表1 社内用ローカルLLMを構成する4要素

ベースモデル	実行環境	RAG（検索拡張生成）	UI（ユーザーインターフェイス）
AIの頭脳にあたる部分。現在は、Gemma、Qwen、Llamaなど、さまざまなオープンモデルが公開されている。モデルによって得意分野、対応言語、必要なメモリ容量、処理速度が異なる。パラメータ数が大きいモデルほど高性能になりやすい一方で、動作に必要なGPUのメモリ（VRAM）も増える。ローカルLLM環境では、使いたいモデルの規模と、用意できるGPUのVRAM容量のバランスが重要。	LLMをGPU上で効率よく動かすには、専用のソフトウェア基盤が必要になる。AMD環境では、GPUコンピューティング向けでオープンソースのROCm（Radeon Open Compute）を利用し、そのうえでvLLMやllama.cpp、PyTorchなどの実行環境を組み合わせる形が代表的だ。特にvLLMは、LLMをサーバーとして動かし、複数のユーザーやアプリから利用する用途で注目されている。	RAGはRetrieval-Augmented Generationの略で、日本語では検索拡張生成と呼ばれる。LLMはそれ自体が大量の知識を持っているが、社内の最新資料や独自文書の内容を最初から知っているわけではない。そこで、社内文書をデータベース化しておき、質問に関連する文書を検索し、その内容をLLMに渡して回答を生成させる。	実行環境を構築しただけでは、一般の社員が使うにはハードルが高い。そこで、Webブラウザからチャット形式で利用できるUIを用意する。代表的な選択肢の1つがOpen WebUIで、社内LAN上のPCからWebブラウザでアクセスし、ChatGPTのような感覚でローカルLLMを利用できる。ユーザーごとの設定と、管理者による全体設定を分けて管理することが可能。

も、基本的にはこのLLMをクラウド上で動かし、ユーザーがブラウザやアプリから利用している。

ローカルLLMは、そのLLMを自社のPCやサーバー上で動かす仕組みだ。オンラインサービスとの一番大きな違いは、処理がどこで行なわれるかにある。オンラインサービスでは、ユーザーが入力した文章がインターネット経由でサーバーへ送られ、そこで処理された結果が返ってくる。これに対してローカルLLMでは、モデルの実行環境をPCに置き、入力した内容も同じPC上で処理をする。

そのため、ローカルLLMでは導入や運用に一定の知識が必要だ。モデルを選び、GPUなどのハードウェアを用意し、実行環境を構築し、社内の利用者が使いやすいUIを整える必要がある。ただし、最近はオープンなLLMや周辺ツールが急速に整備されており、以前に比べると導入のハードルはかなり下がっている。GPUを搭載したPCやワークステーションを用意すれば、中小企業でも現実的なコストで社内AI環境を構築できるようになってきた。

今回のような社内用ローカルLLMを構成する場合の要素は、大きく分けて4つある。LLMのベースモデル、そのモデルを動かす実行環境、社内データを活用するためのRAG（検索拡張生成）、そして利用者がアクセスするためのUIまたはサービスだ（**表1**参照）。

ローカルLLMの仕組みを整理すると、まずGemmaやQwenなどのベースモデルを選び、それをROCmやvLLMといった実行環境でGPU上に展開する。さらに、社内文書をRAGで参照できるようにし、Open WebUIのようなWebブラウザUIを通じて利用者に提供する、という流れになる。

この中でハードウェア、特にGPUのVRAM容量は重要なポイントだ。モデルを動かすための領域だけでなく、入力文や生成中のデータ、複数ユーザーの同時利用に必要な領域もVRAMを消費する。VRAMが少ないと、選べるモデルの規模が小さくなったり、同時利用時の余裕が少なくなったりする。

低予算でローカルLLMを作れるAMDプラットフォーム

ローカルLLM環境を構築する際、重要になるのが大容量VRAMを搭載したグラフィックスカードだ。

VRAM（Video RAM）とは、グラフィックスカードに搭載された専用メモリのこと。LLMの動作にはモデルの重み（パラメータ）をVRAMに展開する必要があり、容量が不足するとモデルを動かせない。一般的なゲーミング向けグラフィックスカードに搭載される8GBや12GBのVRAMでは、実用的な30B（約300億パラメータ）クラスのモデルは動かせず、法人運用するなら24GB以上のVRAMが最低ラインとなる。

VRAM 32GBが魅力の「AMD Radeon™ AI PRO R9700」

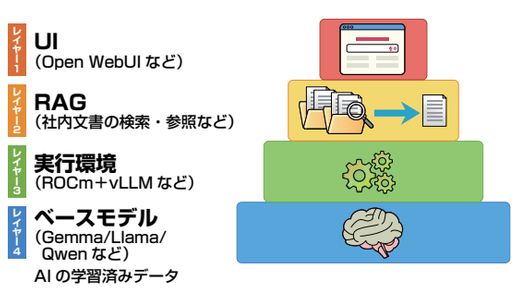
今回は中小企業が、社内にローカルLLMサーバーを導入するというシチュエーションでグラフィックスカードを選びたい。そこでおすすめしたいのが、AMDのGPU「Radeon AI PRO R9700」を搭載したグラフィックスカードだ。

Radeon AI PRO R9700は、RDNA 4世代のアーキテクチャを採用し、ローカルでのAI推論・開発といったワークロードに最適化されたGPUだ。Radeon AI PRO R9700の魅力の1つはVRAMを32GBも搭載していること。前述した30Bクラスのモデルを扱えるのは大きく、しかもRadeon AI PRO R9700搭載グラフィックスカードの実売価格は25万円前後と、AI向けとしては破格の安さなのである。

なお、Radeon AI PRO R9700と演算ユニット数などが同じものとしてコンシューマ向けの「AMD Radeon™ RX 9070 XT」もあるが、こちらはVRAMが16GBと少なくなるため、30Bクラスのモデルを動かすのが難しい。

実際、Radeon RX 9070 XT（VRAM 16GB）のマシンで、26Bのパラメータの LLMを起動しようとすると、VRAMの容量不足で起動できないというメッセージが返ってくる。パラ

ローカルLLMの主な階層



LLMは大まかに学習済みデータを「ベースモデル」とし、それを動かす「実行環境」、AIの回答精度を向上させる「RAG」（検索拡張生成）、そしてWebブラウザなどからアクセスするための「UI」などから成っている