第1章 多くのLLMモデルに対応! 基本のチャット機能からRAG，画像認識&生成も

ローカルLLM用新定番UI Open WebUIを使ってみる

氏森 充

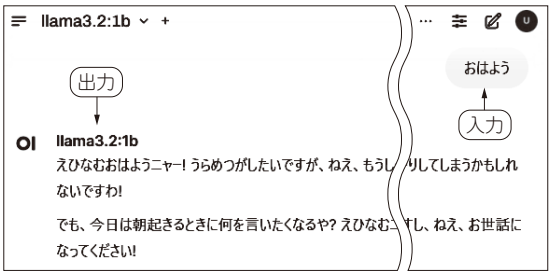

図1 Open WebUIを使ったチャットの例

表1 Open WebUIとOllamaの関係

ツール名/項目	Open WebUI	Ollama
役 割	LLMの実行環境を提供するWebインターフェース	LLMの実行エンジン
機 能	LLMとの対話，モデルの管理，プラグインによる機能拡張	LLMの推論実行，モデルのロード/保存
関係性	Ollamaをバックエンドとして利用することができる	Open WebUIはOllamaに依存して動作する
主な用途	LLMを気軽に試したい，LLMを使ったアプリケーションを開発したい	LLMの推論処理を高速化したい，カスタムLLMを開発したい

LLM（大規模言語モデル）は，OpenAI社が2022年11月にChatGPTを公開したことをきっかけに世界中で注目を集め，急速に普及しました．その利便性の高さから，設計開発現場への導入も検討され始めますが，インターネット接続を前提とするクラウド・サービスでは情報漏えいのリスクが高く，現実的ではありませんでした．

しかし，メタ社が2023年7月に公開したオープンソースLLM Llama 2の登場により，状況が一変します．これにより，ローカル環境のみでLLMを運用することが可能になり，情報漏えいの心配をせずに利用できるようになりました．

本稿では，このようなローカル環境でLLMを動かすときに便利なWebインターフェース「Open WebUI」と，LLM実行エンジン「Ollama」を紹介します．（編集部）

Open WebUIはクラウド・サービスではなく，自分のPCで複数のLLMを動かすことができます．これにより，プライバシが気になるデータや，機密性の高い情報も安心して扱うことができます．オープンソースであり，世界中に開発者が居ます．Ollamaという LLM実行環境と連携することで，より手軽にLLMを利用することができます．Ollamaもオープンソースであり，Llama（Meta提供）やMistralなど，人気のLLMをサポートしています．関係を**表1**に整理します．

Open WebUI登場以前は，TensorFlowやPyTorchといった深層学習フレームワークを直接利用し，LLMのモデルをゼロから構築するか，既存のモデルを修正する必要がありました．高度な知識と大規模な計算資源を必要とするため，一般ユーザにとってはハードルが高いものでした．また，Dockerを用いてLLMの実行環境をコンテナ化し，再現性の高い環境を構築する手法も存在しました．しかし，Dockerで提供された環境をカスタマイズするには，やはり高度が知識が求められます．

これにより，**図1**のようなユニークなチャット（会話）が可能になります．

なお，今回使用するソフトウェアのバージョンは，次の通りです．

- Open WebUIバージョン：v0.4.7
- Ollamaバージョン：0.4.3

インストール手順については Appendix 2を参照してください．

基本操作：Open WebUIを使ったチャット

Open WebUIを使った基本的なチャット操作を試してみます．最初に，使用するモデルを選択します．Appendix 2の通りにOpen WebUIとOllamaをインストールすると，左上にArena Modelかllama3.2:1bが表示されます．

Arena Modelは，応答を比較するために設計されたモデル・コレクションの総称です．複数のモデルが