウドAIと何が違うのか

LLM（Large Language Model）とは、大量のテキストを学習した大規模言語モデルのことで、ChatGPTやClaudeといったサービスの中核をなす技術です。

通常のクラウド型AIは、ユーザーが入力した文章をインターネット越しに外部のサーバーへ送り、回答を受け取る仕組みです。これに対してローカルLLMは、モデルそのものを自社のPCやサーバーにインストールし、処理を自社環境の内部だけで完結させるものです。

> データがネットワーク外に出ない——これがローカルLLMの最大の特徴。

顧客情報・契約書・財務データ・社員の個人情報など、クラウドに送ることをためらうデータを扱う場合に、ローカルLLMは有力な選択肢になります。

なぜ今、中小企業にもローカルLLMが現実的な選択肢になったのか

3〜4年前まで、高品質なLLMを自社で動かすにはデータセンター規模のGPUが必要でした。2026年現在では、次の3つの変化によって中小企業でも導入しやすくなっています。

- オープンソースモデルの品質向上：MetaのLlama 3.3（70B）やAlibabaのQwen 2.5（72B）など、GPT-4oに近い回答品質を持つモデルが無償公開されています。
- 量子化技術の進化：モデルを圧縮する量子化（4〜8bit）により、高性能GPU搭載のワークステーション1台で動かせるサイズに収められるようになりました。
- 起動ツールの整備：Ollamaはコマンドラインからモデルをダウンロード起動できます。GUIアプリのLM Studioはエンジニア以外でも操作しやすい画面を提供しています。

ローカルLLMでできることは何か

- 文書の要約・翻訳・分類：契約書・議事録・報告書・メールを要約・分類する。情報漏洩リスクがある文書を安心して処理できる。
- **Q&A**アシスタント（**RAG**連携）：社内マニュアルや規程集を読み込ませ、社員の質問に答える社内チャットボットを構築できる。
- 文章の下書き・校正：メール文案・報告書の骨格・提案書の草稿を生成する。社内情報を含む文脈をそのまま入力できる。
- コードレビュー・補助：社内システムのソースコードをセキュリティリスクなく解析・レビューできる。

ローカルLLMを動かすには何が必要か（ハードウェア目安）

- **7〜14B**モデル（スモールスタート）：VRAM 8GB以上のGPU搭載PC。20〜40万円台から。

- **32〜70B**モデル（ビジネス品質）：VRAM 24〜48GBのGPU。ワークステーションで50〜150万円前後。

ソフトウェア（Ollama・LM Studio・Dify）はオープンソースで無料利用できます。

クラウド型とローカル**LLM**——どう使い分けるか

- 顧客名・個人情報・未公開の財務情報を含む文書 → ローカル**LLM**
- 一般的なビジネス文書の草稿・アイデア出し・最新情報の検索 → クラウド**AI**
- 社内ナレッジベースを参照した質問応答 → ローカル**LLM**＋**RAG**

中小企業が始める3ステップ

1. 1台の**PC**＋**Ollama**で試験環境を作る：担当エンジニアがOllamaをインストールし、7〜14Bモデルを起動する。数時間で試用環境が立ち上がる。
2. 1業務に絞って試す：「契約書の要約」「問い合わせメールの分類」など1業務を選び、精度と速度を確認する。
3. **RAG**を組み合わせて社内ナレッジと接続する：DifyやLlamaIndexと組み合わせ、社内マニュアルを読み込ませた質問応答システムを構築する。