Flowise-Intro: Chatbot (Service Berlin)

1. Flowise installieren. Dafür ins Terminal gehen und folgende Anweisung: „npm install -g flowise“ und danach darüber starten „npx flowise start“
2. OLLAMA installieren (LLM Framework für den Chat bot, da GPT nicht geht :D )
   1. Erst auf der Webseite Ollama installieren:
   2. Im Terminal lokal starten über: ollama run llama2
   3. Dann „ollama serve“ im Terminal
   4. Jetzt sollte OLLAMA gestartet wurden sein auf: <http://localhost:11434/>
3. Dann die Seite: <http://localhost:3000/> öffnen wo du ins Flowise Dashboard kommst
4. Dort gehe auf Add New -> Einstellung (Zahnrad) -> Load Chatflow. Dann die JSON Dateien einfügen
5. LLAMA\_Chat.json
   1. Hier wird einmal kurz demonstriert wie llama2 kommuniziert. Hierfür musst du bei ChatOllama in die Base URL: <http://localhost:11434/> und als Model Name: llama2
   2. Dann kannst du einen Chat starten und iwas schreiben
6. LLama2\_PDF\_Wohnung.json
   1. Fast derselbe Aufbau nur das wir eine memory vector eigenschaft hinzufügen und einen PDF-File-Reader. In der File-Reader musst du einmal die pdf: [info\_anmeldung\_einer\_wohnung.pdf](blob:chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/431d5323-cac3-437d-aadb-e65df4e5007d) hinzufügen und im OLLama-Chat bzw. Ollama-Embedding wieder den localhost und den Model Name: llama2
   2. Dann auf das grüne Datenbank Icon. Um die Daten zu vektorisieren
   3. Sobald das erfolgreich war eine spezifische Frage übers Dokument stellen
7. Rag\_Berlin\_Service.json
   1. Ist dasselbe wie LLama2\_PDF\_Wohnung.json nur mir einer Vectordatabase verknüpft (Pinecone). Da einmal die Credits eintragen: Name: pinecone; Key: 5afb36c9-cf50-47aa-a39b-335de452afdb und als Pinecone-index: test oder pdf benutzen
   2. Den Rest wie oben mit dem File-Reader und Ollama
   3. Und dann Upserten und den Chat starten
8. GPT\_AS
   1. Das wäre die Version über chatgpt, die leider nicht funktioniert hat, aufgrund der Limit Überschreitung.
   2. Ähnlicher Flow wie bei LLama nur das der Inhalt von Wohnungsanmeldung gescraped wird und GPT nutzen würden. Wenn die Überschreitung aufgehoben ist, vlt mal die Alternative testen, da dies auch im FLowise Tutorial vom City Lab fast genauso gemacht wurde und die Ergebnisse besser als in Llama ausschauen

Sollte die Anfragen bei dir besser und schneller verarbeitet werden, wäre das sehr gut und wir können sogar noch finetunen