ローカルLLMの導入と活用の試み

水矢　亨（情報・生産技術部）

ローカルLLM，エージェント，MCP

はじめに

生成AI(LLM：大規模言語モデル)の利用の動向

・対話(チャット)から**エージェント**へ

　― MCP や Skill による外部ツール／外部データ等の利用

　― 「開発して使う」(LangCain など)から「設定して使う」(OpenClaw など)へ

　― 自律的な Agent Loop でトークン量が増加

　― 機密保持しながらユーザデータを活用

→ **ローカルLLM**

・フィジカルAI（現実世界と相互作用するAI）

ローカルLLMの実行用ソフトウェアとモデル

・無償ソフト(Ollama，LM Studio など)あり

・tool calling 対応で3B～8B程度のローカルLLM

　qwen3.5，granite4.1 gemma4，llama3.2 など

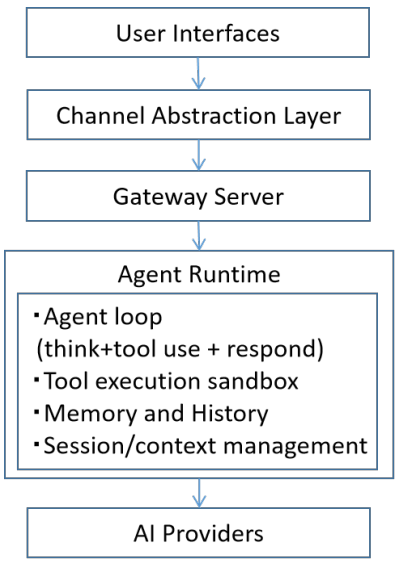
エージェント実行環境

・ 生成AIを使うエージェントをUI(ファイル操作含む)とを連携させて動作する

・ OpenClaw や Hermes など

　(NanoBot，ZeroClaw，PicoClaw などの派生もある)

・FastMCP などで作成したMCPサーバを設定する



MoltBot(OpenClawの前身）のアーキテクチャ

https://github.com/az9713/moltbot/ のdvanced-tutorials の図を参考に作成

軽量エージェント実行環境による事例

・OpenClaw 派生とされる ZeroClaw を利用

　(Rust製で軽量・マルチチャンネル対応)

・ローカル LLM：Ollama 経由で Granite4.1(3B)

・画像ファイルから文字情報を抽出するOCR機能を提供するMCPサーバを用意（FastMCPを利用）

・エージェントは，ワークスペース内の画像ファイルを探し、MCPサーバが提供するOCR機能を使って、そのファイルから文字情報を抽出する

→ PC(Core i7/メモリ32GB/Intel UHD Graphics)でも動作させることができた(右図)

地方独立行政法人
KISTEC 神奈川県立産業技術総合研究所

問い合わせ先
情報・生産技術部
TEL 046-236-1500