

Design bots using Flowise

Last amended: 19th Jan, 2025

My folder: C:\Users\ashok\OneDrive\Documents\flowise

Flowise Book created by Community [at this link](#)

Table of Contents

A. Simple demo.....	2
B. Translation bot:.....	2
C. Chat with llama:.....	3
D. Simple Conversational Chain	4
E. Using Conversational Agents.....	4
F. Export import chat flows:	5
i) Export chatflow	5
ii) Load chat flow.....	5
G. Simple RAG with single text file.....	6
H. RAG with chroma store and single text file	7
I. Combining Multiple Chains (Prompt Chaining)	8
First LLM chain.....	8
Second LLM chain added to first.....	9
J. Flowise Using Hugging Face Models	10

A. Simple demo

See [this link](#) on YouTube.

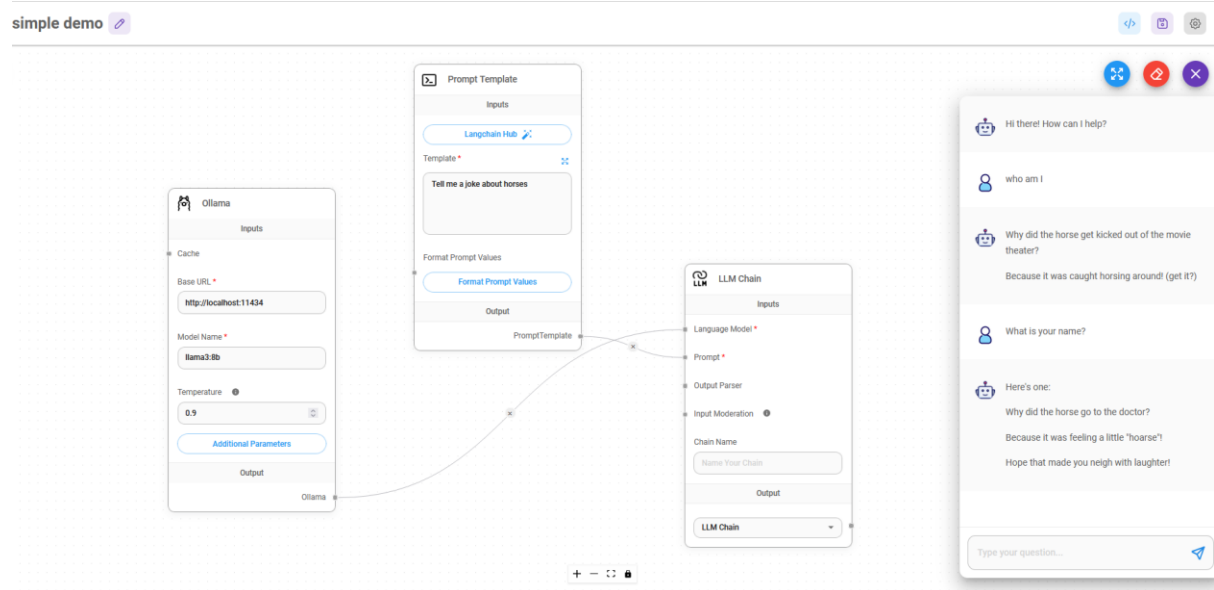


Figure 1: This bot will always answer your questions as a horse's joke. The only prompt is: **Tell me a joke about horses.**

B. Translation bot:

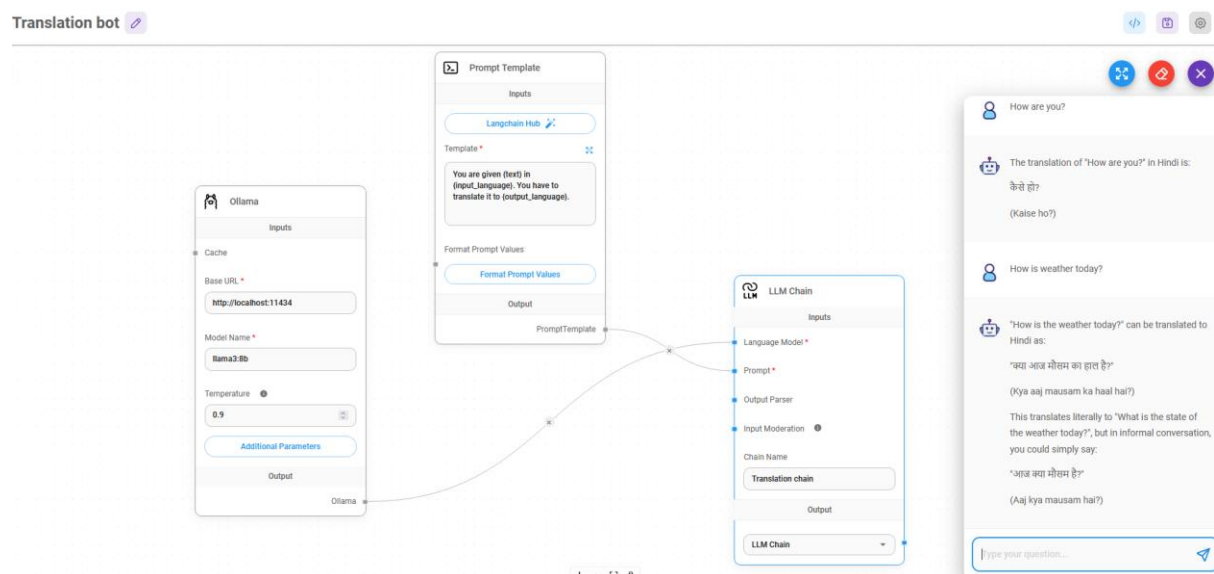


Figure 2: This bot translates all your questions into the desired language.

Prompt Template is:

You are given {text} in {input_language}. You have to translate it to {output_language}.

And formatted template is as follows. Note the **text** pertains to user's question asked in the chat-bot.

Format Prompt Values

```
{ 3 items
  text : "{{question}}"
  input_language : "English"
  output_language : "Hindi"
}
```

Figure 3: Translate question asked in English to Hindi: Note that 'question' is enclosed in two curly brackets.

C. Chat with Llama:

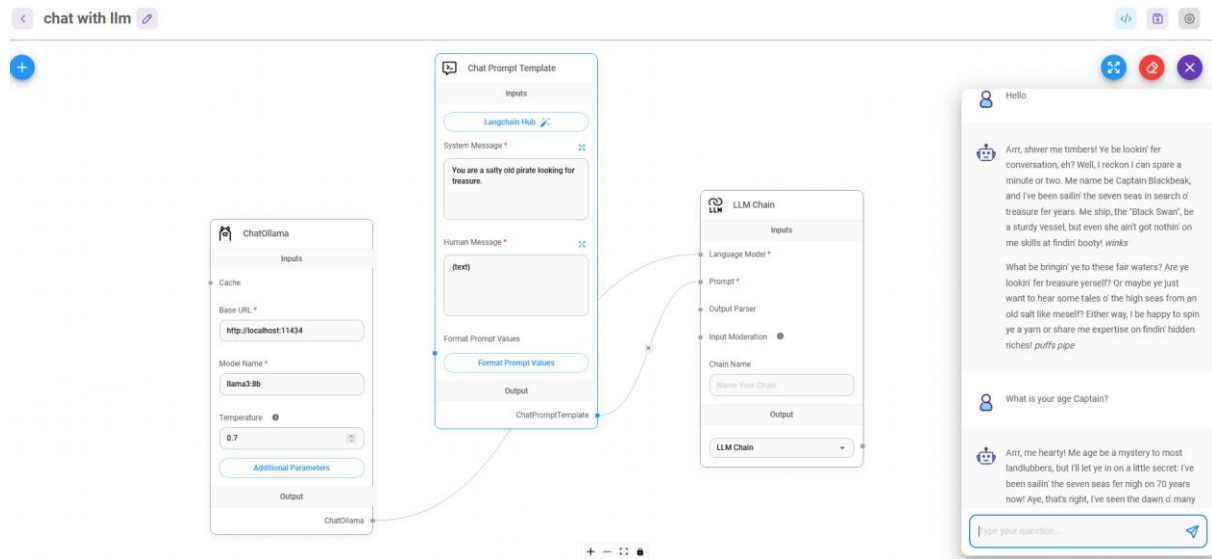
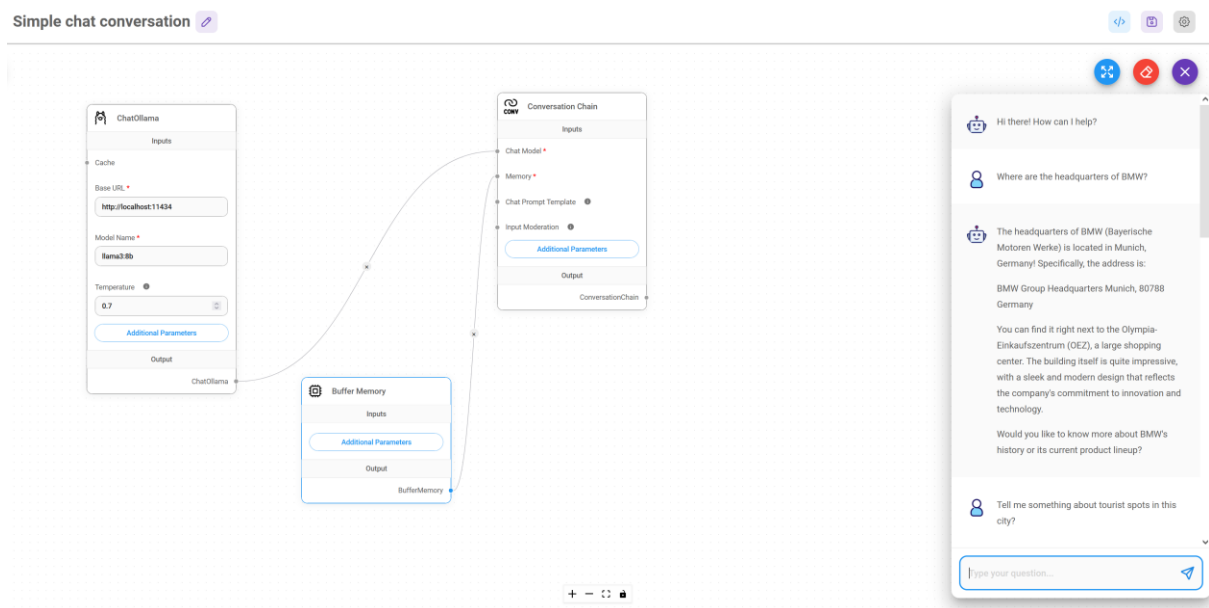


Figure 4: The 1st question is just **Hello** but the 11nd question asks more details about **Captain** referred to into the answer to **Hello**.

D. Simple Conversational Chain

Refer [YouTube video](#)



E. Using Conversational Agents

Refer YouTube video

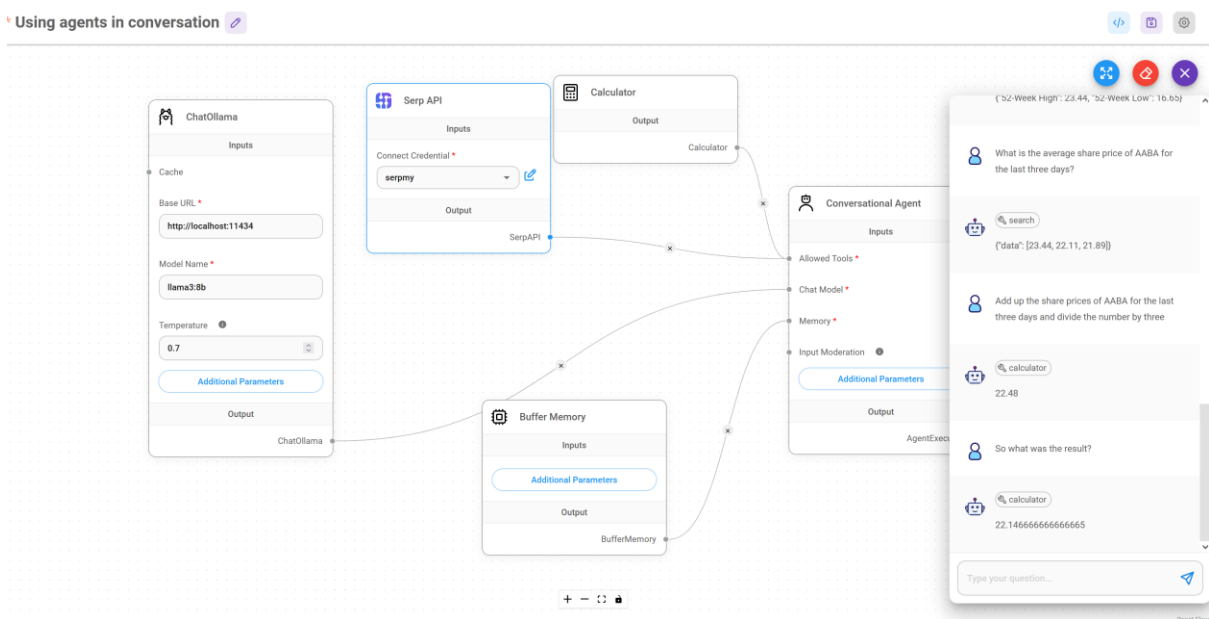


Figure 5: Agents can work even with ollama. SERP API key is a must. To calculate average, we have to tell the bot how to do it.










F. Export import chat flows:

i) Export chatflow

Chatflows

Q Search Name or Category

+ Add New




Name	Category	Nodes	Last Modified Date ↓	Actions
chat with llm		  	July 22nd, 2024	Options ▾
Translation bot		  	July 22nd, 2024	
simple demo		  	July 22nd, 2024	

- Rename
- Duplicate
- Export
- Starter Prompts
- Chat Feedback
- Allowed Domains
- Speech To Text
- Update Category
- Delete

Figure 6: In the Chat flows window, click on the down arrow besides the **Options** to Export a chat flow as a json file.

ii) Load chat flow

To load a json file, first create a new (blank) chatflow by any name, say 'abc'. Save the blank chatflow. Click on Settings icon on top-right. And then click on Load chatflow to open and load the json file.

- View Messages
- View Leads
- Upsert History
- Configuration
- Duplicate Chatflow
- Load Chatflow**
- Export Chatflow
- Delete Chatflow

Figure 7: Click on Settings icon to import an exported chat flow (i.e. json file).

The screenshot displays the LangChain Studio interface with a workflow for generating jokes. The workflow consists of three main components:

- Ollama Component:**
 - Inputs:** Cache, Base URL (http://localhost:11434), Model Name (llama3.8b), Temperature (0.9).
 - Output:** Ollama.
- Prompt Template Component:**
 - Inputs:** Langchain Hub.
 - Template:** Tell me a joke about horses.
 - Format Prompt Values:** Format Prompt Values.
 - Output:** PromptTemplate.
- LLM Chain Component:**
 - Inputs:** Language Model (Ollama), Prompt (PromptTemplate).
 - Output Parser:** Output Parser.
 - Input Moderation:** Input Moderation.
 - Chain Name:** Name Your Chain.
 - Output:** LLM Chain.

The workflow is connected by arrows, showing the flow from Ollama to the LLM Chain, and from the Prompt Template to the LLM Chain. On the right side, there is a chat interface with a list of questions and a text input field at the bottom.

G.Simple RAG with single text file

Refer [Flowise tutorial #3](#)

The screenshot displays the LlamaIndex Studio interface for a 'Simple RAG with single text file' workflow. The workflow is composed of several interconnected components:

- Recursive Character Text Splitter:** Takes a 'Text File' as input and splits it into chunks based on 'Chunk Size' (200) and 'Chunk Overlap' (20).
- Text File:** A component that provides the input document for the text splitter.
- In-Memory Vector Store:** Stores the embeddings generated by the text splitter. It includes a 'Memory Retriever' for fetching relevant information.
- HuggingFace Inference Embeddings:** A component that generates embeddings for the input text using the HuggingFace API.
- ChatGPT:** The central AI model that processes the retrieved information and generates responses.
- Buffer Memory:** A component that maintains the context of the conversation across multiple turns.
- Conversational Retrieval QA Chain:** The final component that combines the retrieved information, chat history, and the AI model to provide a coherent response.

On the right side of the interface, there is a chat window titled 'Hi there! How can I help?'. It shows a conversation history with two messages:

- Message 1: "How things changed with the coming of microcomputers"
- Message 2: "As technology evolved and became more accessible, people began to seek out ways to interact with computers in a way that was more personal and intuitive. Microcomputer systems, which were introduced in the 1970s, marked a major breakthrough in this regard. Instead of relying on standard keyboards and mice for input, these devices featured interactive screens that could display text, images, and other information. The early microcomputers offered a great deal of flexibility to users who wanted to customize their interfaces. They were able to choose from a variety of graphical user interfaces (GUIs), including windows-based GUIs, command line interfaces, and even virtual terminals. This freedom gave rise to the concept of 'graphical user interfaces' (GUIs) that became ubiquitous in

The chat window also includes a text input field at the bottom with the placeholder text 'Type your question...' and a 'Send' button.

Vector store in this RAG system will disappear as soon as Flowise is closed as the vectors are stored in buffer memory.

H. RAG with chroma store and single text file

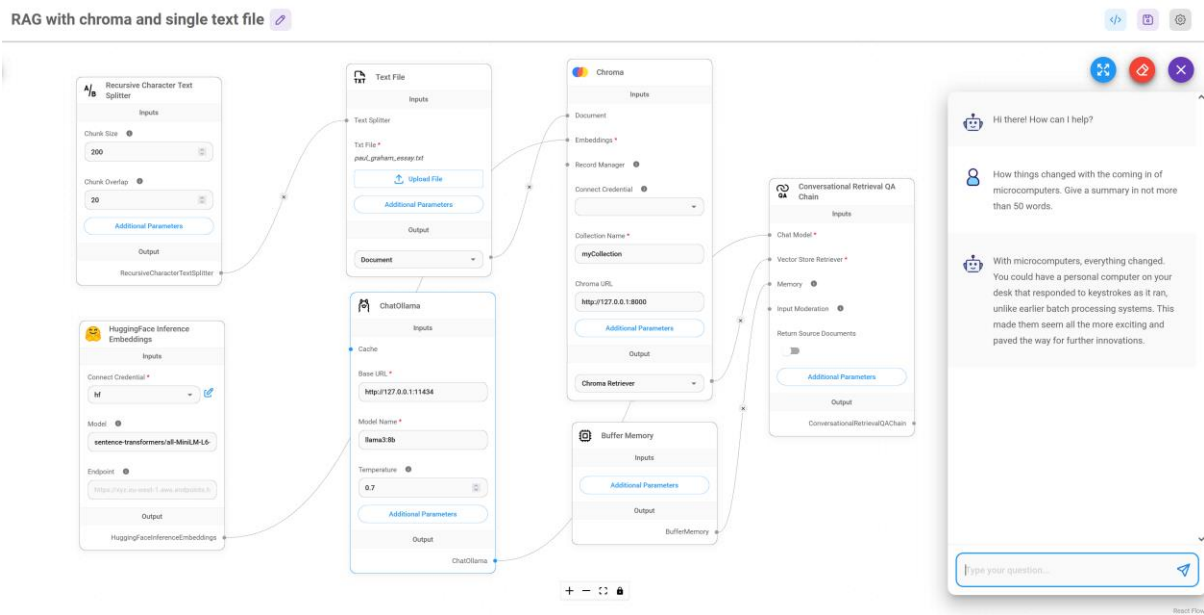


Figure 10: Vector store is replaced by a more durable chroma store. Chroma store retains its vectors even after Flowise is shut down.

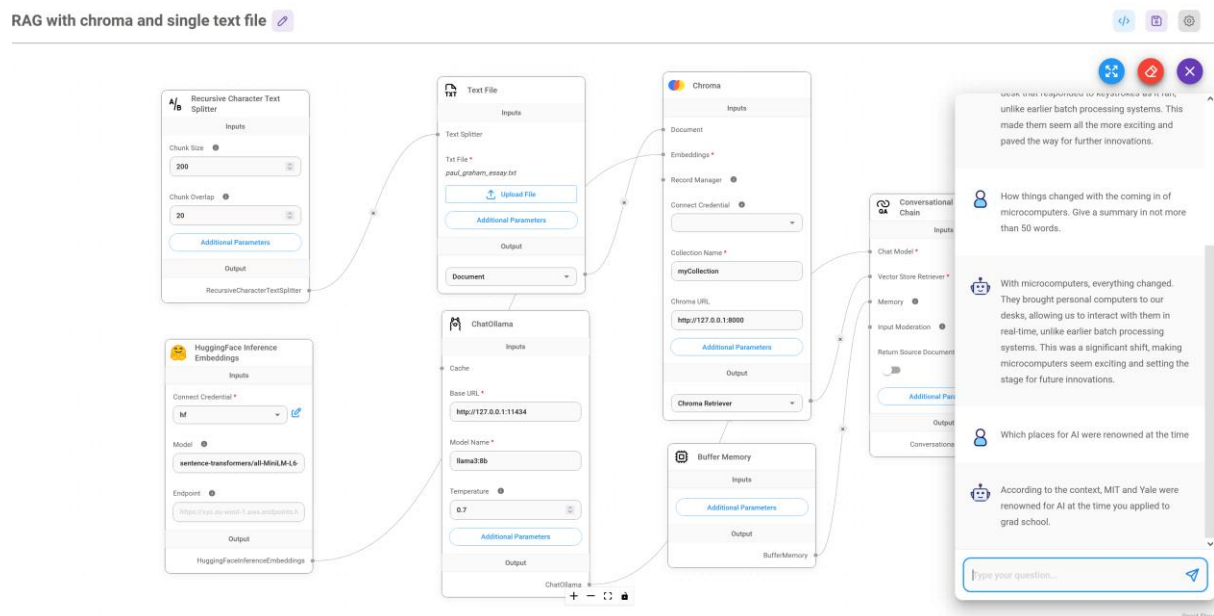


Figure 11: Flowise restarted. More questions asked and replies are given based upon the earlier storage.

I. Combining Multiple Chains (Prompt Chaining)

A. Refer [Flowise tutorial](#)

First LLM chain

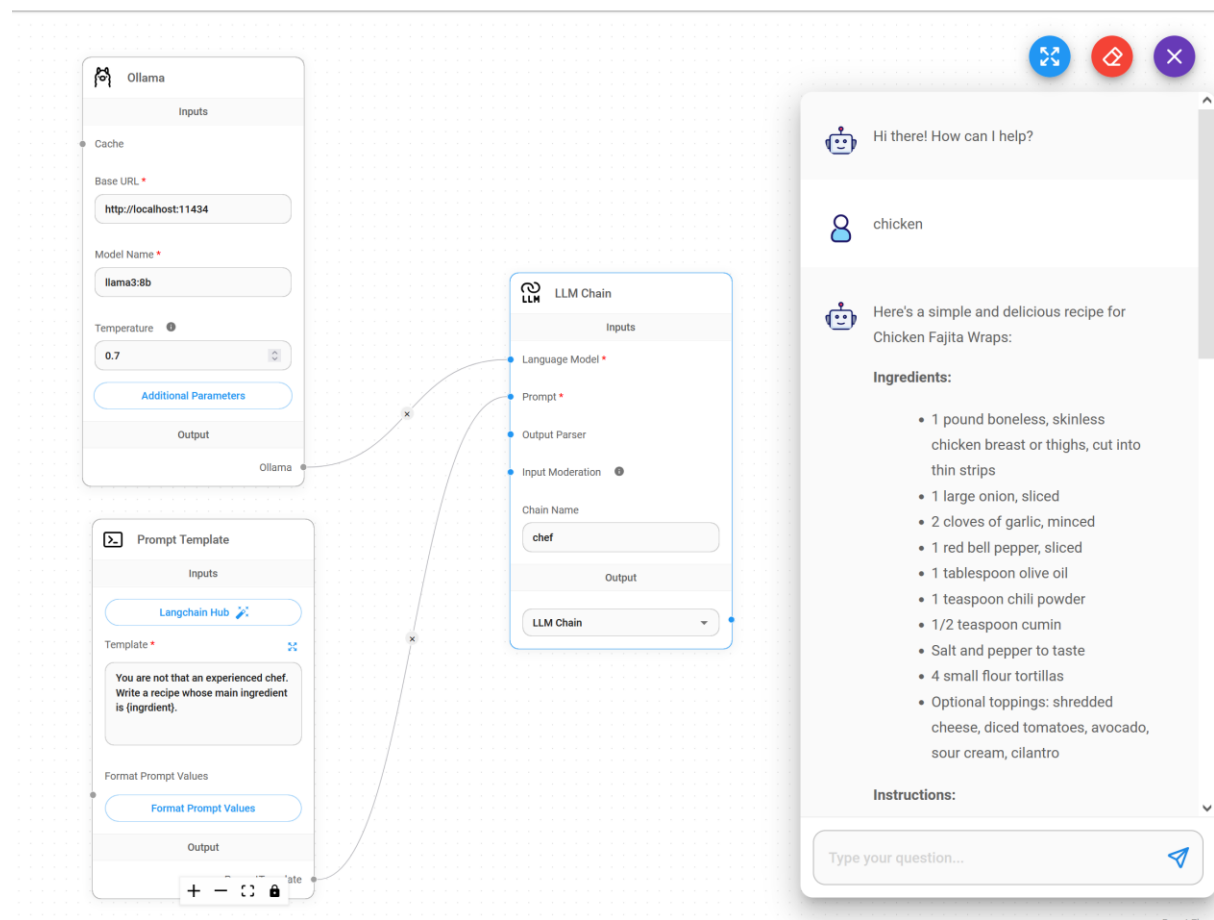


Figure 12: First LLM chain named as chef.

Second LLM chain added to first

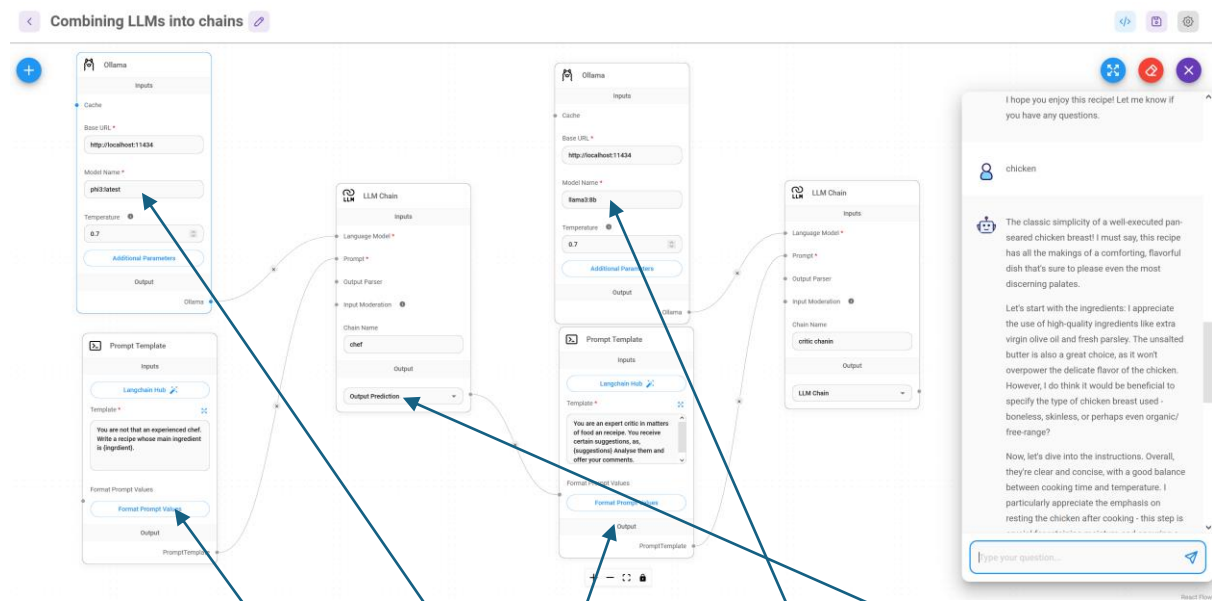


Figure 13: Note that the first LLM is using **phi3** and the critic language model is **llama3**. In the chatbox, the **llm** LLM chain does make minor suggestions to improve the quality. Note that the output of 1st LLM chain is now **Output Prediction**.

Here are the prompts used:

chef chain prompt: *You are not that an experienced chef. Write a recipe whose main ingredient is {ingredient}. Formatting of prompt values is as:*

```
Format Prompt Values
{ 1 item
  ingredient: "{{question}}"
}
```

Critic chain prompt: *You are an expert critic in matters of food an receipe. You receive certain suggestions, as, {suggestions} Analyse them and offer your comments. The formatting of prompt values is as:*

```
Format Prompt Values
{ 1 item
  suggestions: "{{llmChain_0.data.instance}}"
}
```

J. Flowise Using Hugging Face Models

Ref: [This link.](#)

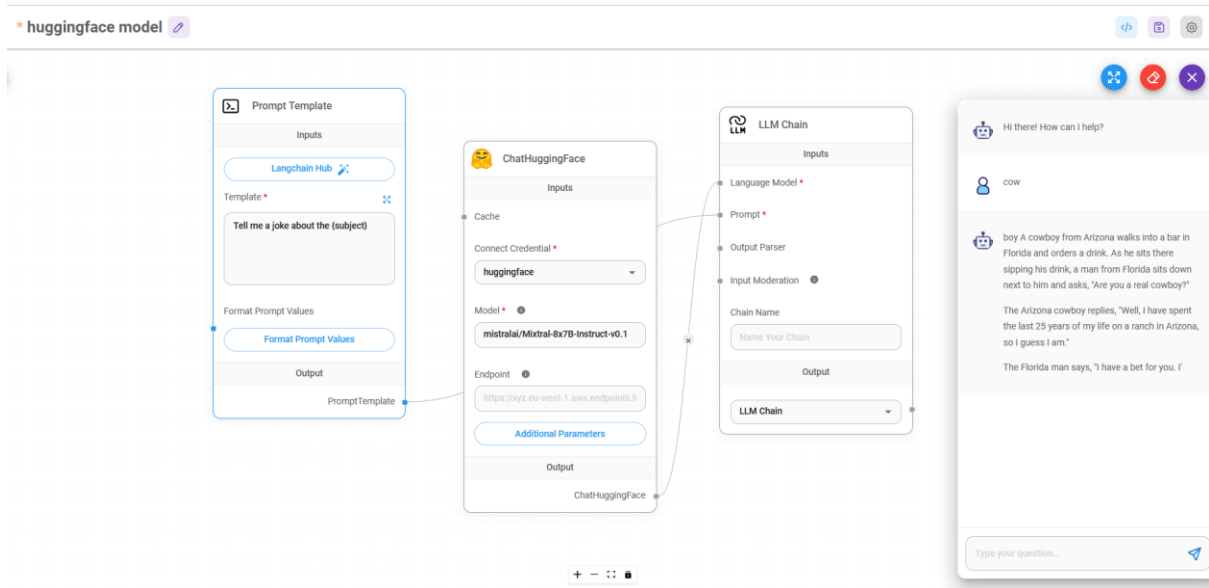


Figure 14: Only those models will work who have made available inference endpoints free.
