社内向けローカルLLMサーバー構築、意外と難しくない

コマンドがコピーできる⇒
Web版も公開中

UbuntuマシンにvLLMを導入する

個人利用や家庭内LANでLLMを動かす場合、llama.cppやLM Studio（llama.cppをGUIで操作できるツール）が一般的な選択肢だ。社内向けサーバーにもこれらを使うことはできるが、複数人が同時にアクセスする並列処理を考慮すると、現時点では「vLLM」が適している。

vLLMとは、カリフォルニア大学バークレー校のSky Computingラボが開発したオープンソースのLLM推論エンジンで、本番環境向けの高スループット推論に特化している。

vLLMの最大の特徴は、KVキャッシュをOSのページングメモリになぞらえて動的管理する独自技術「PagedAttention」を搭載していること。KVキャッシュとはLLMが過去のトークン（入力・出力履歴）を処理する際に生成する中間データを保存しておく領域のことだ。会話の文脈を保持するために必要で、コンテキスト長（会話の長さ）に比例してVRAMを消費する。vLLMの動的KVキャッシュは、従来の静的KVキャッシュと比べてVRAM利用効率が大幅に向上し、長いコンテキストや大きなバッチの並列処理で真価を発揮するのである。

vLLMとllama.cppを用途別に整理すると**表4**のようになる。llama.cppはローカル/シングルユーザー用途では非常に優秀だが、オフィスのオンプレミス環境や複数人が同時利用する用途、Claude Codeなどのツールとの連携を想定するならvLLMが適している。そのため、ここではvLLMを使ってローカルLLMサーバーを構築する。

ステップ1｜AMDドライバ+ROCmのインストール

それでは、AMD Radeon™ AI PRO R9700とAMD Ryzen™ 7 9850X3Dを搭載するPCが用意されていることを前提に話を進めていこう。

まずローカルLLMサーバーを作るにあたり、AMDドライバとROCmのインストールを行なう。なお、今回はOSに「Ubuntu 24.04.4 LTS」を使用した。サーバー OSとして広く普及しており、安定した動作が期待できる。

表4 vLLMとllama.cppの違い

	vLLM	llama.cpp
主な用途	サーバー/API/複数同時リクエスト	ローカル/シングルユーザー
対応ハード	ROCm/CUDA	CPU/GPU/Apple Silicon等
並列処理	本格的なバッチ処理	限定的
対応フォーマット	safetensors/AWQ/FP8等	主にGGUF
OpenAI互換API	標準搭載	llama-server
メモリ効率	やや重い	軽い
速度（シングル）	普通	速い場合もある
速度（並列）	強い	弱い
セットアップ	やや複雑	非常に簡単

検証に使用したPC
今回は、CPUにRyzen 7 9850X3D、GPUにRadeon AI PRO R9700を搭載したPCをベースに、ローカルLLMサーバーを構築・検証した

ドライバはAMD公式サイトのセットアップ手順（https://www.amd.com/ja/support/download/linux-drivers.html）に従ってインストールする。ここで使用したドライバは「Radeon Software for Linux version 25.35.1 for Ubuntu 24.04.4 HWE with ROCm 7.2.1」だ。

```
sudo apt update
wget https://repo.radeon.com/amdgpu-install/7.2.1/ubuntu/noble/amdgpu-install_7.2.1.70201-1_all.deb
sudo apt install ./amdgpu-install_7.2.1.70201-1_all.deb
sudo amdgpu-install -y --usecase=graphics,rocm
sudo usermod -a -G render,video $LOGNAME
sudo reboot
```

再起動後、ROCkカーネルモジュールが正しくロードされているかを確認する。

```
rocminfo | grep "ROCk module"
```

以下のように表示されれば正常にインストールされている。

```
ROCk module version 6.16.13 is loaded
```

ステップ2｜uvとHugging Faceのセットアップ

Pythonのパッケージ管理ツール「uv」と、機械学習モデルの共有プラットフォーム「Hugging Face」をセットアップする。uvなら「pip」と同様に使えて高速で、仮想環境の