Carnegie Mellon University

Lecture 5: LLM Application Design (Cont'd)

SPRING 2025

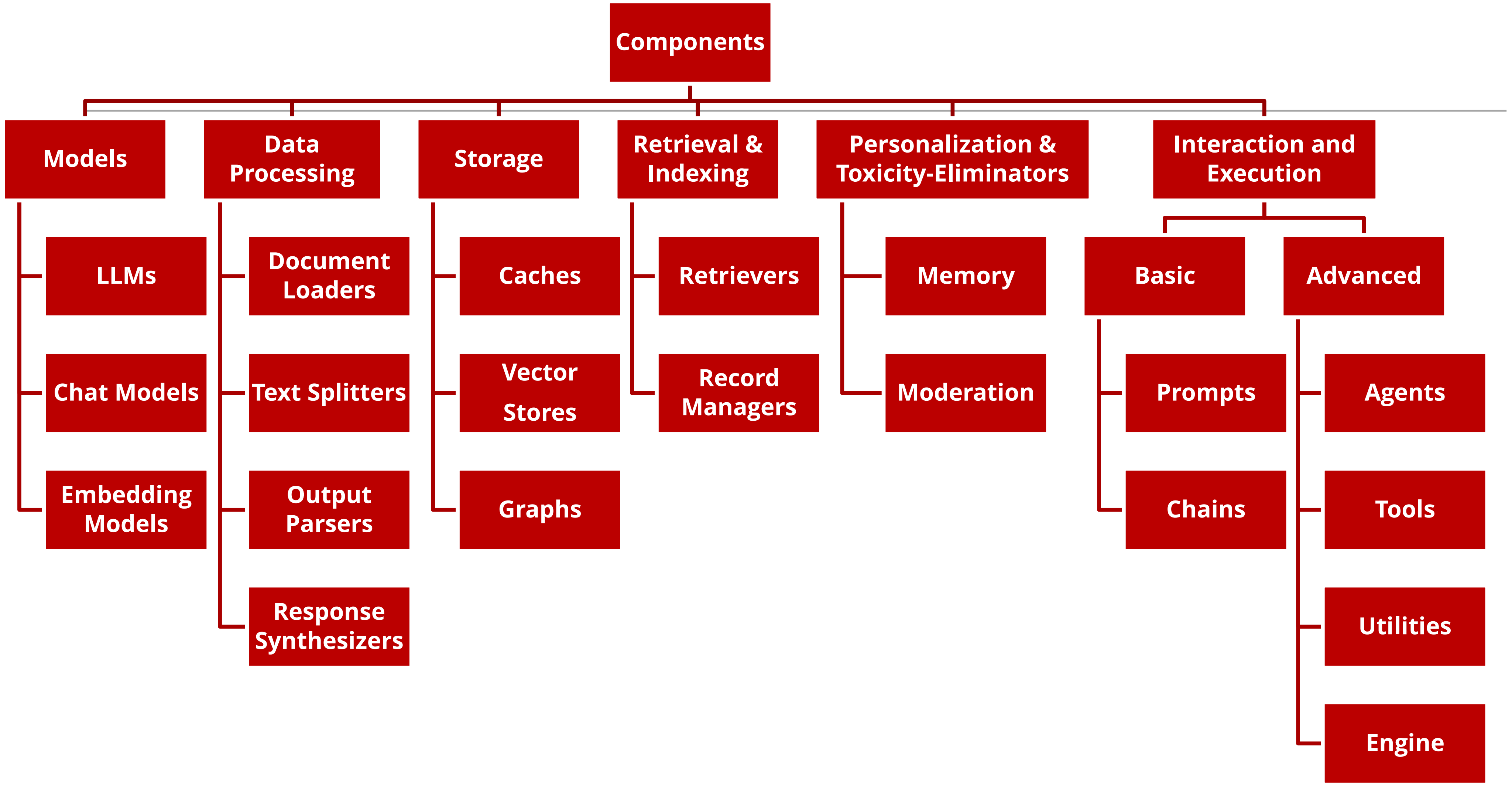
MOHAMED FARAG

FARAG@CMU.EDU

Agenda

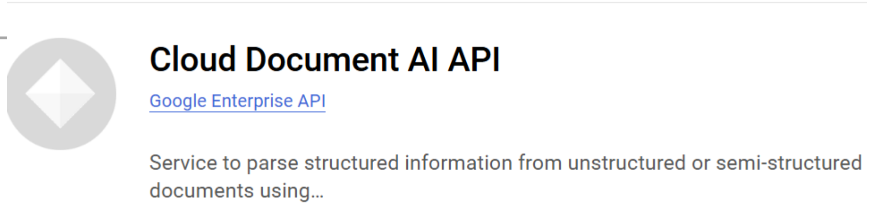
- Enable Vertex AI for Document Processing
- Back to Basics: LLM Chain Design
  - Moderation
  - Output Parsers
  - If-Else Logic
- Multi-agent LLM Applications
- Connect Flowise AI to Local Applications
- Deploy Flowise AI Application
- FlowiseAI Limitations

Reminder: FlowiseAI Components



Enable Vertex AI for Document Processing

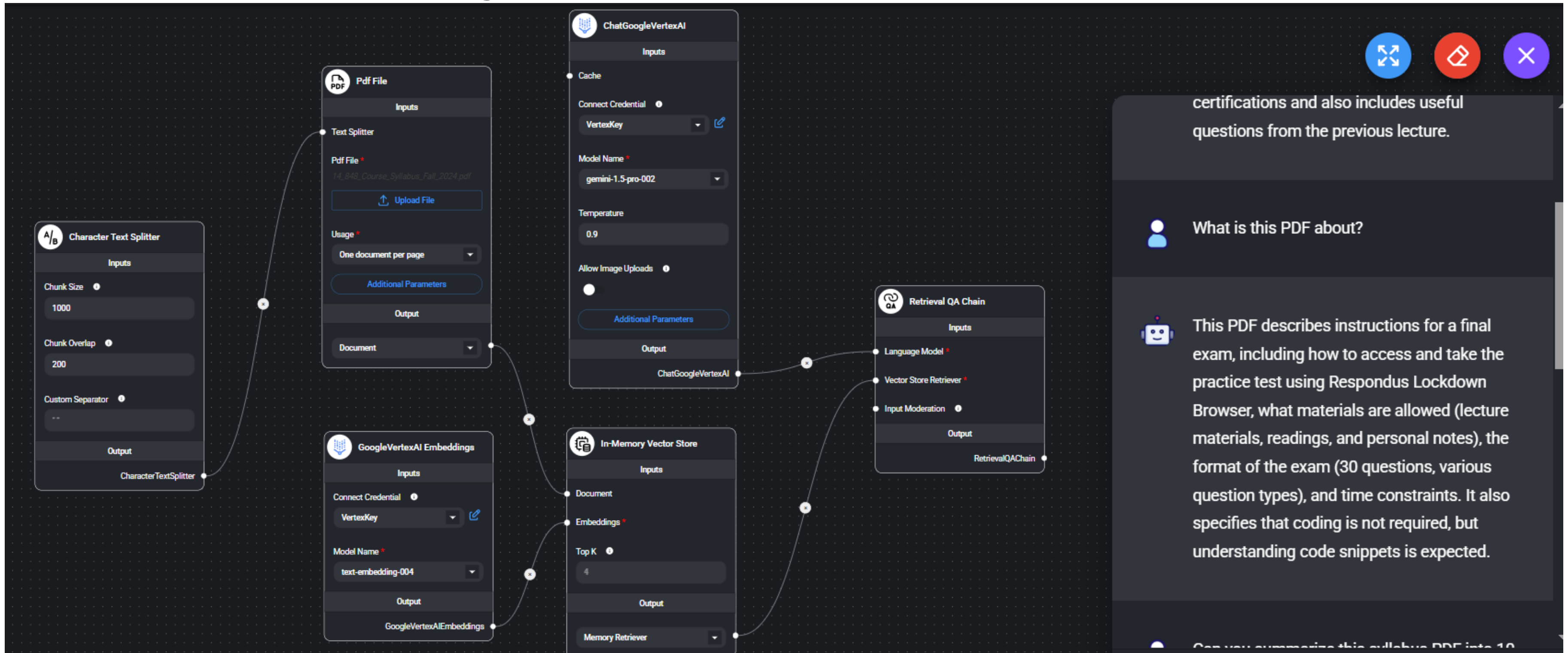
- Enable Cloud Documents AI API
- Use Gemini-models only.



Cloud Document AI API

Google Enterprise API

Service to parse structured information from unstructured or semi-structured documents using...



Character Text Splitter
Inputs
Chunk Size: 1000
Chunk Overlap: 200
Custom Separator: --
Output: CharacterTextSplitter

Pdf File
Inputs
Text Splitter
Pdf File *: 14_848_Course_Syllabus_Fall_2024.pdf
Upload File
Usage *: One document per page
Additional Parameters
Output: Document

GoogleVertexAI Embeddings
Inputs
Connect Credential: VertexKey
Model Name *: text-embedding-004
Output: GoogleVertexAIEmbeddings

ChatGoogleVertexAI
Inputs
Cache
Connect Credential: VertexKey
Model Name *: gemini-1.5-pro-002
Temperature: 0.9
Allow Image Uploads
Additional Parameters
Output: ChatGoogleVertexAI

In-Memory Vector Store
Inputs
Document
Embeddings *
Top K: 4
Output: Memory Retriever

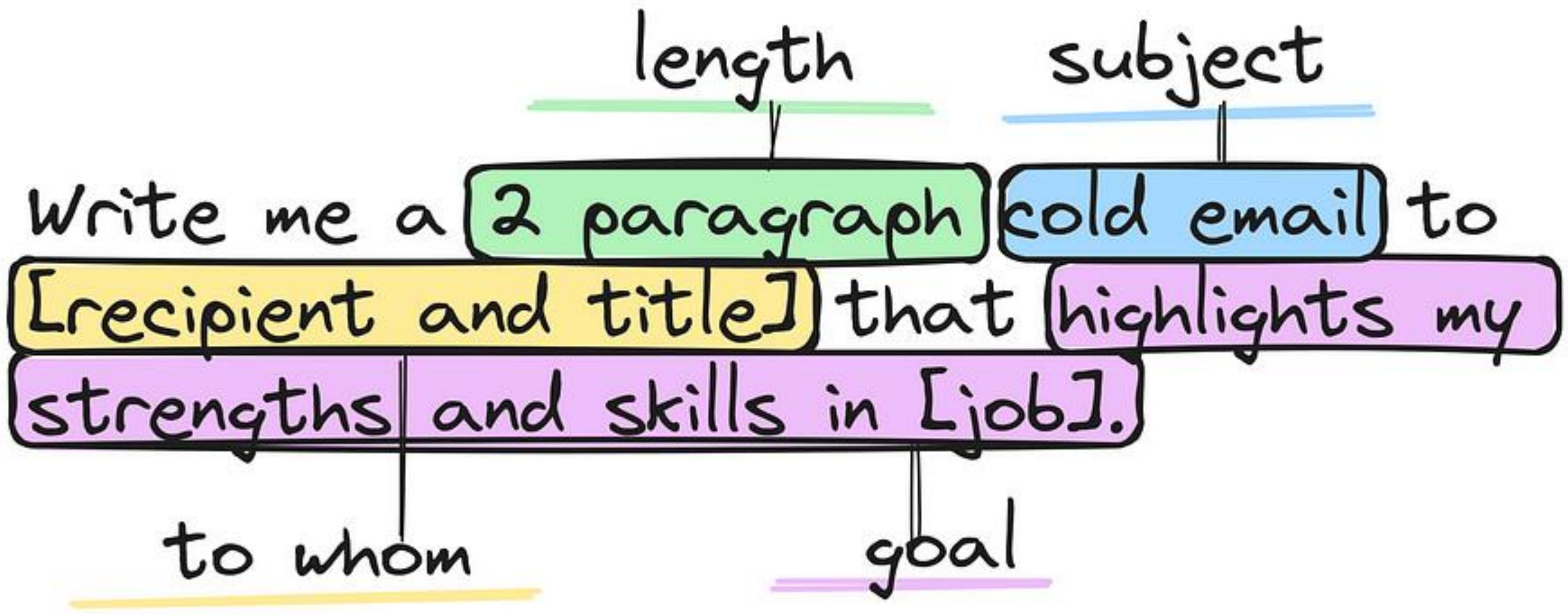
Retrieval QA Chain
Inputs
Language Model *
Vector Store Retriever *
Input Moderation
Output: RetrievalQAChain

certifications and also includes useful questions from the previous lecture.

What is this PDF about?

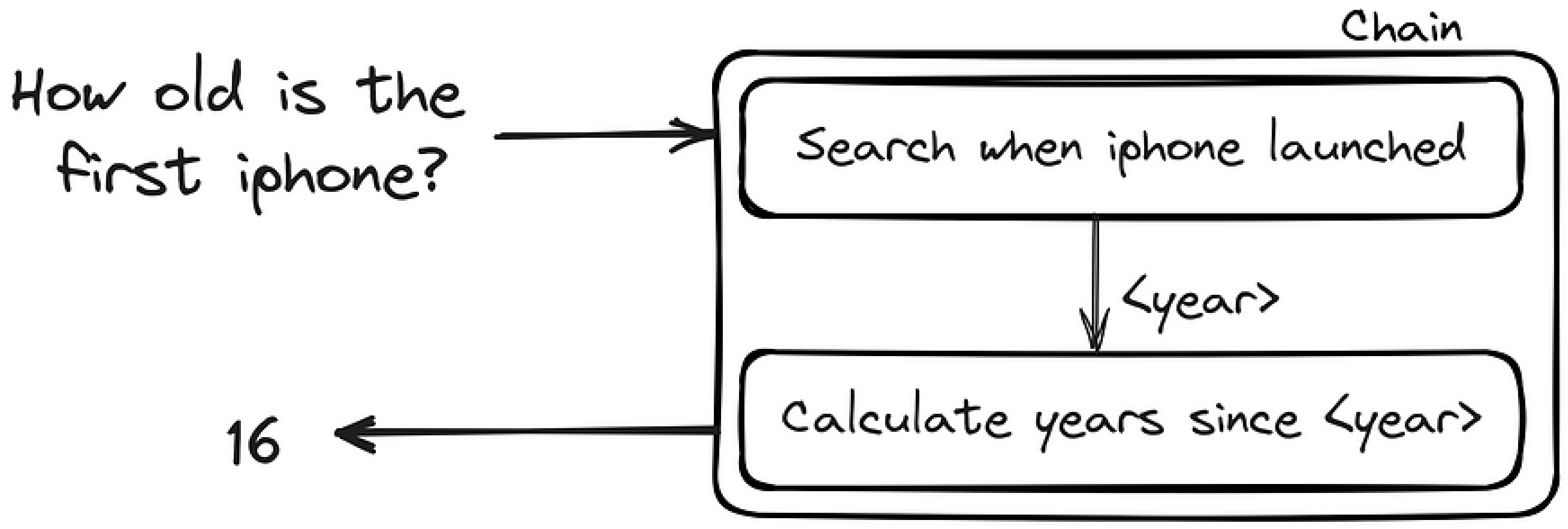
This PDF describes instructions for a final exam, including how to access and take the practice test using Respondus Lockdown Browser, what materials are allowed (lecture materials, readings, and personal notes), the format of the exam (30 questions, various question types), and time constraints. It also specifies that coding is not required, but understanding code snippets is expected.

FlowiseAI Components: Prompts


length
subject
Write me a 2 paragraph cold email to [recipient and title] that highlights my strengths and skills in [job].
to whom
goal

FlowiseAI Components: Chains

Combine LLMs and prompt template to enhance multistep workflows.


Chain
How old is the first iphone?
Search when iphone launched
<year>
Calculate years since <year>
16

Steps to Create LLM Chain

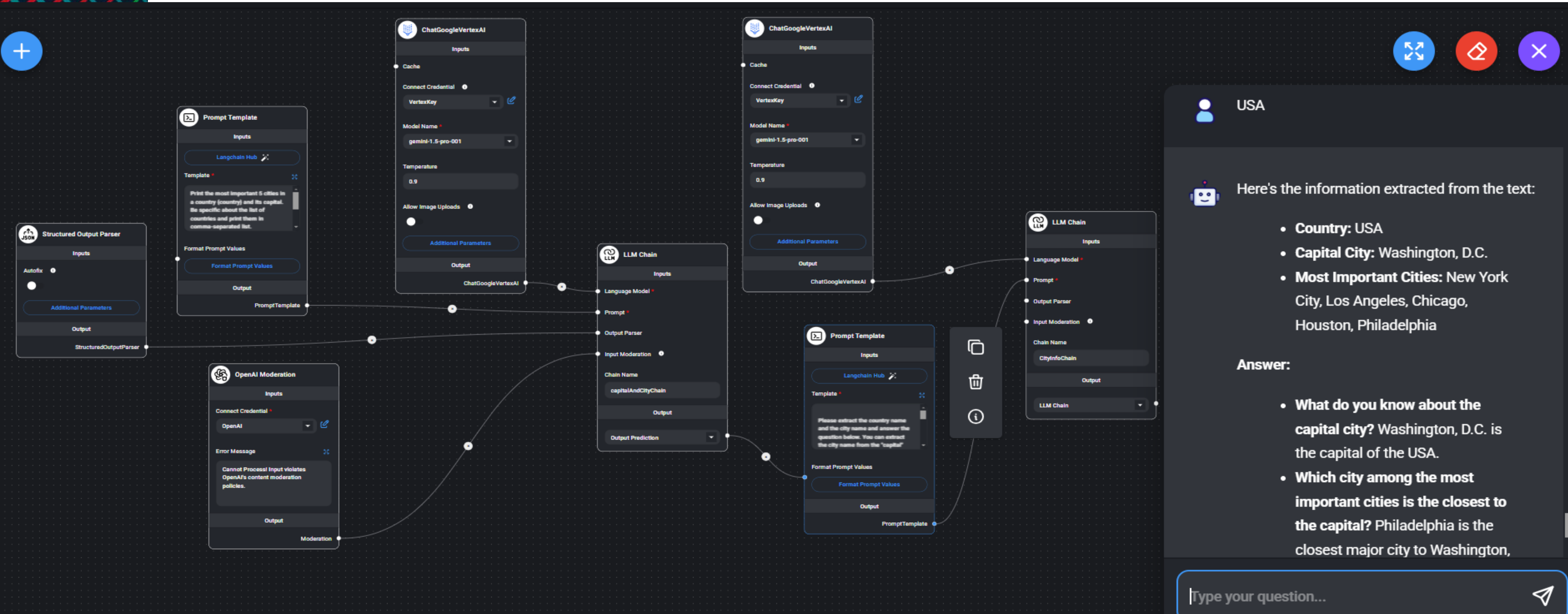
- Identify your LLM Model.
- Create a detailed prompt for your application.
  - A good prompt should provide information on the output format, structure, and any other useful details. LLM models have different output expectations.
- Assess your need for output parsers
- Add input moderation
- Debug FlowiseAI issues as needed.

Carnegie Mellon University

Chain Exercise

- Create an LLM Chain that identifies the capital and the most important 5 cities in a specific country.
- Then, list the closest city among the important cities to the capital

Chain Exercise – Sample Solution
(With Output Parsers)



Structured Output Parser
Inputs
Autofix
Additional Parameters
Output
StructuredOutputParser

Prompt Template
Inputs
Langchain Hub
Template *
Print the most important 5 cities in a country {country} and its capital. Be specific about the list of countries and print them in comma-separated list.
Format Prompt Values
Output
PromptTemplate

OpenAI Moderation
Inputs
Connect Credential *
OpenAI
Error Message
Cannot Process! Input violates OpenAI's content moderation policies.
Output
Moderation

ChatGoogleVertexAI
Inputs
Cache
Connect Credential
VertexKey
Model Name *
gemini-1.5-pro-001
Temperature
0.9
Allow Image Uploads
Additional Parameters
Output
ChatGoogleVertexAI

LLM Chain
Inputs
Language Model *
Prompt *
Output Parser
Input Moderation
Chain Name
capitalAndCityChain
Output
Output Prediction

ChatGoogleVertexAI
Inputs
Cache
Connect Credential
VertexKey
Model Name *
gemini-1.5-pro-001
Temperature
0.9
Allow Image Uploads
Additional Parameters
Output
ChatGoogleVertexAI

Prompt Template
Inputs
Langchain Hub
Template *
Please extract the country name and the city name and answer the question below. You can extract the city name from the "capital"
Format Prompt Values
Output
PromptTemplate

LLM Chain
Inputs
Language Model *
Prompt *
Output Parser
Input Moderation
Chain Name
CityInfoChain
Output
LLM Chain

USA

Here's the information extracted from the text:

- **Country:** USA
- **Capital City:** Washington, D.C.
- **Most Important Cities:** New York City, Los Angeles, Chicago, Houston, Philadelphia

Answer:

- **What do you know about the capital city?** Washington, D.C. is the capital of the USA.
- **Which city among the most important cities is the closest to the capital?** Philadelphia is the closest major city to Washington,

Type your question...

Chain Exercise – Sample Solution – Cont’d

- **Selected LLM Model**: gemini-1.5-pro-001

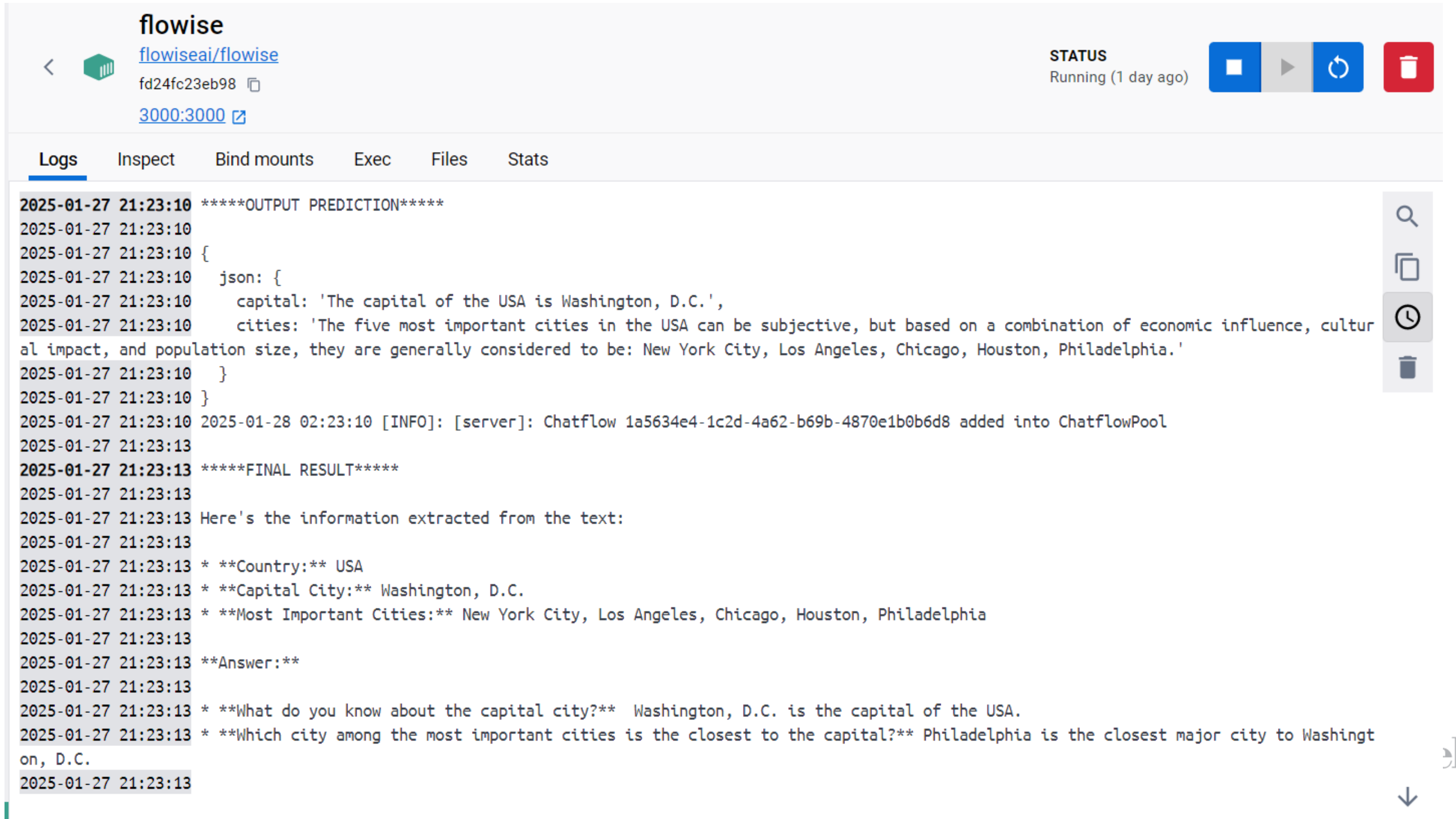
- **LLM Chain-1 Prompt Template:**
  Print the most important 5 cities in a country {country} and its capital. Be specific about the list of countries and print them in comma-separated list. Produce the output in the format of two paragraphs. First paragraph starts with "capital:" and second paragraph starts with "cities:“

- **LLM Chain-2 Prompt Template:**
  Please extract the country name and the city name and answer the question below. You can extract the city name from the "capital" attribute in the string. Also, please extract the most important cities in this country from the "cities" attribute in the string.
  Country: {country}
  City: {cityChain}
  Most Important Cities: {cityChain}

  What do you know about the capital city? Which city among the most important cities is the closest to the capital?

Food For Thought:
Why would you use multiple prompt templates instead of one?

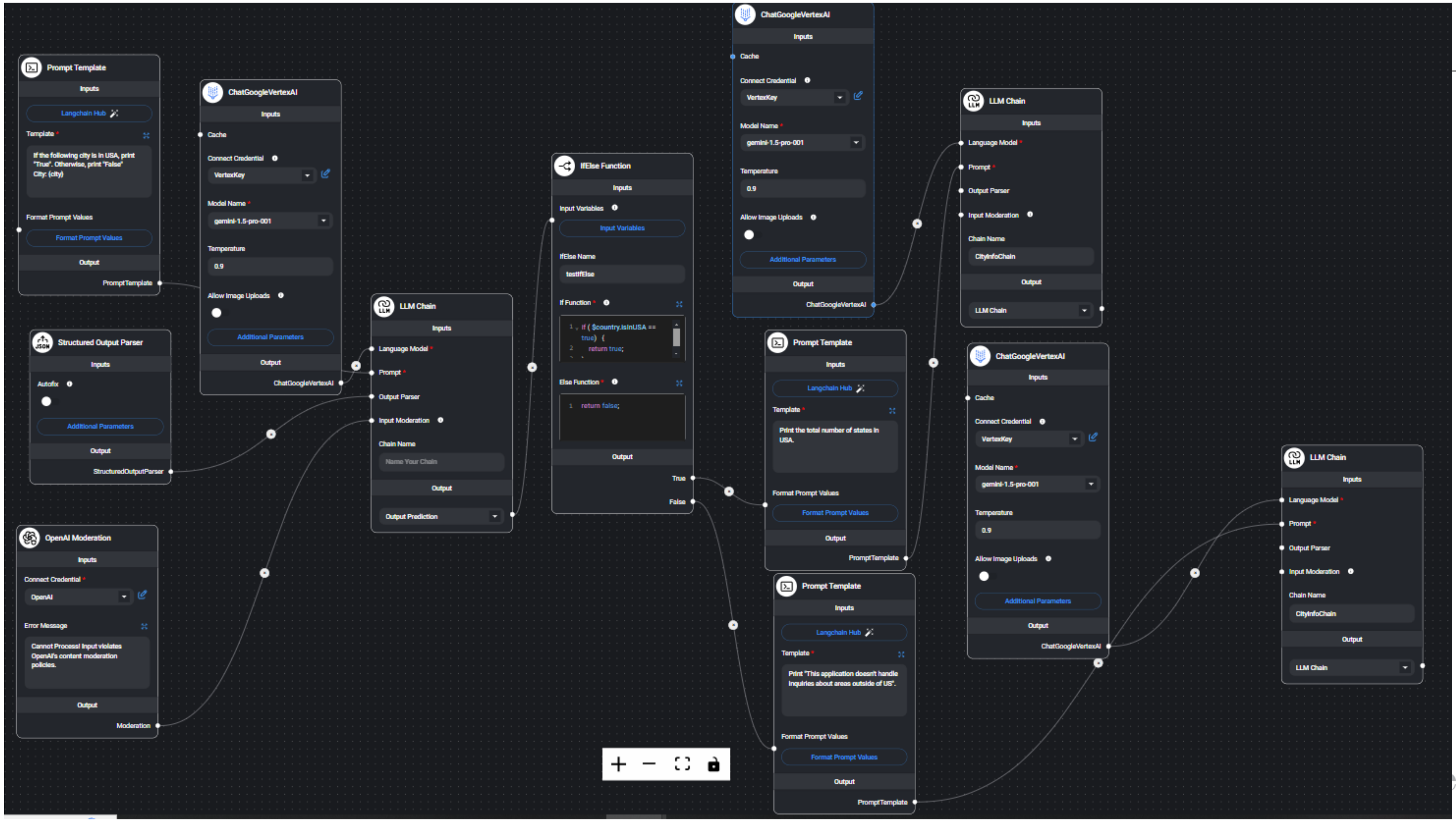
LLM Application Debugging in FlowiseAI



```
2025-01-27 21:23:10 *****OUTPUT PREDICTION*****
2025-01-27 21:23:10
2025-01-27 21:23:10 {
2025-01-27 21:23:10   json: {
2025-01-27 21:23:10     capital: 'The capital of the USA is Washington, D.C.',
2025-01-27 21:23:10     cities: 'The five most important cities in the USA can be subjective, but based on a combination of economic influence, cultur
al impact, and population size, they are generally considered to be: New York City, Los Angeles, Chicago, Houston, Philadelphia.'
2025-01-27 21:23:10   }
2025-01-27 21:23:10 }
2025-01-27 21:23:10 2025-01-28 02:23:10 [INFO]: [server]: Chatflow 1a5634e4-1c2d-4a62-b69b-4870e1b0b6d8 added into ChatflowPool
2025-01-27 21:23:13
2025-01-27 21:23:13 *****FINAL RESULT*****
2025-01-27 21:23:13
2025-01-27 21:23:13 Here's the information extracted from the text:
2025-01-27 21:23:13
2025-01-27 21:23:13 * **Country:** USA
2025-01-27 21:23:13 * **Capital City:** Washington, D.C.
2025-01-27 21:23:13 * **Most Important Cities:** New York City, Los Angeles, Chicago, Houston, Philadelphia
2025-01-27 21:23:13
2025-01-27 21:23:13 **Answer:**
2025-01-27 21:23:13
2025-01-27 21:23:13 * **What do you know about the capital city?**  Washington, D.C. is the capital of the USA.
2025-01-27 21:23:13 * **Which city among the most important cities is the closest to the capital?** Philadelphia is the closest major city to Washingt
on, D.C.
2025-01-27 21:23:13
```

LLM Application Design using IfElse Utility

Note: Define Input Variables on ifElse Utility and Access Variables in JSON format



Prompt Template

Template: If the following city is in USA, print "True". Otherwise, print "False"
City: {city}

ChatGoogleVertexAI

Connect Credential: VertexKey
Model Name: gemini-1.5-pro-001
Temperature: 0.9

Structured Output Parser

OpenAI Moderation

Connect Credential: OpenAI
Error Message: Cannot Process! Input violates OpenAI's content moderation policies.

LLM Chain

Chain Name: Name Your Chain
Output: Output Prediction

IfElse Function

IfElse Name: testIfElse

If Function:

```
if ( $country.isInUSA == true) {
    return true;
```

Else Function:

```
return false;
```

Output: True / False

ChatGoogleVertexAI

Connect Credential: VertexKey
Model Name: gemini-1.5-pro-001
Temperature: 0.9

Prompt Template

Template: Print the total number of states in USA.

LLM Chain

Chain Name: CityInfoChain
Output: LLM Chain

Prompt Template

Template: Print "This application doesn't handle inquiries about areas outside of US".

ChatGoogleVertexAI

Connect Credential: VertexKey
Model Name: gemini-1.5-pro-001
Temperature: 0.9

LLM Chain

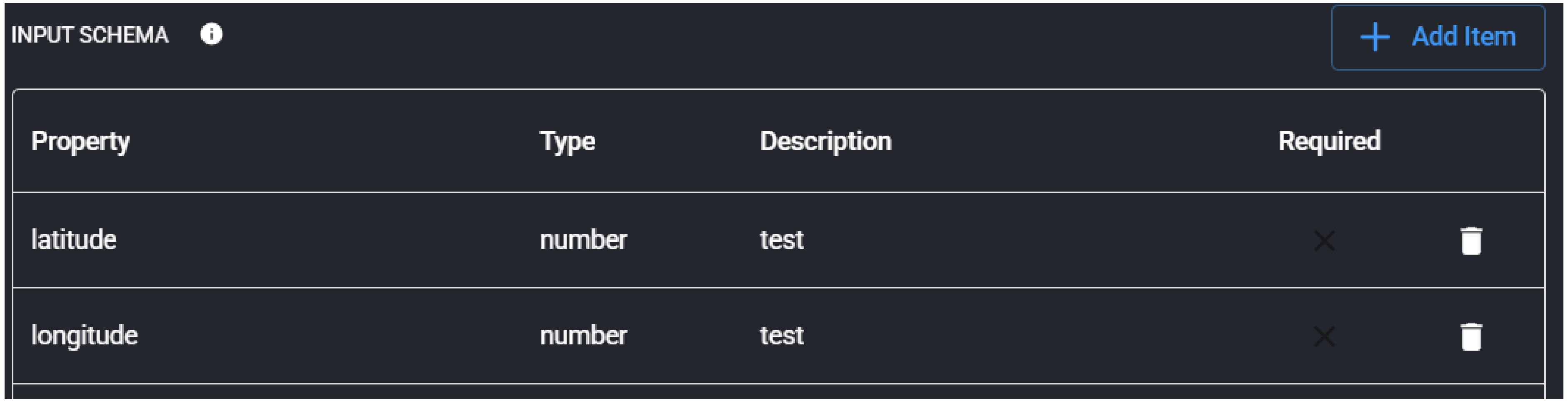
Chain Name: CityInfoChain
Output: LLM Chain

Reminder: Custom Tool Exercise

Build Custom Tool for retrieving weather by Lat/Long

Custom Tools Creation

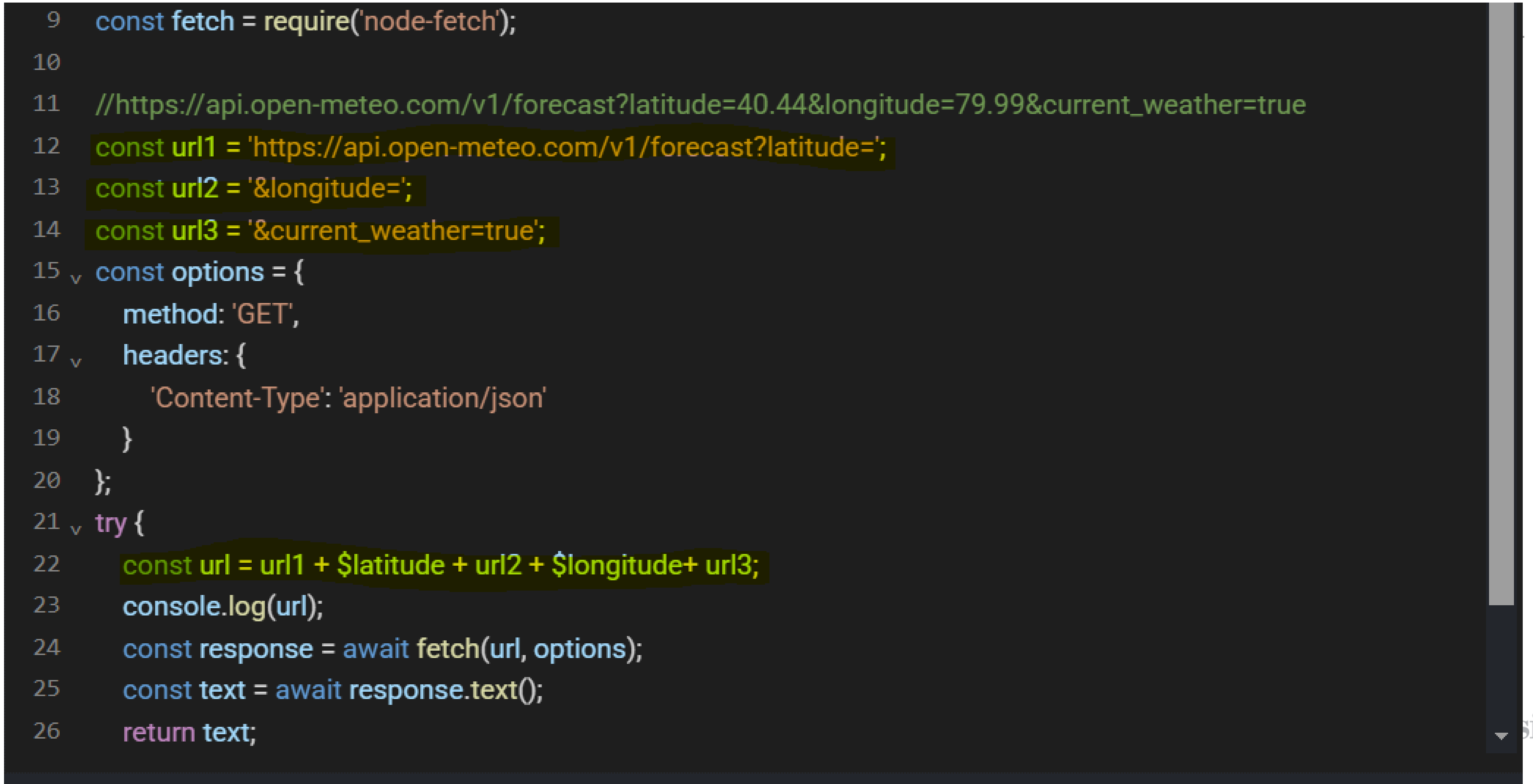
- Create a custom tool with the API URL and proper arguments.
  API URL: http://api.open-meteo.com/v1/forecast?latitude=40.44&longitude=79.99&current_weather=true



Property	Type	Description	Required
latitude	number	test	✗
longitude	number	test	✗

Carnegie Mellon University

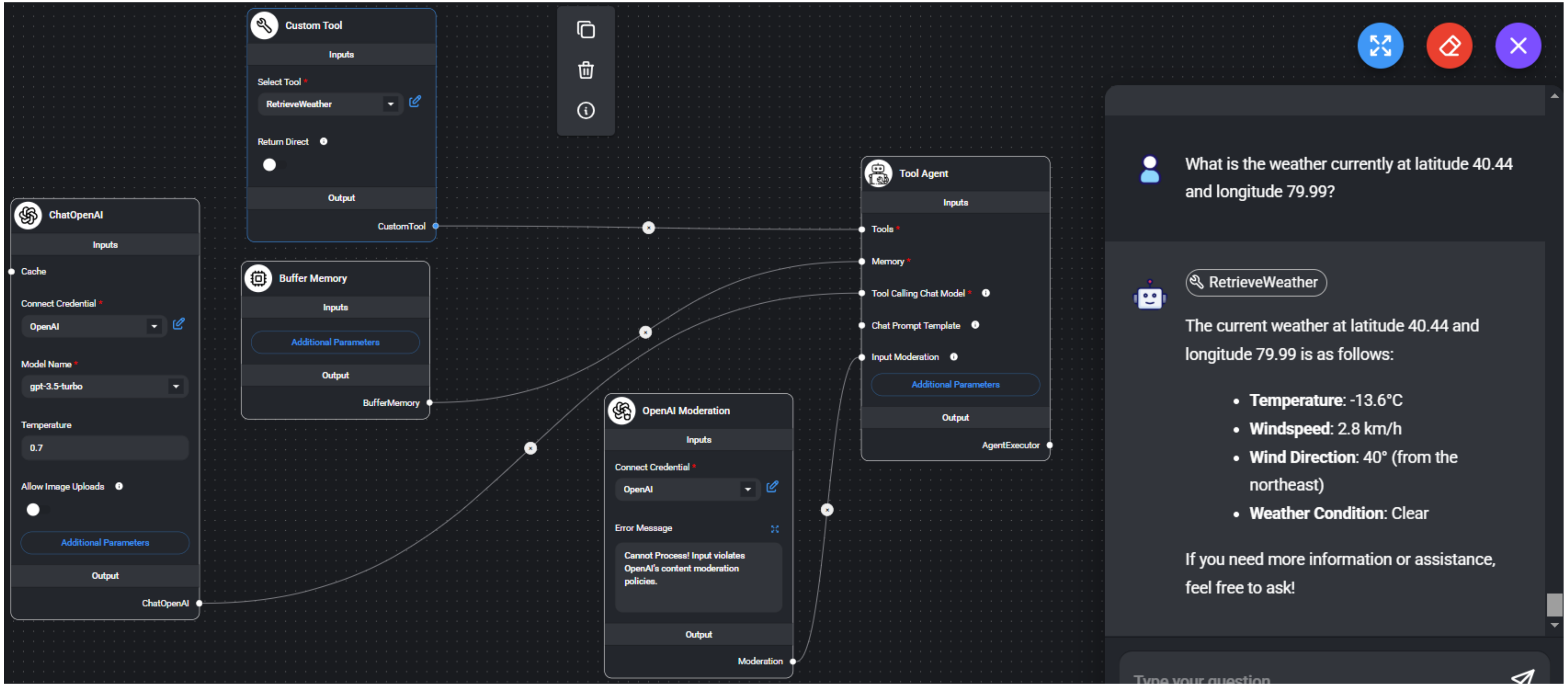
Custom Tools Creation – Cont’d



```
const fetch = require('node-fetch');

//https://api.open-meteo.com/v1/forecast?latitude=40.44&longitude=79.99&current_weather=true
const url1 = 'https://api.open-meteo.com/v1/forecast?latitude=';
const url2 = '&longitude=';
const url3 = '&current_weather=true';
const options = {
  method: 'GET',
  headers: {
    'Content-Type': 'application/json'
  }
};
try {
  const url = url1 + $latitude + url2 + $longitude+ url3;
  console.log(url);
  const response = await fetch(url, options);
  const text = await response.text();
  return text;
```

sity

Weather Application Design



Cannot Process! Input violates OpenAI's content moderation policies.

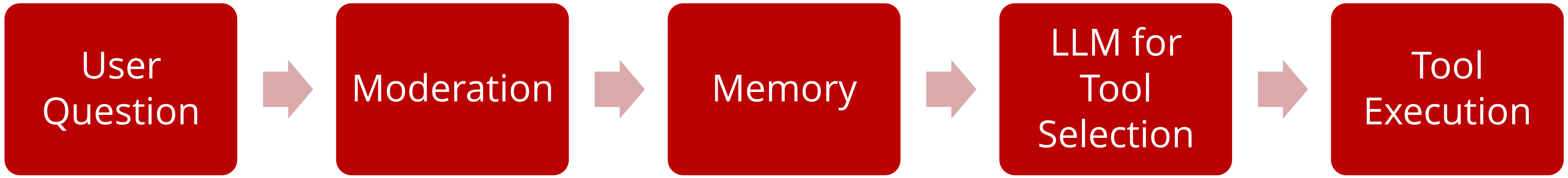
What is the weather currently at latitude 40.44 and longitude 79.99?

The current weather at latitude 40.44 and longitude 79.99 is as follows:

- **Temperature**: -13.6°C
- **Windspeed**: 2.8 km/h
- **Wind Direction**: 40° (from the northeast)
- **Weather Condition**: Clear

If you need more information or assistance, feel free to ask!

Weather Application Design – Cont'd Execution Flow

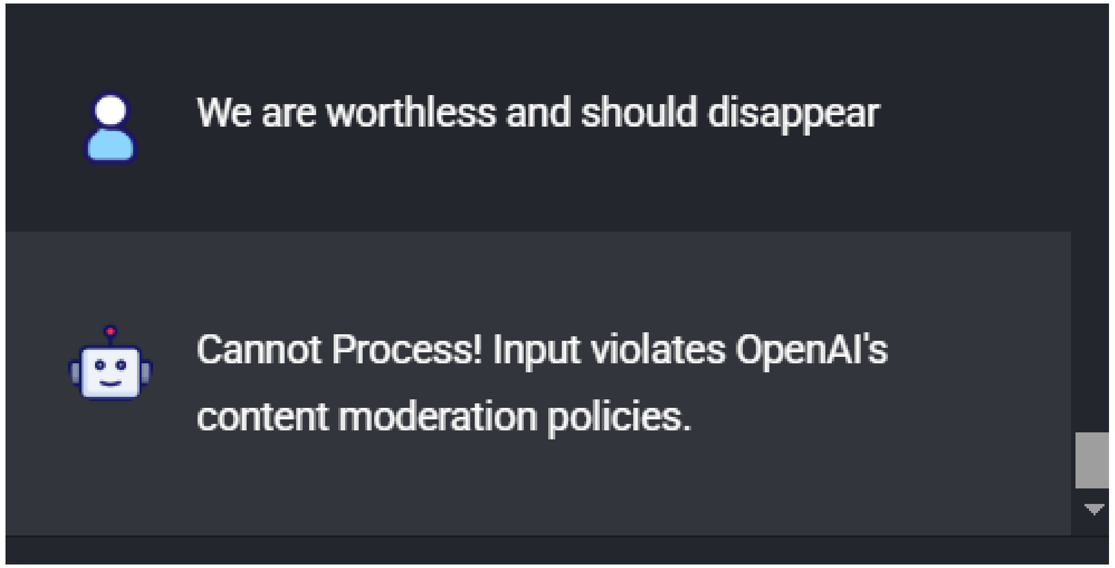


Carnegie Mellon University

Weather Application Design – Cont’d

- **Input Moderation**

“We are worthless and should disappear”



We are worthless and should disappear

Cannot Process! Input violates OpenAI's content moderation policies.

Food For Thought:
How would you connect a chain to an agent?

Carnegie Mellon University

What about recursive Prompt Requests?

It’s counter intuitive to manually require your LLM to generate better quality responses. To handle complex tasks and build recursive systems that can be composed of subsystems:

- Multi-agent System: A scalable framework that breaks down complex projects into manageable tasks, controlled by an agent.
- Sequential Agent System: similar to multi-agent systems but offers higher control and larger capabilities to build more complex systems.

Multi-agent Systems

- User: Initiates requests.
- Supervisor AI: Orchestrates the workflow, breaking down requests into subtasks and assigning them to specialized workers.
- Worker AI Team: Specialized agents that independently execute specific tasks, using necessary tools and returning results to the Supervisor.

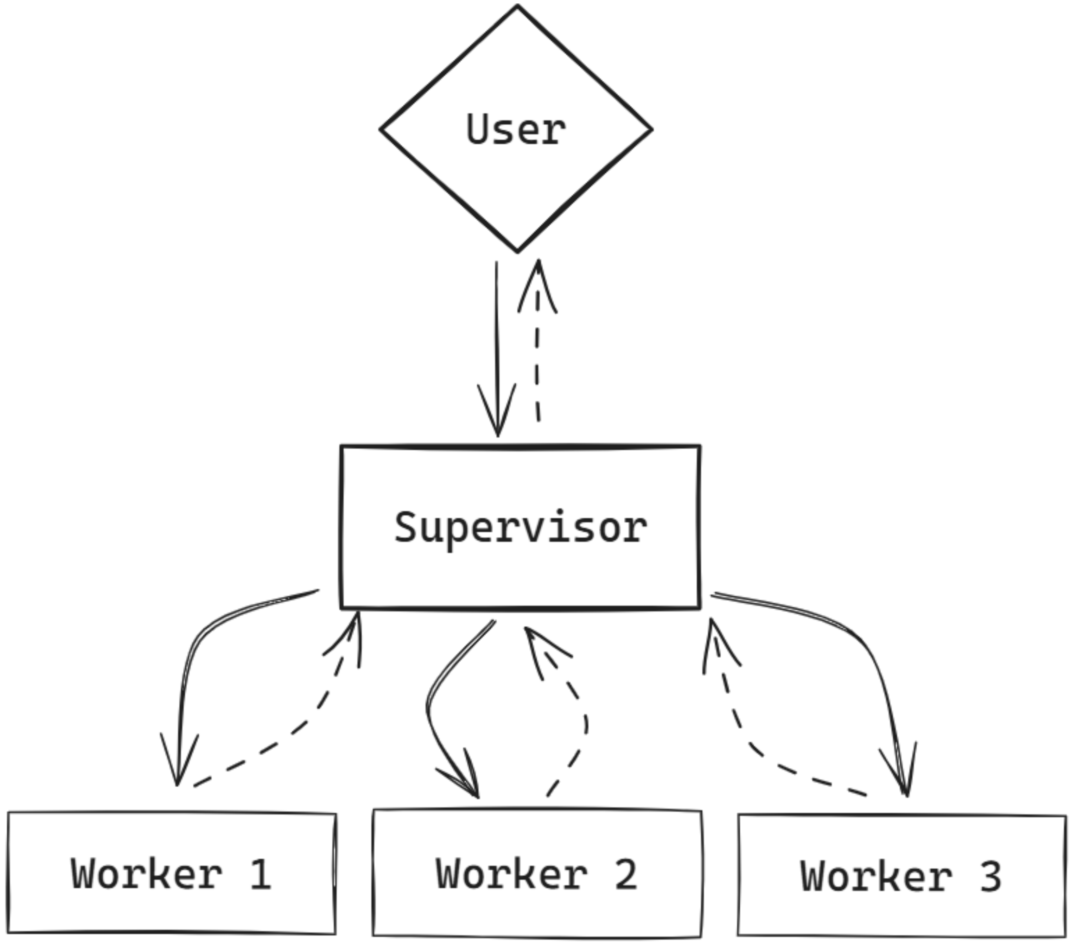
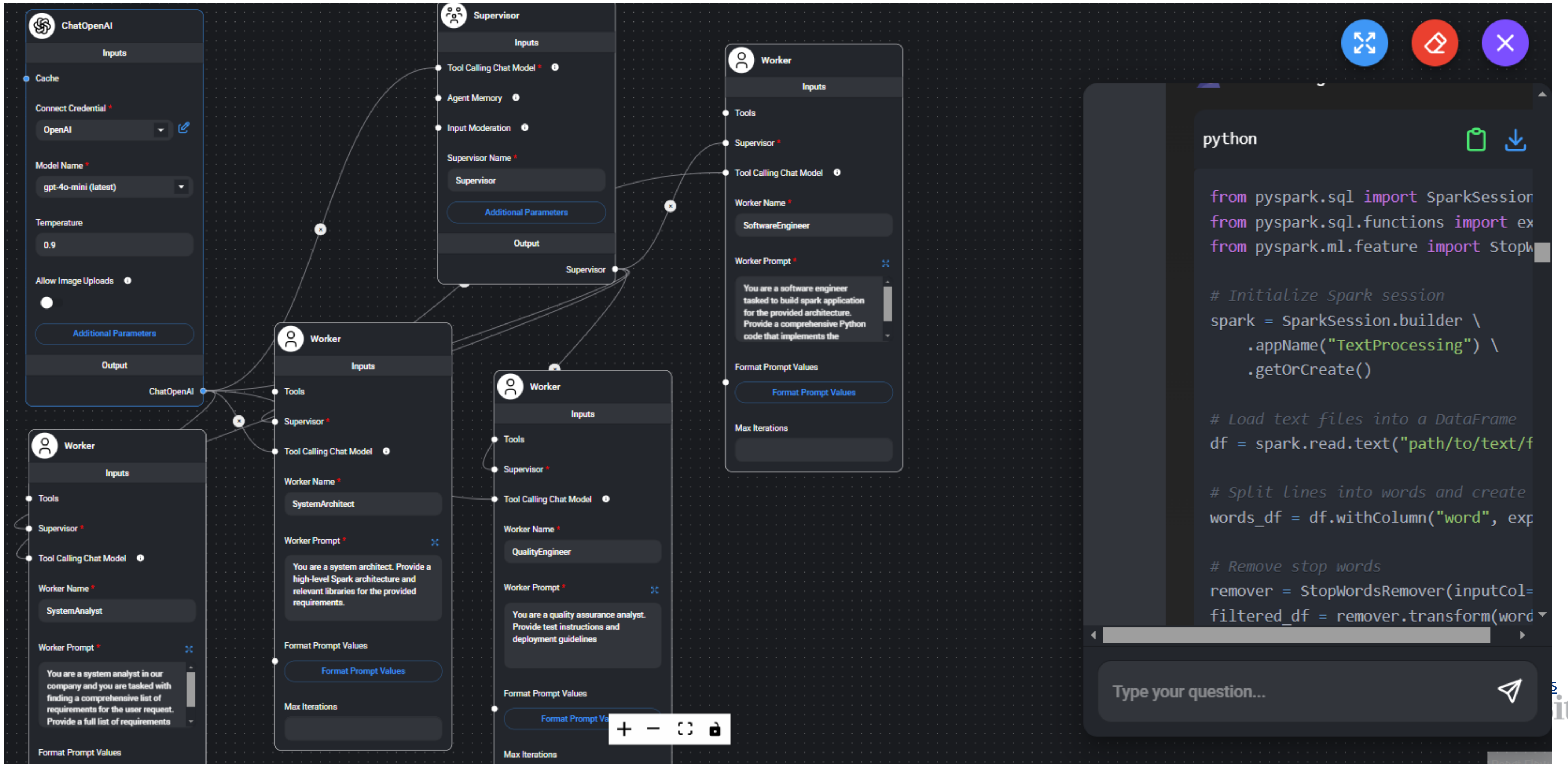


Image source: https://docs.flowiseai.com/using-flowise/agentflows/multi-agents

Multi-agent Systems Example



ChatOpenAI

Inputs

Cache

Connect Credential *: OpenAI

Model Name *: gpt-4o-mini (latest)

Temperature: 0.9

Allow Image Uploads

Additional Parameters

Output

ChatOpenAI

Supervisor

Inputs

Tool Calling Chat Model *

Agent Memory

Input Moderation

Supervisor Name *: Supervisor

Additional Parameters

Output

Supervisor

Worker

Inputs

Tools

Supervisor *

Tool Calling Chat Model

Worker Name *: SystemAnalyst

Worker Prompt *: You are a system analyst in our company and you are tasked with finding a comprehensive list of requirements for the user request. Provide a full list of requirements

Format Prompt Values

Worker

Inputs

Tools

Supervisor *

Tool Calling Chat Model

Worker Name *: SystemArchitect

Worker Prompt *: You are a system architect. Provide a high-level Spark architecture and relevant libraries for the provided requirements.

Format Prompt Values

Format Prompt Values

Max Iterations

Worker

Inputs

Tools

Supervisor *

Tool Calling Chat Model

Worker Name *: QualityEngineer

Worker Prompt *: You are a quality assurance analyst. Provide test instructions and deployment guidelines

Format Prompt Values

Format Prompt Va

Max Iterations

Worker

Inputs

Tools

Supervisor *

Tool Calling Chat Model

Worker Name *: SoftwareEngineer

Worker Prompt *: You are a software engineer tasked to build spark application for the provided architecture. Provide a comprehensive Python code that implements the

Format Prompt Values

Format Prompt Values

Max Iterations

python

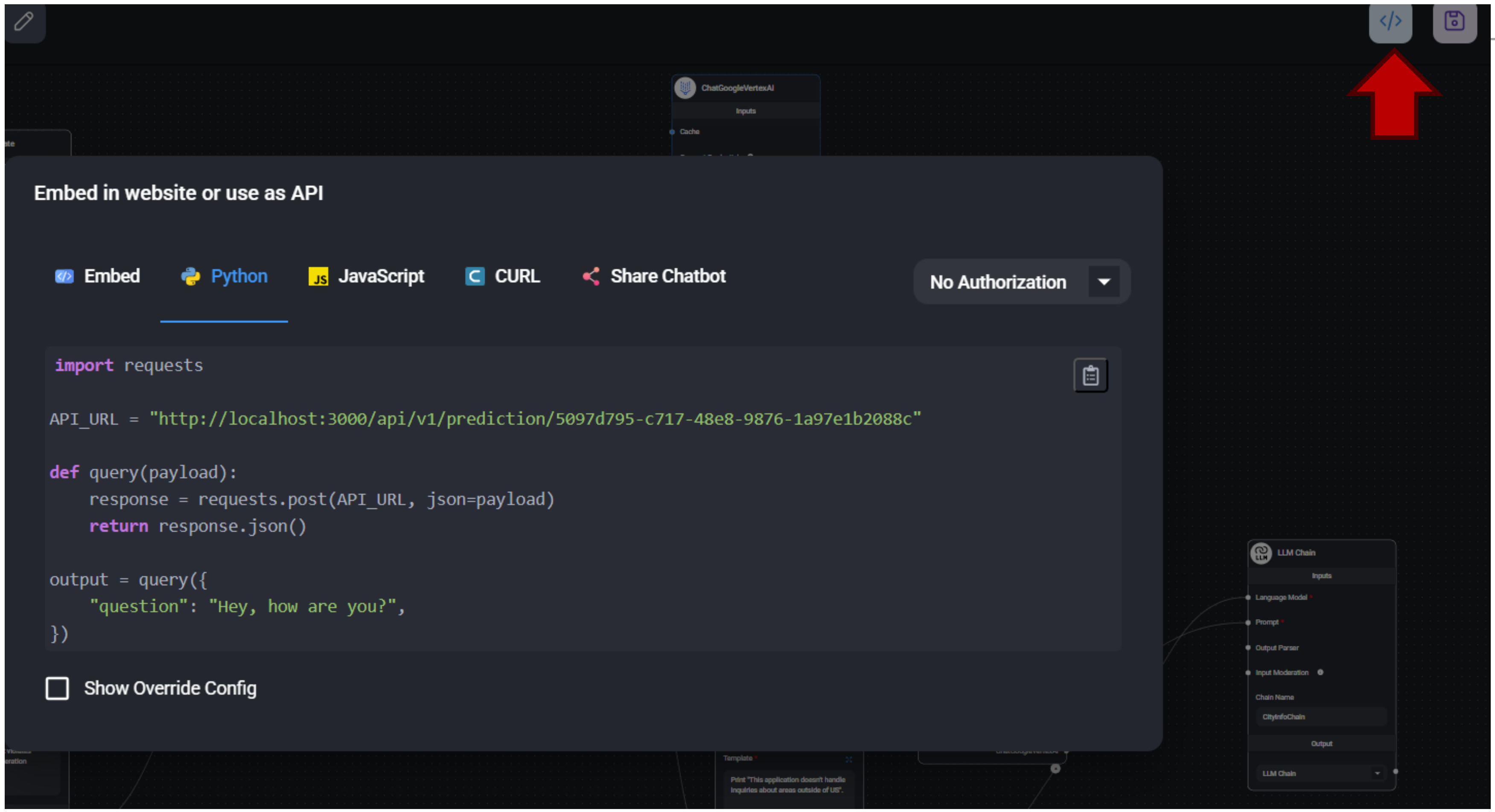
```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import ex
from pyspark.ml.feature import StopW

# Initialize Spark session
spark = SparkSession.builder \
    .appName("TextProcessing") \
    .getOrCreate()

# Load text files into a DataFrame
df = spark.read.text("path/to/text/f

# Split lines into words and create
words_df = df.withColumn("word", exp

# Remove stop words
remover = StopWordsRemover(inputCol=
filtered_df = remover.transform(word
```

Type your question...

Take Home Exercise:
Deploy FlowiseAI to Google Kubernetes Engine

Follow the instructions on this page:
https://docs.flowiseai.com/configuration/deployment/gcp

Connect FlowiseAI Apps to Local Applications



```python
import requests

API_URL = "http://localhost:3000/api/v1/prediction/5097d795-c717-48e8-9876-1a97e1b2088c"

def query(payload):
    response = requests.post(API_URL, json=payload)
    return response.json()

output = query({
    "question": "Hey, how are you?",
})
```

Flowise AI Limitations

- Lack of output moderation
- Limited UI capabilities
- Limited deployment and scalability options

To overcome these limitations, **building LLM applications using coding frameworks is needed**

Readings

- Multi-Agent Systems: https://docs.flowiseai.com/using-flowise/agentflows/multi-agents
- Sequential Agents: https://docs.flowiseai.com/using-flowise/agentflows/sequential-agents