|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tiêu Chí | Ollama agent | Jan.ia | LM Studio | marimo notebook | jupyter notebook |
| 1. Tính năng, Chức năng, Khả năng tùy biến prompt, mở rộng điều chỉnh tham số chuyển thành Form fields | **Chức năng chính:** Serve LLMs cục bộ với API. **Tùy biến:** Tùy biến prompt thông qua Modelfile (system prompt, template). Tham số qua API. UI (như Open WebUI) thêm form. | **Chức năng chính:** Ứng dụng desktop chạy LLM. **Tùy biến:** UI thân thiện, chatbox, prompt templates, điều chỉnh tham số cơ bản (nhiệt độ, top\_k). | **Chức năng chính:** Ứng dụng desktop tải, chạy LLM. **Tùy biến:** UI trực quan, chatbox, nhiều tùy chọn prompt, điều chỉnh tham số chi tiết qua sliders/fields. | **Chức năng chính:** Notebook reactive, xây dựng app web. **Tùy biến:** Hoàn toàn tùy biến bằng Python, có thể tạo form input động cho prompt/params. | **Chức năng chính:** Môi trường code tương tác. **Tùy biến:** Hoàn toàn tùy biến bằng Python, có thể dùng ipywidgets tạo form cho prompt/params. |
| 2. Tải models từ hãng: huggingface, kaggle, openai, ollama, anthorpic, qwen, deepseek | **Ollama Hub:** Tải trực tiếp các model đã chuyển đổi sang định dạng Ollama (nhiều model từ HF). **Không trực tiếp:** từ HF, Kaggle, OpenAI, Anthropic (API). | **Hugging Face:** Tải GGUF trực tiếp. **Local Files:** Chạy model từ file cục bộ. **Không trực tiếp:** từ Ollama Hub, OpenAI, Anthropic. | **Hugging Face:** Tải GGUF trực tiếp. **Local Files:** Chạy model từ file cục bộ. **Không trực tiếp:** từ Ollama Hub, OpenAI, Anthropic. | **Qua Code:** Có thể tải model từ HF (thư viện transformers), Kaggle (API), OpenAI/Anthropic/Ollama (qua API). | **Qua Code:** Tải model từ HF (thư viện transformers), Kaggle (API), OpenAI/Anthropic/Ollama (qua API). |
| 3. GGLF, llm, qwen, dataset | **GGUF:** *Sử dụng* GGUF (đã được đóng gói trong định dạng Ollama). Không phải công cụ chuyển đổi. | **GGUF:** *Sử dụng* GGUF. Không phải công cụ chuyển đổi. | **GGUF:** *Sử dụng* GGUF. Không phải công cụ chuyển đổi. | **Dataset:** Có thể xử lý, phân tích dataset. **Transform:** Có thể viết code Python để chuyển đổi định dạng model (ví dụ: PyTorch -> GGUF). | **Dataset:** Có thể xử lý, phân tích dataset. **Transform:** Có thể viết code Python để chuyển đổi định dạng model (ví dụ: PyTorch -> GGUF). |
| 4. Web rest api local, Tích hợp webui các loại AI Agent chạy localhost (Offline) | **Web REST API Local:** CÓ (mặc định có API). **Tích hợp WebUI:** CÓ (thường dùng với Open WebUI). **Offline:** CÓ. | **Web REST API Local:** KHÔNG (là ứng dụng desktop). **Tích hợp WebUI:** KHÔNG. **Offline:** CÓ. | **Web REST API Local:** CÓ (có chế độ OpenAI-compatible API server). **Tích hợp WebUI:** CÓ (UI nội bộ hoặc với các tool hỗ trợ OpenAI API). **Offline:** CÓ. | **Web REST API Local:** CÓ (có thể viết code để expose API). **Tích hợp WebUI:** CÓ (có thể tự xây dựng UI). **Offline:** CÓ. | **Web REST API Local:** CÓ (có thể viết code để expose API). **Tích hợp WebUI:** CÓ (có thể tự xây dựng UI). **Offline:** CÓ. |
| 5. Chạy local trên win 11 pro, win 2019 dc server, linux ubuntu 24.04 | **Windows 11 Pro:** CÓ. **Windows 2019 DC Server:** CÓ. **Linux Ubuntu 24.04:** CÓ. | **Windows 11 Pro:** CÓ. **Windows 2019 DC Server:** CÓ (dưới dạng ứng dụng desktop). **Linux Ubuntu 24.04:** CÓ. | **Windows 11 Pro:** CÓ. **Windows 2019 DC Server:** CÓ (dưới dạng ứng dụng desktop). **Linux Ubuntu 24.04:** CÓ. | **Windows 11 Pro:** CÓ. **Windows 2019 DC Server:** CÓ. **Linux Ubuntu 24.04:** CÓ (yêu cầu Python). | **Windows 11 Pro:** CÓ. **Windows 2019 DC Server:** CÓ. **Linux Ubuntu 24.04:** CÓ (yêu cầu Python). |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Open WebUI | Vicuna | Llama-2-7B-Chat-HF | Text Generation WebUI Oobabooga | FastChat |
| Chức năng chính: Web UI cho Ollama/LocalAI. Tùy biến: UI chat hiện đại, prompt templates, system prompt, điều chỉnh tham số qua form/slider. | **Bản thân là model:** Khả năng do model định nghĩa. Tương tác qua công cụ khác (Oobabooga, FastChat). | **Bản thân là model:** Khả năng do model định nghĩa. Tương tác qua công cụ khác (Ollama, LM Studio). | **Chức năng chính:** Web UI toàn diện cho text generation. **Tùy biến:** Rất mạnh mẽ, nhiều chế độ chat, prompt templates, tùy chỉnh hàng loạt tham số (nhiệt độ, top\_p, top\_k, repetition penalty, Mirostat, v.v.), hỗ trợ extensions. | **Chức năng chính:** Framework huấn luyện/phục vụ LLM. **Tùy biến:** Web UI chat đơn giản, prompt cơ bản, vài tham số thông qua UI/API. |
| Không trực tiếp: Là UI. Tải model qua các service nó kết nối (Ollama, LocalAI, OpenAI API). | **Hugging Face:** Là nơi chứa model này (thường là PyTorch, Safetensors). Cần công cụ để tải/chạy. | **Hugging Face:** Là nơi chứa model này (thường là PyTorch, Safetensors). Cần công cụ để tải/chạy. | **Hugging Face:** Tải trực tiếp nhiều định dạng (GGUF, Safetensors, PyTorch). **Local Files:** Chạy từ file cục bộ. | **Hugging Face:** Tải các model tương thích. **Không trực tiếp:** Ollama, OpenAI, Anthropic. |
| Transform: Không, nó là UI. | **Bản thân là model:** Tồn tại ở định dạng PyTorch/Safetensors trên HF. Cần công cụ khác để chuyển sang GGUF. | **Bản thân là model:** Tồn tại ở định dạng PyTorch/Safetensors trên HF. Cần công cụ khác để chuyển sang GGUF. | **GGUF/khác:** *Hỗ trợ tải và chạy* nhiều định dạng (GGUF, PyTorch, Safetensors). Có thể có script chuyển đổi đi kèm. | **LLM formats:** Hỗ trợ các định dạng model của Hugging Face (PyTorch, Safetensors). Không chuyển đổi. |
| Web REST API Local: KHÔNG (chỉ là UI). Tích hợp WebUI: CÓ (chính nó là WebUI). Offline: CÓ (khi kết nối với Ollama/LocalAI offline). | **Bản thân là model:** KHÔNG. Cần một công cụ để phục vụ API/UI. | **Bản thân là model:** KHÔNG. Cần một công cụ để phục vụ API/UI. | **Web REST API Local:** CÓ (có chế độ OpenAI-compatible API server). **Tích hợp WebUI:** CÓ (chính nó là WebUI). **Offline:** CÓ. | **Web REST API Local:** CÓ (có API server). **Tích hợp WebUI:** CÓ (có WebUI mẫu). **Offline:** CÓ. |
| Windows 11 Pro: CÓ (Docker/Python). Windows 2019 DC Server: CÓ (Docker/Python). Linux Ubuntu 24.04: CÓ (Docker/Python). | **Bản thân là model:** CÓ (chạy trên bất kỳ OS nào thông qua một framework phục vụ). | **Bản thân là model:** CÓ (chạy trên bất kỳ OS nào thông qua một framework phục vụ). | **Windows 11 Pro:** CÓ (Python). **Windows 2019 DC Server:** CÓ (Python). **Linux Ubuntu 24.04:** CÓ (Python). | **Windows 11 Pro:** CÓ (Python). **Windows 2019 DC Server:** CÓ (Python). **Linux Ubuntu 24.04:** CÓ (Python). |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| litellm | anythingllm | llama.cpp | PrivateGPT | LocalAI |
| Chức năng chính: Proxy API thống nhất cho nhiều LLM. Tùy biến: Ánh xạ tham số giữa các API, cho phép tùy biến prompt theo chuẩn API gốc. Không có UI. | **Chức năng chính:** Nền tảng RAG, Chatbot. **Tùy biến:** Web UI cho chat, prompt templates, quản lý tài liệu, có thể chỉnh system prompt. | **Chức năng chính:** Core engine inference GGUF. **Tùy biến:** CLI, prompt trực tiếp qua dòng lệnh, tham số cơ bản qua flag. Server mode có API. | **Chức năng chính:** RAG riêng tư. **Tùy biến:** Web UI cho chat, tải tài liệu. Prompt thông qua API/UI, chỉnh system prompt. | **Chức năng chính:** Serve LLMs (OpenAI API-compatible). **Tùy biến:** API-centric, tùy biến prompt và tham số thông qua request body JSON. |
| Proxy: Không tải, mà làm proxy cho API của OpenAI, Anthropic, Cohere, Azure, Ollama, Hugging Face Inference API. | **Tích hợp API/Local:** OpenAI, Anthropic, Hugging Face (API), Ollama, LocalAI. **Không tải trực tiếp:** từ Kaggle. | **Không tải:** Nó là engine chạy GGUF. Cần tải file GGUF từ nguồn khác (ví dụ: HF). | **Tích hợp API/Local:** OpenAI, Anthropic (API), Hugging Face (local/API), Ollama, LocalAI. | **Tải/Chạy:** Hỗ trợ tải và chạy các model từ Hugging Face (GGUF, Safetensors, Pytorch), Ollama. |
| Transform: Không, nó xử lý API requests/responses. Không liên quan đến định dạng model file hay dataset. | **Dataset:** Xử lý dataset cho RAG (tải tài liệu, tạo embeddings). Không transform model. | **GGUF:** *Là* tiêu chuẩn để chạy GGUF. Có script (convert.py) để chuyển đổi model PyTorch sang GGUF. | **Dataset:** Xử lý dataset cho RAG (tải tài liệu, tạo embeddings). Không transform model. | **GGUF/khác:** *Hỗ trợ chạy* nhiều định dạng (GGUF, PyTorch, ONNX, TF). Không phải công cụ chuyển đổi chính. |
| Web REST API Local: CÓ (chính nó là một proxy API server). Tích hợp WebUI: CÓ (nếu webUI kết nối tới API của nó). Offline: CÓ (nếu backend LLM cũng offline). | **Web REST API Local:** CÓ (có API internal và external). **Tích hợp WebUI:** CÓ (có sẵn WebUI mạnh mẽ). **Offline:** CÓ. | **Web REST API Local:** CÓ (với llama.cpp/server). **Tích hợp WebUI:** CÓ (với các UI hỗ trợ OpenAI API hoặc llama.cpp API). **Offline:** CÓ. | **Web REST API Local:** CÓ (có API internal). **Tích hợp WebUI:** CÓ (có sẵn WebUI). **Offline:** CÓ. | **Web REST API Local:** CÓ (chính nó là một OpenAI-compatible API server). **Tích hợp WebUI:** CÓ (với Open WebUI, anythingLLM, v.v.). **Offline:** CÓ. |
| Windows 11 Pro: CÓ. Windows 2019 DC Server: CÓ. Linux Ubuntu 24.04: CÓ (yêu cầu Python). | **Windows 11 Pro:** CÓ (Docker/Native). **Windows 2019 DC Server:** CÓ (Docker/Native). **Linux Ubuntu 24.04:** CÓ (Docker/Native). | **Windows 11 Pro:** CÓ. **Windows 2019 DC Server:** CÓ. **Linux Ubuntu 24.04:** CÓ (biên dịch). | **Windows 11 Pro:** CÓ (Docker/Python). **Windows 2019 DC Server:** CÓ (Docker/Python). **Linux Ubuntu 24.04:** CÓ (Docker/Python). | **Windows 11 Pro:** CÓ (Docker/Native). **Windows 2019 DC Server:** CÓ (Docker/Native). **Linux Ubuntu 24.04:** CÓ (Docker/Native). |