2025

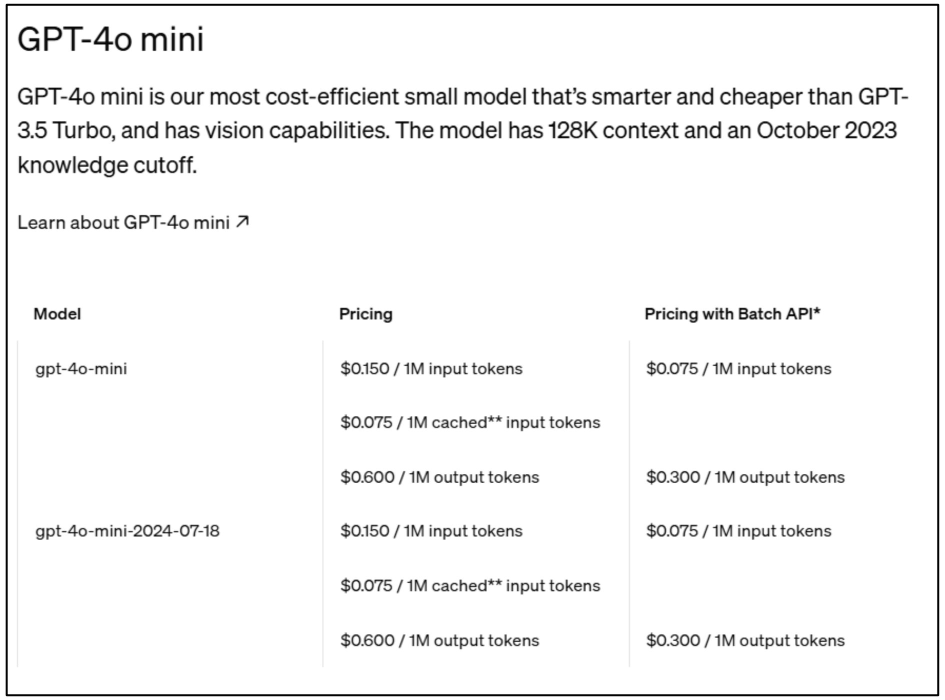
Lesson 08

Low Code LLM App Builder (Part 2)

Tokens and Embeddings

Tokenization (Recap)

- The computers aren't very good with words but are great with numbers.
- Tokenization breaks down a continuous stream of text into smaller, discrete units called tokens.
- Token helps to provides a structured way to break down text into manageable pieces for the model to process.
- Tokenizer is an integral part of a language model used during training.
- Tokens are basic units that the model processes and different tokenizers can produce different token sequences for the same text.
- Understanding token is important because the service providers will charge you by tokens for API calls.



GPT-4o mini

GPT-4o mini is our most cost-efficient small model that's smarter and cheaper than GPT-3.5 Turbo, and has vision capabilities. The model has 128K context and an October 2023 knowledge cutoff.

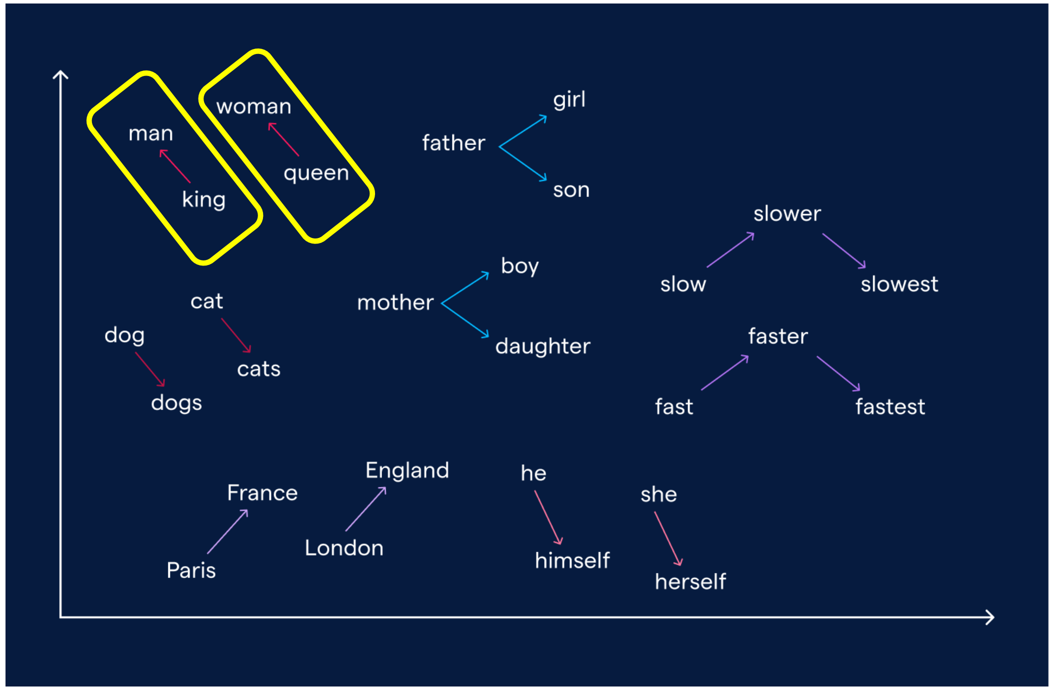
Learn about GPT-4o mini ↗

Model	Pricing	Pricing with Batch API*
gpt-4o-mini	$0.150 / 1M input tokens	$0.075 / 1M input tokens
	$0.075 / 1M cached** input tokens	
	$0.600 / 1M output tokens	$0.300 / 1M output tokens
gpt-4o-mini-2024-07-18	$0.150 / 1M input tokens	$0.075 / 1M input tokens
	$0.075 / 1M cached** input tokens	
	$0.600 / 1M output tokens	$0.300 / 1M output tokens

Source: https://openai.com/api/pricing/

Embedding (Recap)

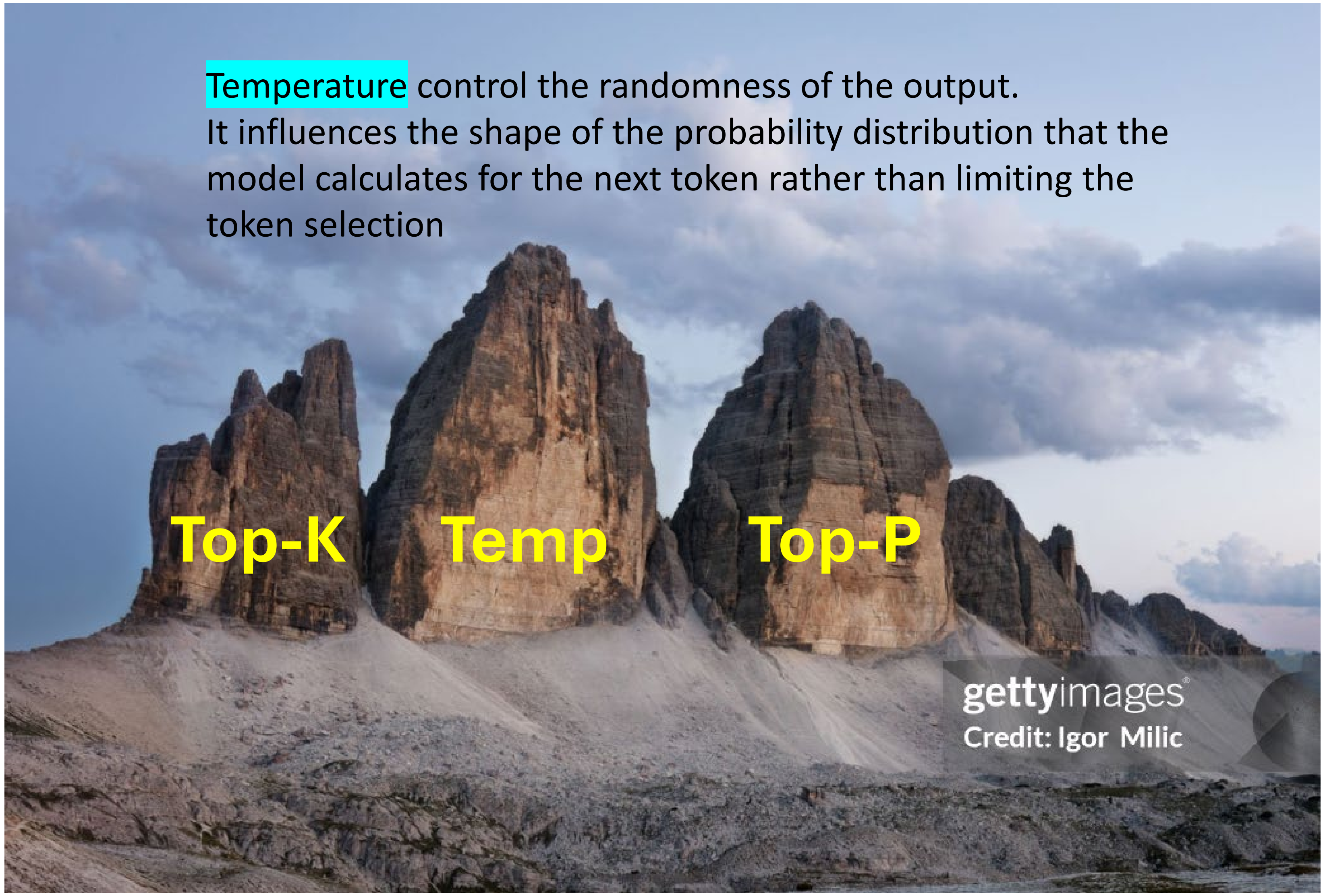
- The next step after tokenizing the text is to convert the words to numbers.
- Transform text into something that machines can interpret is called “word embedding”.
- Word embeddings are numerical illustrations of a text.
- Sentence and texts include organized sequences of information.
- Goal is to develop a representation of words that capture their meanings, meaningful connections plus other sorts of situations in which the words are employed.



LLM Randomness

Temperature control the randomness of the output.
It influences the shape of the probability distribution that the model calculates for the next token rather than limiting the token selection

Top-K provides a controlled randomness by considering a fixed number of top probable tokens.



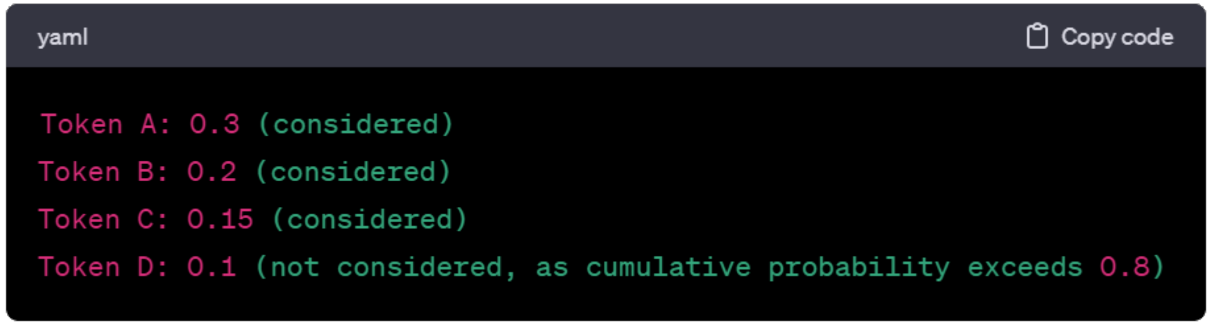
Top-P allows for dynamic control of the number of tokens considered, leading to different levels of diversity in the generated text.

Top-K

- Limit the model’s predictions to the top k most probable tokens at each step of generation.
- Instructing the model to consider only the k most likely tokens, truncating the less likely ones.
- For example, if k=5, the model predicts the probabilities [0.1, 0.3, 0.2, 0.15, 0.05, 0.2], it will only consider the top 5 probabilities, ignoring the rest.
- This helps in generating more focused responses while still allowing for some diversity.

Top-P

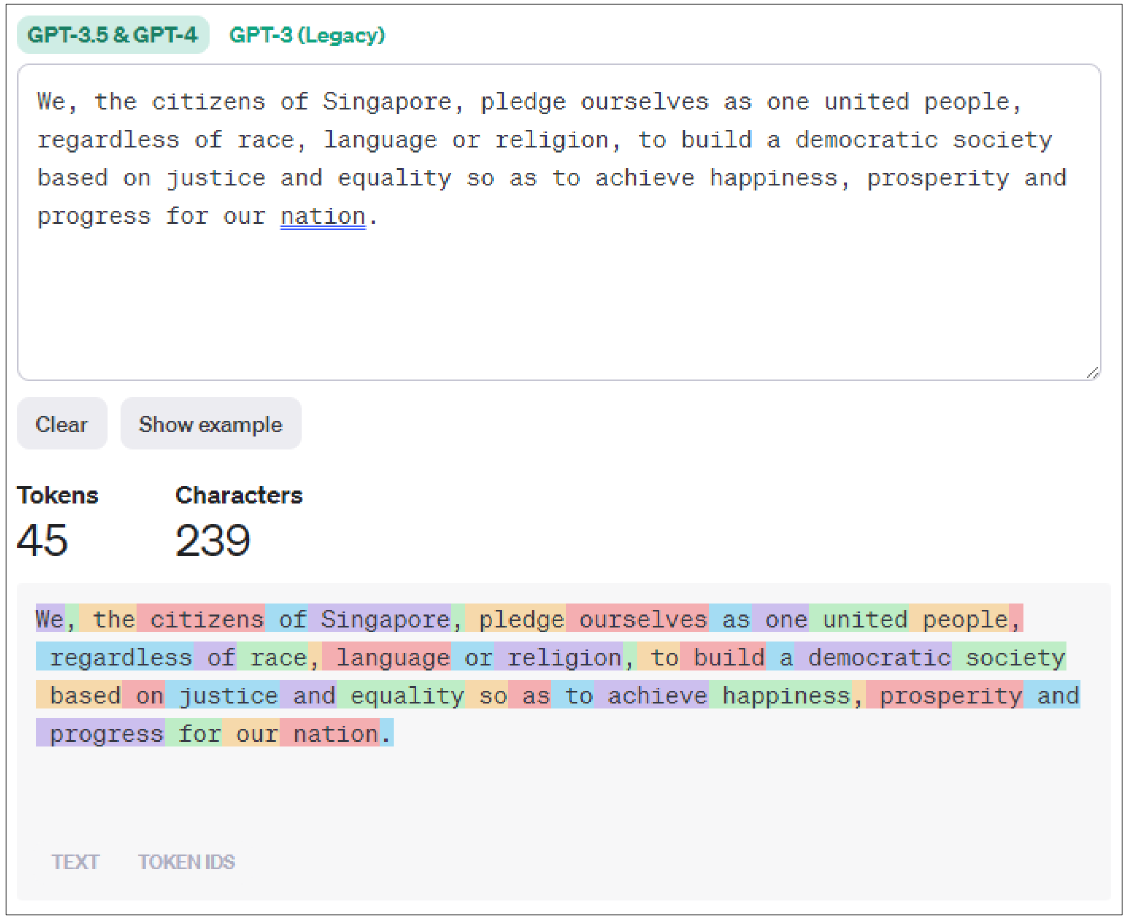
- Also known as nucleus sampling.
- Considers a cumulative probability distribution and includes tokens until the cumulative probability exceeds a certain threshold p.
- Allows dynamic adjustment based on the changing probabilities of tokens.
- Assuming p=0.8, top-P sampling will consider tokens until the cumulative probability exceeds 0.8.



```yaml
Token A: 0.3 (considered)
Token B: 0.2 (considered)
Token C: 0.15 (considered)
Token D: 0.1 (not considered, as cumulative probability exceeds 0.8)
```

Maximum Length

- Maximum length
  - Set the maximum number of tokens of the output.
  - With GPT-3.5, the maximum allowed is 2K (2048) or approximately 1500 words.
  - Token calculator: https://platform.openai.com/tokenizer



We, the citizens of Singapore, pledge ourselves as one united people, regardless of race, language or religion, to build a democratic society based on justice and equality so as to achieve happiness, prosperity and progress for our nation.

Tokens 45

Characters 239

We, the citizens of Singapore, pledge ourselves as one united people, regardless of race, language or religion, to build a democratic society based on justice and equality so as to achieve happiness, prosperity and progress for our nation.

Stop Sequence

- Stop Sequences
  - Tell the model when to stop generating an output.
  - For example: If you want a one-sentence answer, you can use "." as the Stop sequence.
  - Alternatively, for a one-paragraph answer, you can use New Line "\n" as the Stop sequence.
- Useful when trying to generate a dialogue, Q&A or any kind of structured format.
- Most of the time, it is not used frequently.

Frequency and Presence Penalties

- A lot is done under the hood to make sure that it does not just generate the same text again and again.
- One of the ways is by automatically penalizing tokens that is already used.
- API provides two penalty controls.
  - Frequency penalty
    - Penalizes tokens based on how many times the tokens appeared in the text i.e. more time the tokens appear, the more they are penalized
  - Presence penalty
    - Penalizes tokens based on whether the tokens have already appeared in the text.

Chunking & Document Splitting

- Splitting a long document into smaller chunks that can fit into the model's context window.
- Splitting documents into smaller chunks is important and tricky as we need to maintain meaningful relationships between the chunks.
- When you chunk data, overlapping a small amount of text between chunks can help preserve context. We recommend starting with an overlap of approximately 10%.
- For example, given a fixed chunk size of 256 tokens, you would begin testing with an overlap of 25 tokens. The actual amount of overlap varies depending on the type of data and the specific use case, but we have found that 10-15% works for many scenarios[1].

Reference: 1. https://learn.microsoft.com/en-us/azure/search/vector-search-how-to-chunk-documents

Chaining

Prompt Chaining

- One of the important prompt engineering techniques is to break tasks into its subtasks.
- The LLM is prompted with a subtask and then its response is used as input to another prompt, creating a sequence of prompts that lead to the final results.
- Prompt chaining is like assembling a series of building blocks to construct a complete solution.
- Instead of overwhelming the LLM instance with a single detailed prompt, we can use the sub-prompts to guide it through multiple steps, making the process more efficient and transparent.

Prompt Chaining Advantages

- Simplified instructions
  - Write less complicated instruction instead of trying to express a complex task in a single prompt.
  - Break it down into smaller, more straightforward steps.
  - Increase the chances of getting accurate results.
- Help troubleshooting
  - Pinpoint specific prompt in the chain that needs adjustment when LLM has difficulty getting accurate or desired result.
- Incremental validation
  - Check LLM's output in stages rather than wait till the end.
  - Allow to assess the correctness of responses as it progress through the prompt stages.

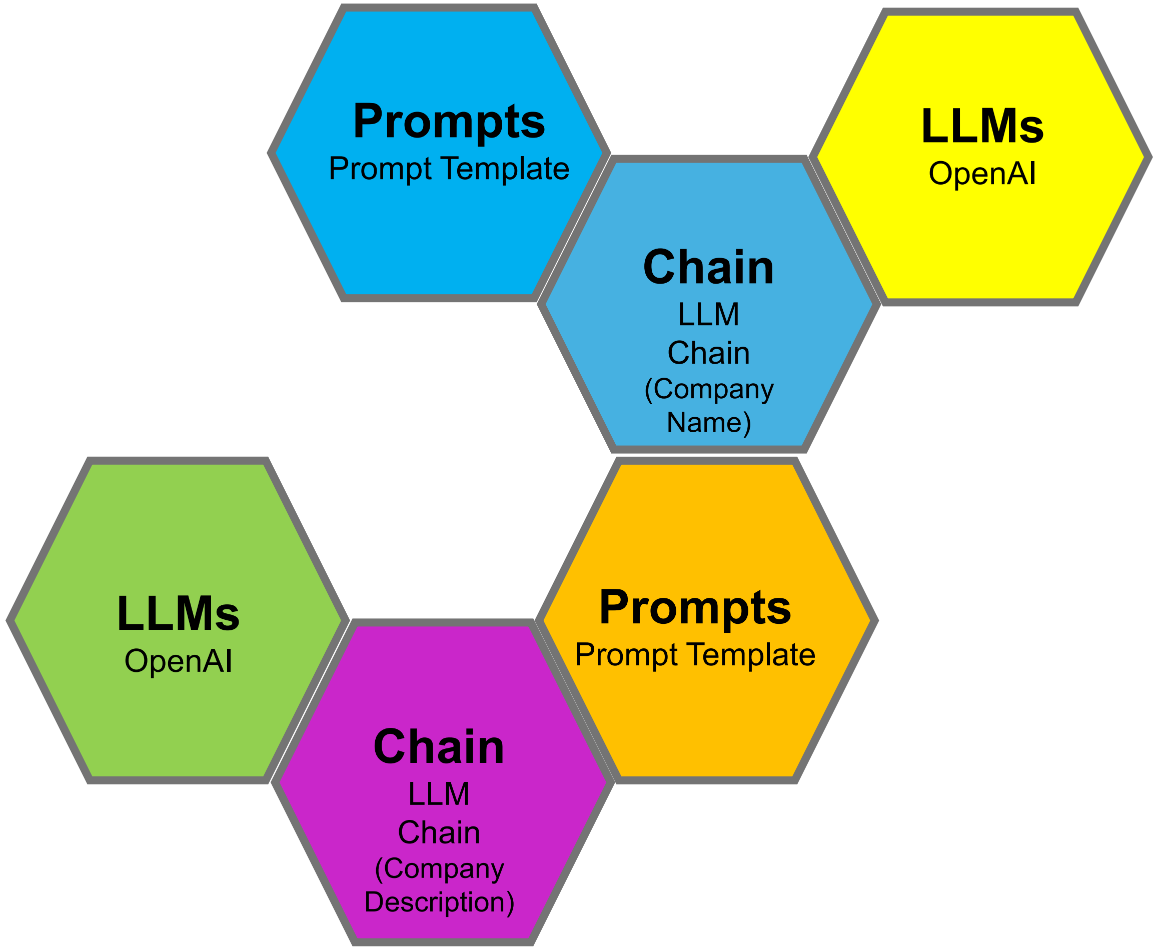
Chain Types

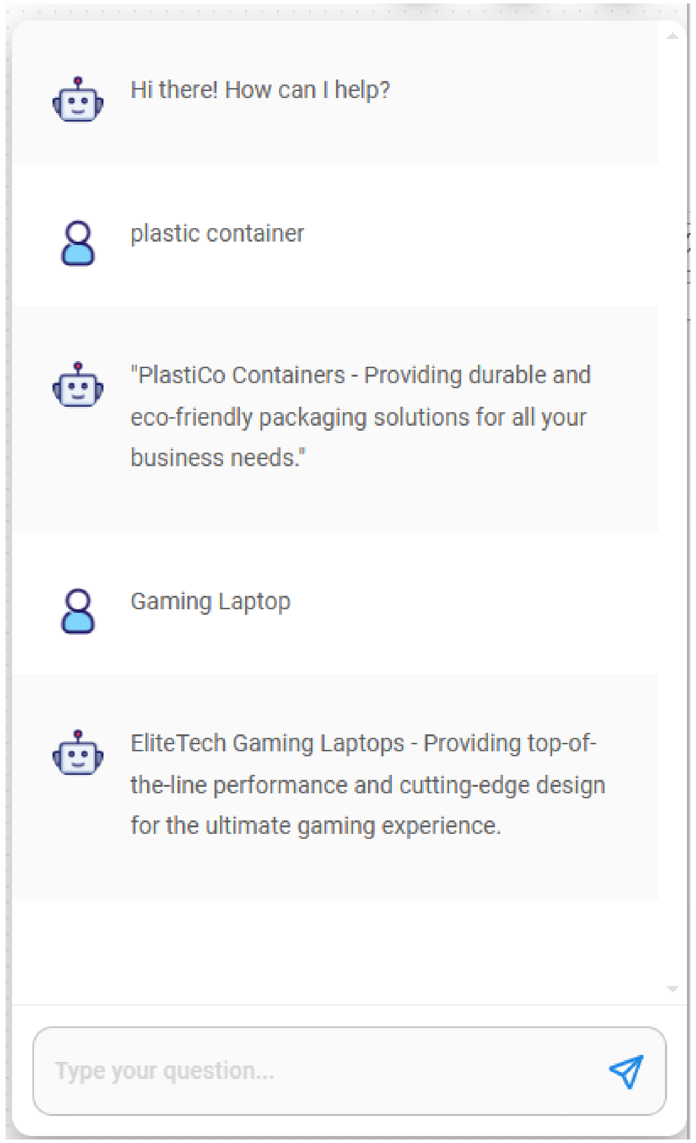
- Simplest Chain
  - User Input ➔ Prompt Template ➔ LLM Model
- Sequential Chain
  - Simple sequential chain: handle a single input and output.
  - Sequential chain: manage multiple inputs and outputs simultaneously.
- Router Chain
  - Allows routing inputs to different destination chains based on the input text.
  - Allows the building of chatbots to handle diverse inputs.

Activity: Sequential Chain

Activity Map (Sequential Chain)

- You will need the following:
  - ☑ OpenAI API Key [OpenAI]




Hi there! How can I help?

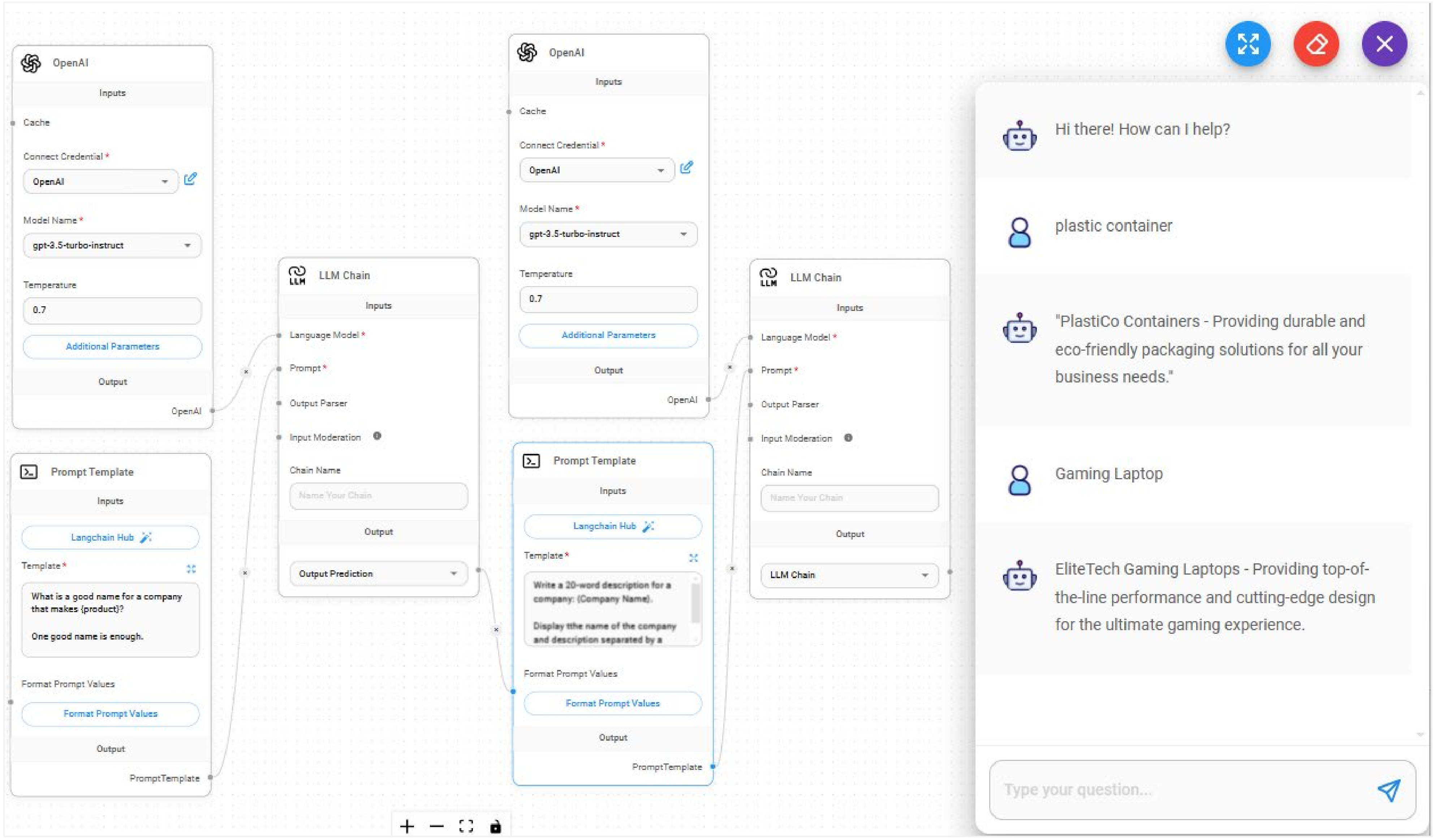
plastic container

"PlastiCo Containers - Providing durable and eco-friendly packaging solutions for all your business needs."

Gaming Laptop

EliteTech Gaming Laptops - Providing top-of-the-line performance and cutting-edge design for the ultimate gaming experience.

Sequential Chain



OpenAI
Inputs
Cache
Connect Credential *: OpenAI
Model Name *: gpt-3.5-turbo-instruct
Temperature: 0.7
Additional Parameters
Output
OpenAI

Prompt Template
Inputs
Langchain Hub
Template *

```
What is a good name for a company that makes {product}?

One good name is enough.
```

Format Prompt Values
Output
PromptTemplate

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

OpenAI
Inputs
Cache
Connect Credential *: OpenAI
Model Name *: gpt-3.5-turbo-instruct
Temperature: 0.7
Additional Parameters
Output
OpenAI

Prompt Template
Inputs
Langchain Hub
Template *

```
Write a 20-word description for a company: {Company Name}.

Display tthe name of the company and description separated by a
```

Format Prompt Values
Output
PromptTemplate

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

Hi there! How can I help?

plastic container

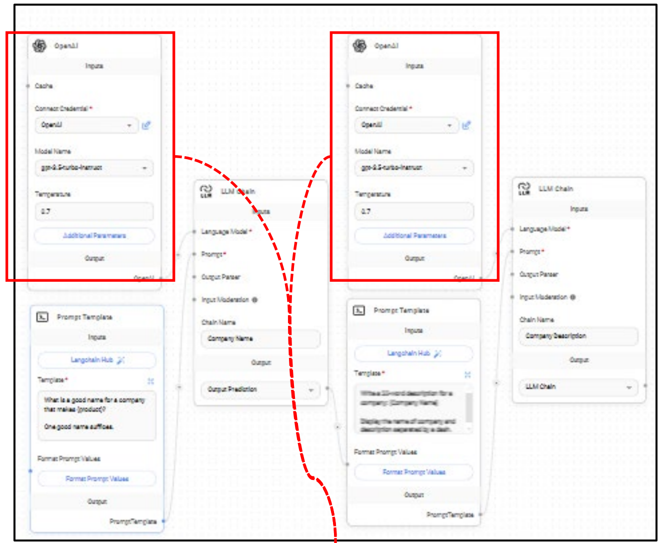
"PlastiCo Containers - Providing durable and eco-friendly packaging solutions for all your business needs."

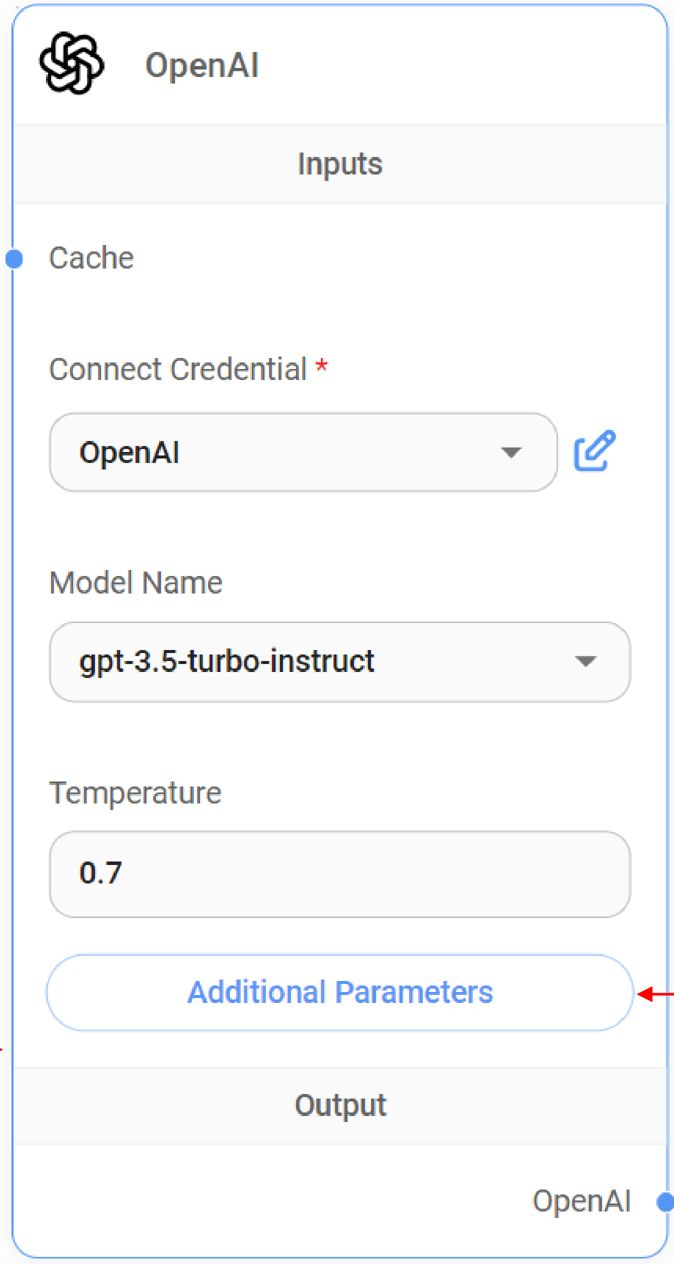
Gaming Laptop

EliteTech Gaming Laptops - Providing top-of-the-line performance and cutting-edge design for the ultimate gaming experience.

Type your question...

Sequential Chain




OpenAI

Inputs

Cache

Connect Credential *

OpenAI

Model Name

gpt-3.5-turbo-instruct

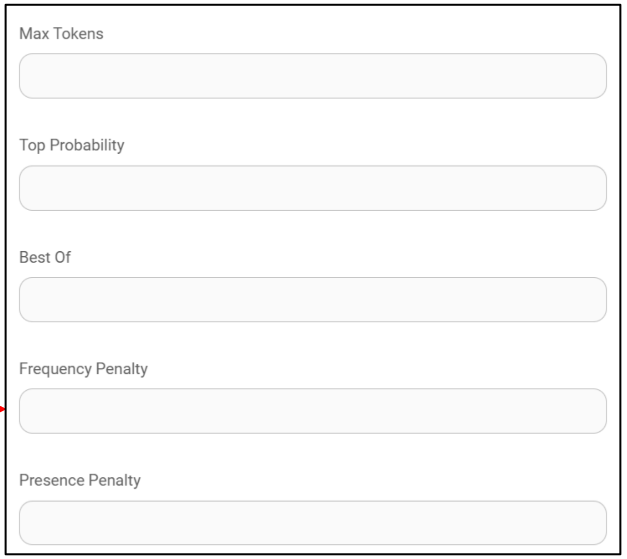
Temperature

0.7

Additional Parameters

Output

OpenAI


Max Tokens

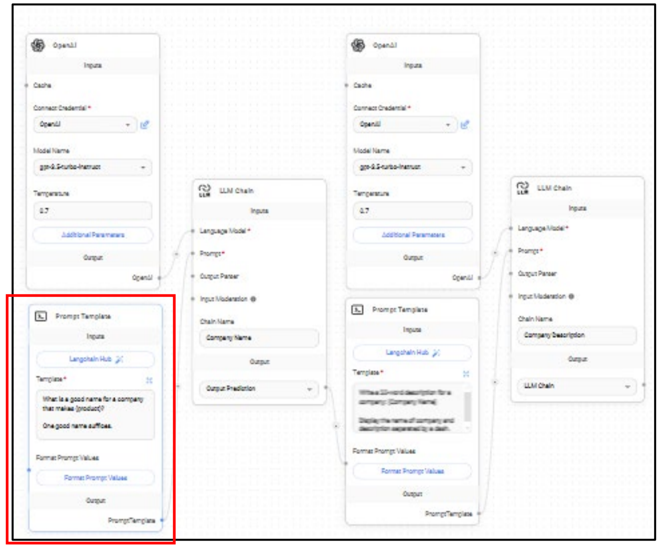
Top Probability

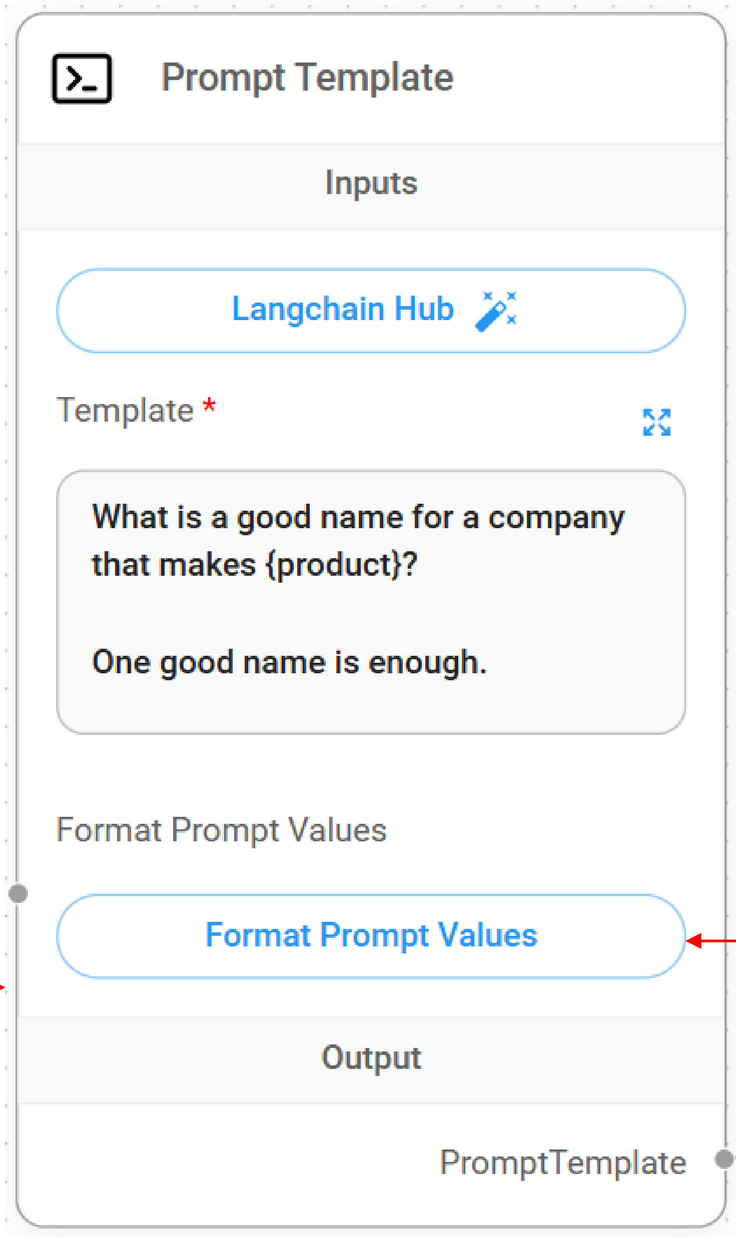
Best Of

Frequency Penalty

Presence Penalty

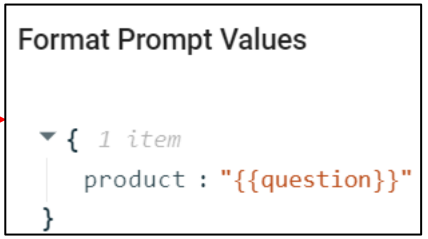
Sequential Chain





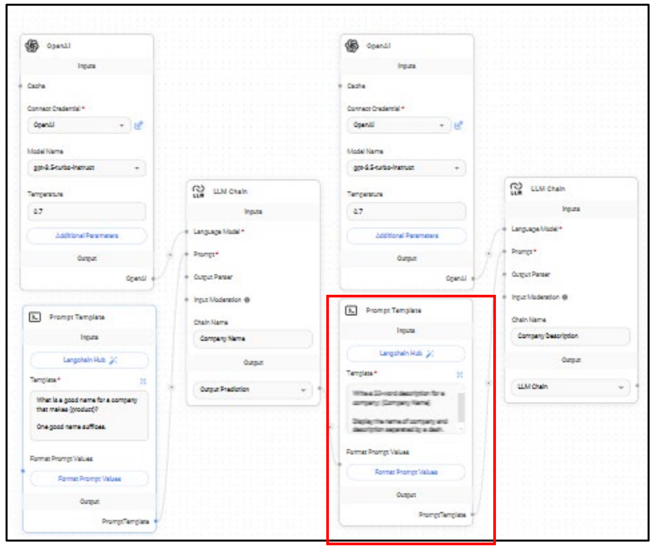
What is a good name for a company that makes {product}?

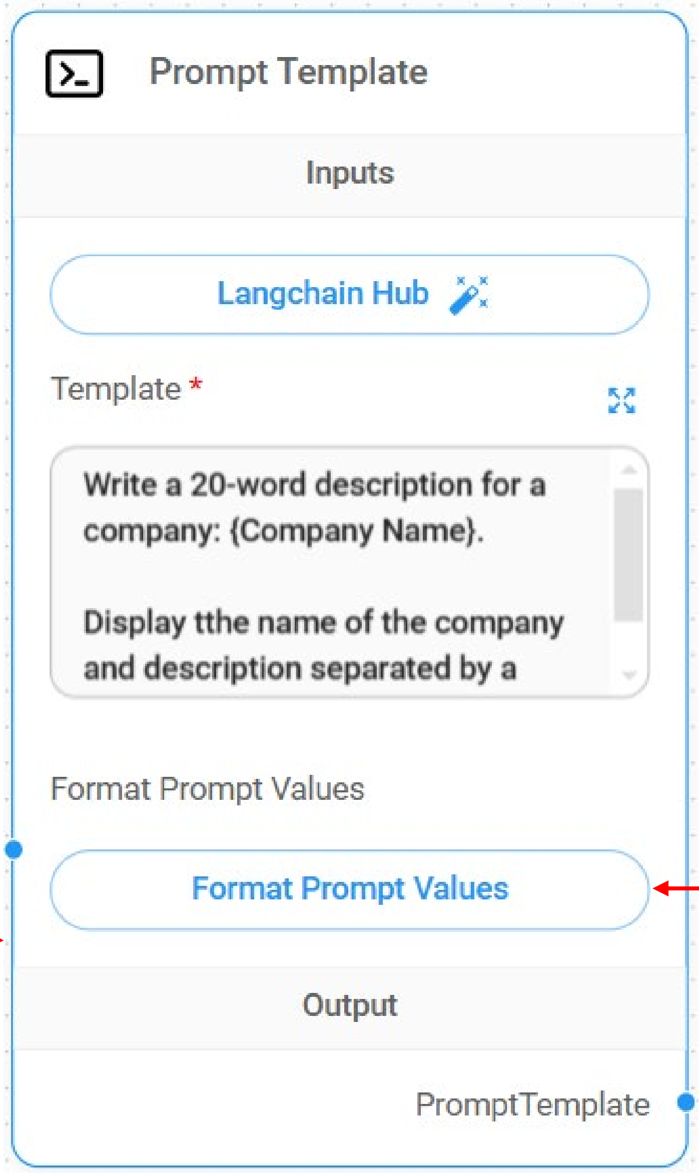
One good name is enough.


Format Prompt Values

```
▾ { 1 item
    product : "{{question}}"
  }
```

Sequential Chain




Prompt Template

Inputs

Langchain Hub

Template *

Write a 20-word description for a company: {Company Name}.

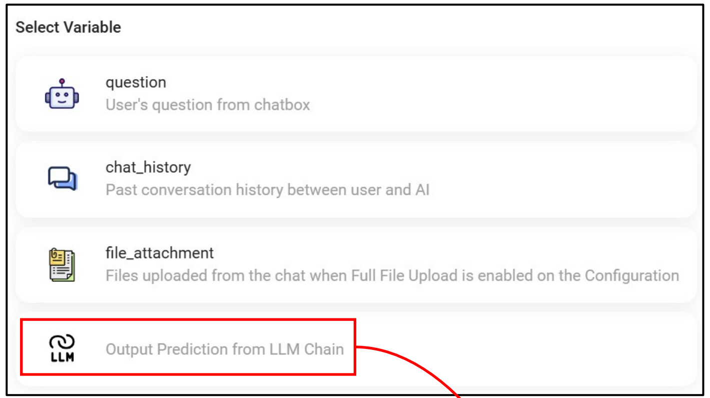
Display tthe name of the company and description separated by a

Format Prompt Values

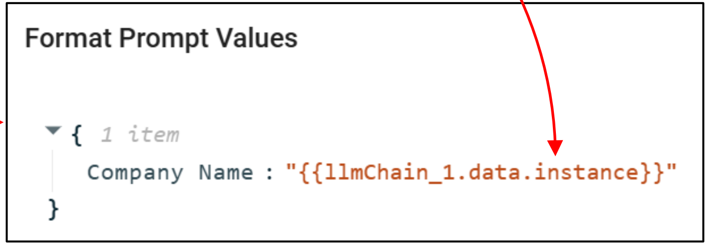
Format Prompt Values

Output

PromptTemplate


Select Variable

question	User's question from chatbox
chat_history	Past conversation history between user and AI
file_attachment	Files uploaded from the chat when Full File Upload is enabled on the Configuration
	Output Prediction from LLM Chain


Format Prompt Values

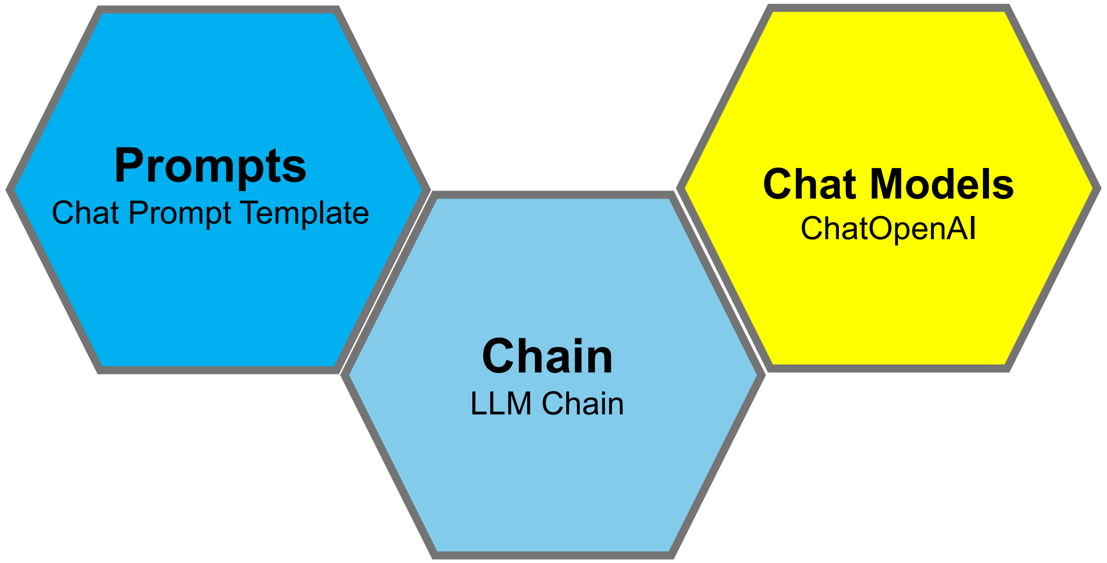
```
{ 1 item
  Company Name : "{{llmChain_1.data.instance}}"
}
```

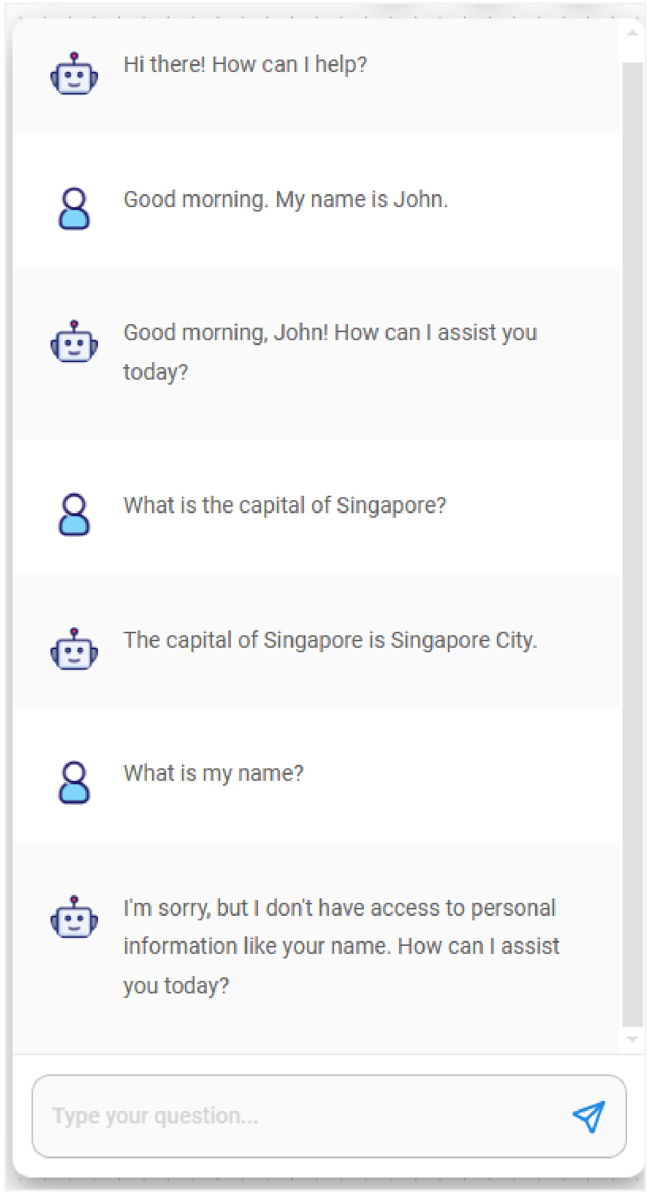
Reminder: Save your work before testing

Activity: Memory

Activity Map (Forgetful Assistant)

- You will need the following:
  - ☑ OpenAI API Key [OpenAI]





Hi there! How can I help?

Good morning. My name is John.

Good morning, John! How can I assist you today?

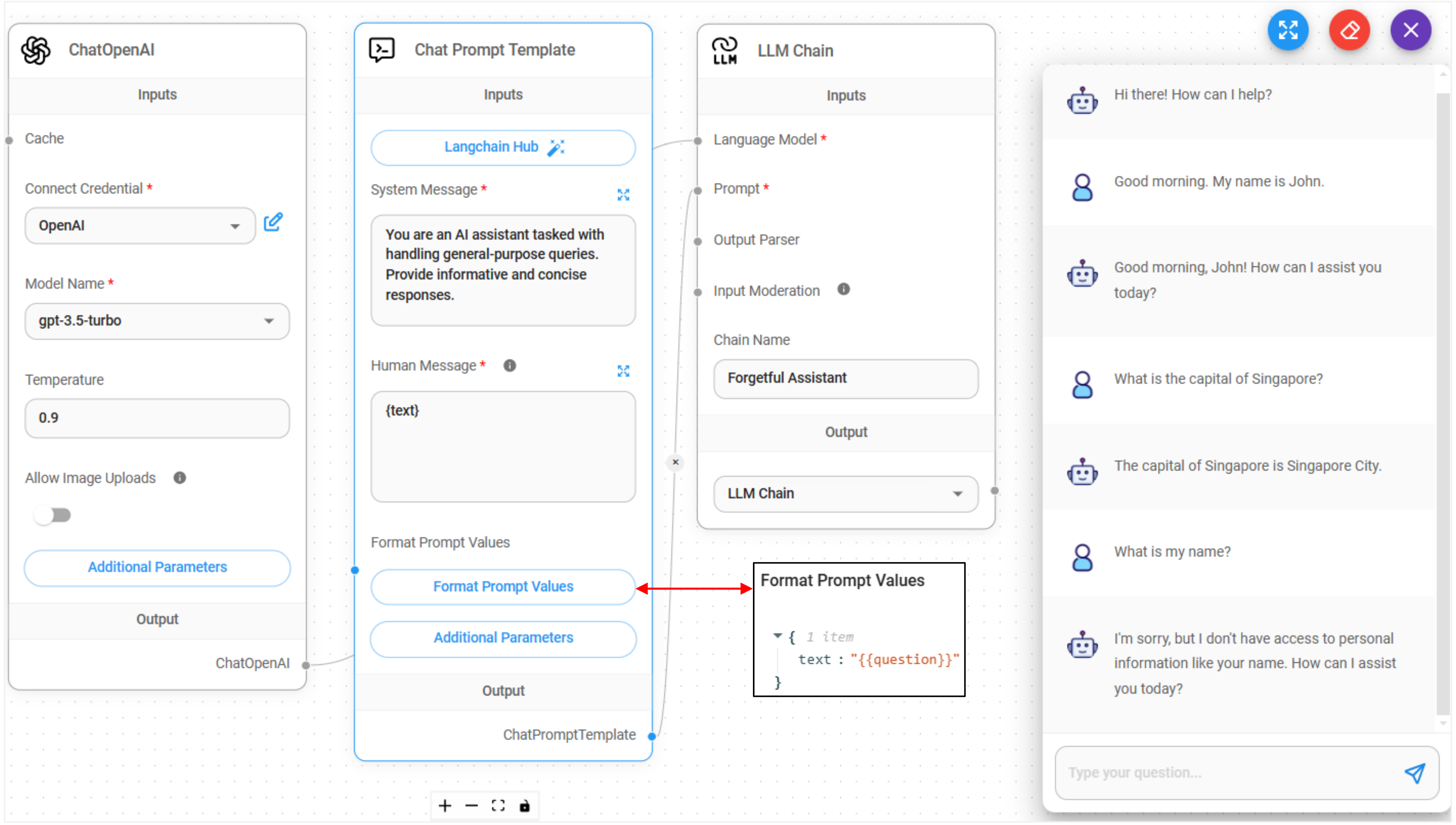
What is the capital of Singapore?

The capital of Singapore is Singapore City.

What is my name?

I'm sorry, but I don't have access to personal information like your name. How can I assist you today?

Forgetful Assistant



ChatOpenAI
Inputs
Cache
Connect Credential *
OpenAI
Model Name *
gpt-3.5-turbo
Temperature
0.9
Allow Image Uploads
Additional Parameters
Output
ChatOpenAI

Chat Prompt Template
Inputs
Langchain Hub
System Message *
You are an AI assistant tasked with handling general-purpose queries. Provide informative and concise responses.
Human Message *
{text}
Format Prompt Values
Format Prompt Values
Additional Parameters
Output
ChatPromptTemplate

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

Format Prompt Values

```
▾ { 1 item
  text : "{{question}}"
}
```

Hi there! How can I help?

Good morning. My name is John.

Good morning, John! How can I assist you today?

What is the capital of Singapore?

The capital of Singapore is Singapore City.

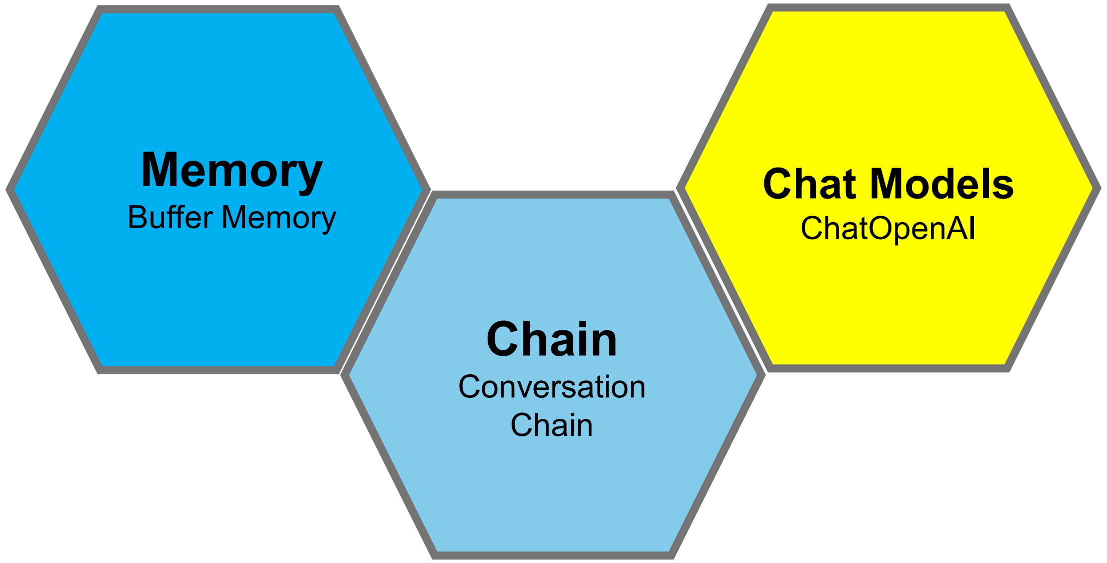
What is my name?

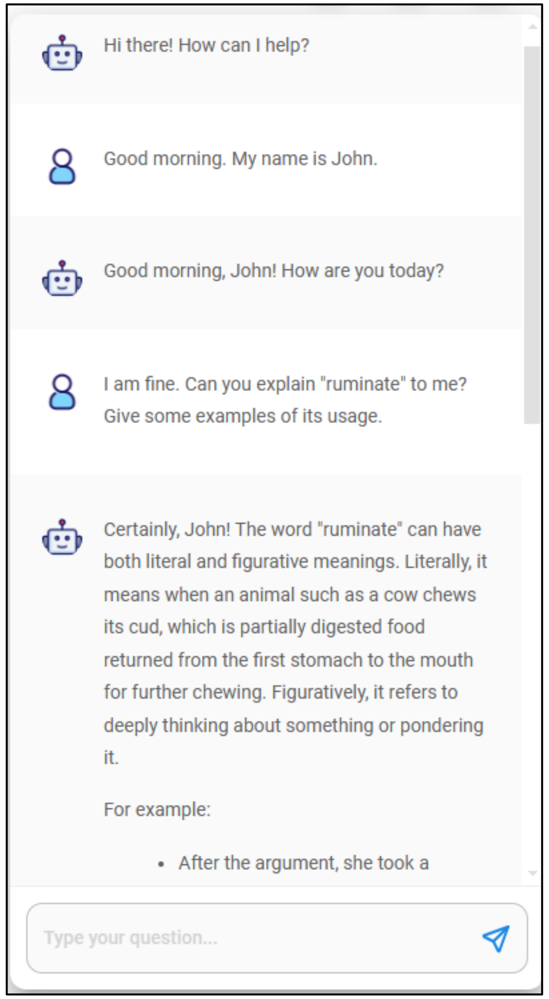
I'm sorry, but I don't have access to personal information like your name. How can I assist you today?

Type your question...

Activity Map (Attentive Assistant)

- You will need the following:
  - ☑ OpenAI API Key [OpenAI]





Hi there! How can I help?

Good morning. My name is John.

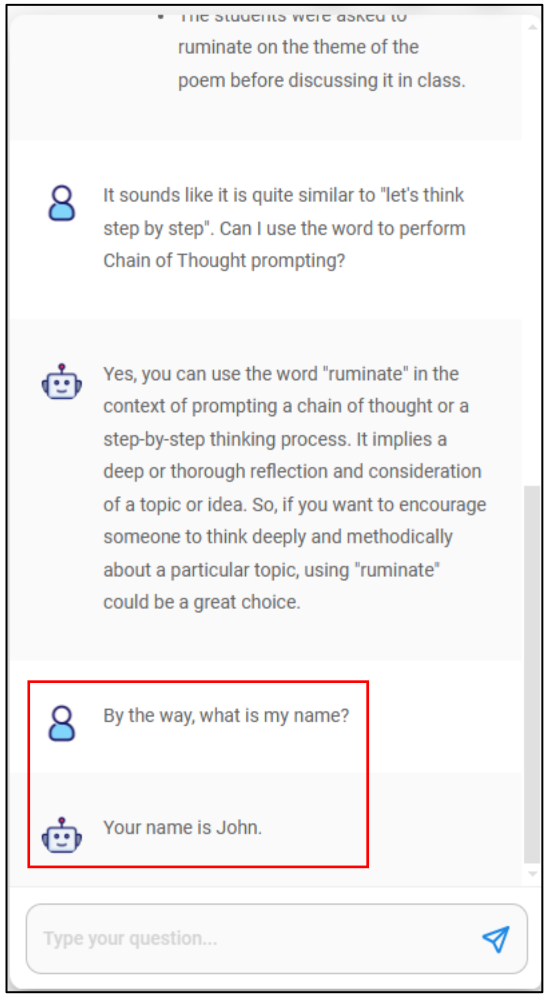
Good morning, John! How are you today?

I am fine. Can you explain "ruminate" to me? Give some examples of its usage.

Certainly, John! The word "ruminate" can have both literal and figurative meanings. Literally, it means when an animal such as a cow chews its cud, which is partially digested food returned from the first stomach to the mouth for further chewing. Figuratively, it refers to deeply thinking about something or pondering it.

For example:

- After the argument, she took a



ruminate on the theme of the poem before discussing it in class.

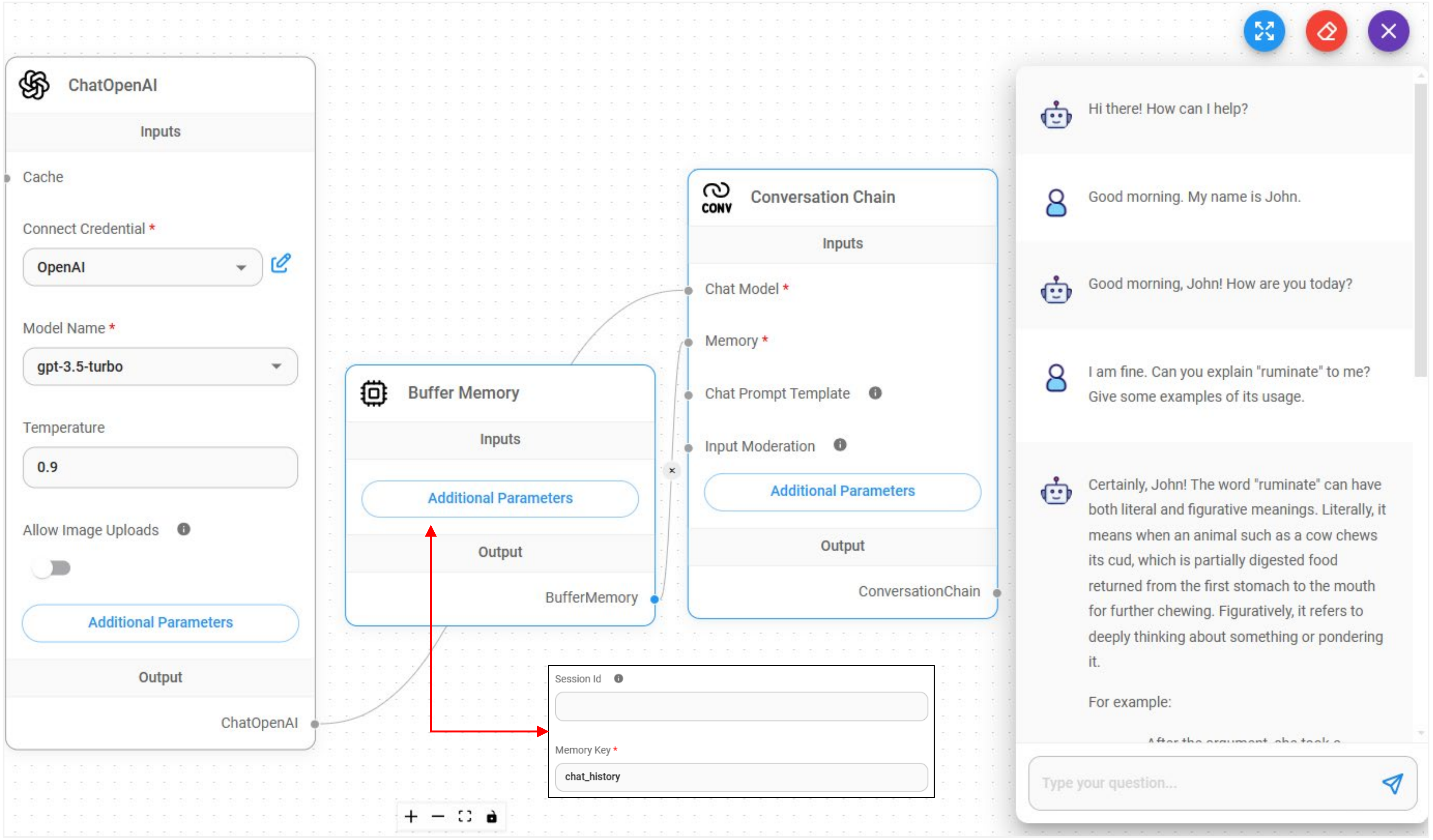
It sounds like it is quite similar to "let's think step by step". Can I use the word to perform Chain of Thought prompting?

Yes, you can use the word "ruminate" in the context of prompting a chain of thought or a step-by-step thinking process. It implies a deep or thorough reflection and consideration of a topic or idea. So, if you want to encourage someone to think deeply and methodically about a particular topic, using "ruminate" could be a great choice.

By the way, what is my name?

Your name is John.

Attentive Assistant



Hi there! How can I help?

Good morning. My name is John.

Good morning, John! How are you today?

I am fine. Can you explain "ruminate" to me? Give some examples of its usage.

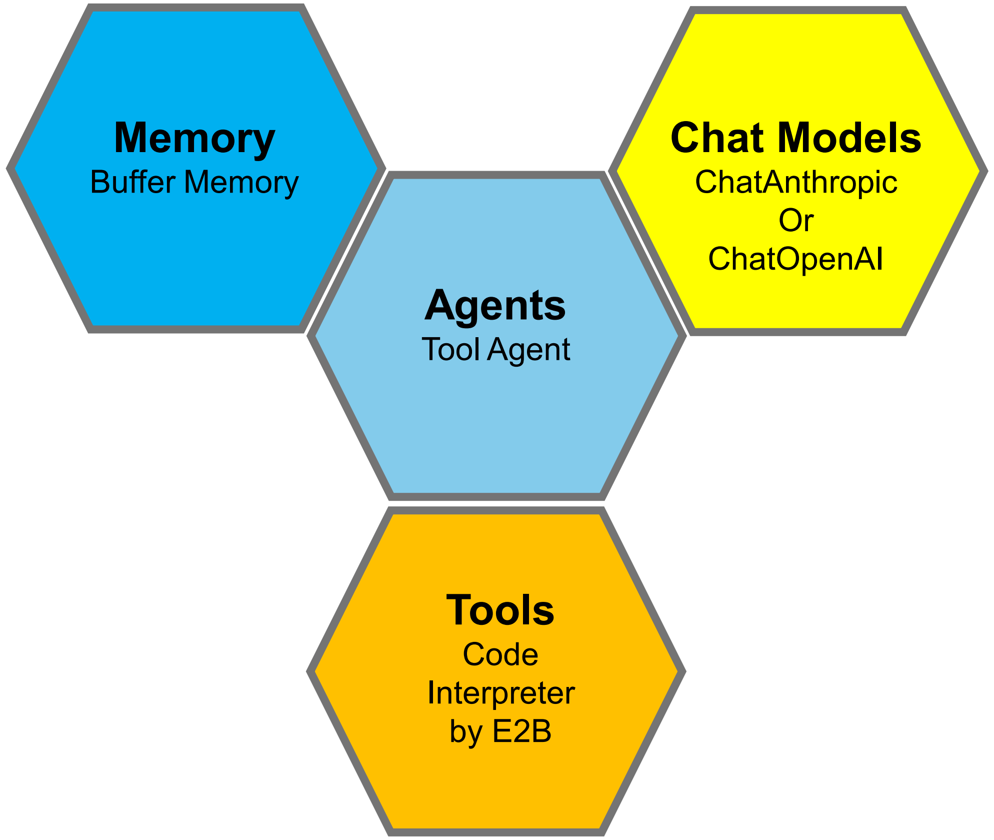
Certainly, John! The word "ruminate" can have both literal and figurative meanings. Literally, it means when an animal such as a cow chews its cud, which is partially digested food returned from the first stomach to the mouth for further chewing. Figuratively, it refers to deeply thinking about something or pondering it.

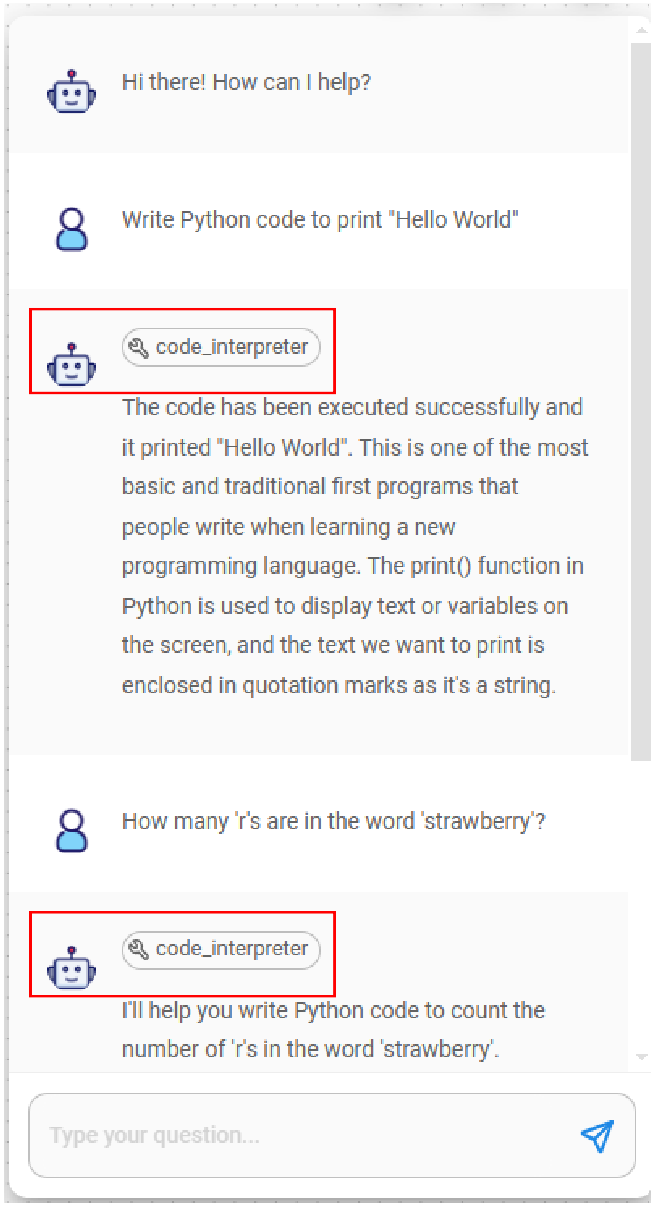
For example:

Activity: Agents

Activity Map (Tool Agent)

- You will need the following:
  - ☑ OpenAI API Key [OpenAI]




Hi there! How can I help?

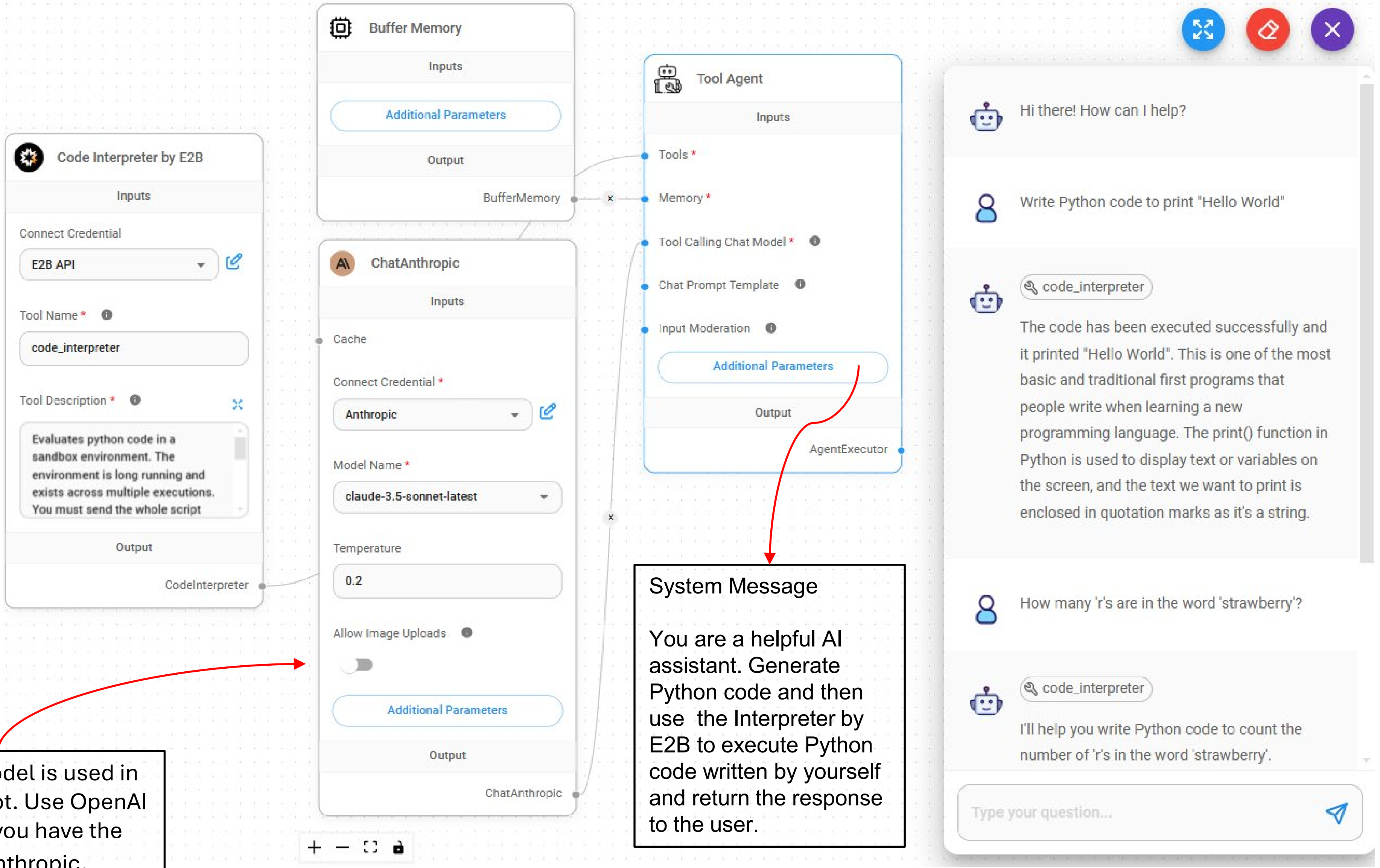
Write Python code to print "Hello World"

The code has been executed successfully and it printed "Hello World". This is one of the most basic and traditional first programs that people write when learning a new programming language. The print() function in Python is used to display text or variables on the screen, and the text we want to print is enclosed in quotation marks as it's a string.

How many 'r's are in the word 'strawberry'?

I'll help you write Python code to count the number of 'r's in the word 'strawberry'.

Tool Agent



Code Interpreter by E2B

Inputs

Connect Credential: E2B API

Tool Name *: code_interpreter

Tool Description *: Evaluates python code in a sandbox environment. The environment is long running and exists across multiple executions. You must send the whole script

Output: CodeInterpreter

Buffer Memory

Inputs

Additional Parameters

Output: BufferMemory

ChatAnthropic

Inputs

Cache

Connect Credential *: Anthropic

Model Name *: claude-3.5-sonnet-latest

Temperature: 0.2

Allow Image Uploads

Additional Parameters

Output: ChatAnthropic

Tool Agent

Inputs

Tools *

Memory *

Tool Calling Chat Model *

Chat Prompt Template

Input Moderation

Additional Parameters

Output: AgentExecutor

Hi there! How can I help?

Write Python code to print "Hello World"

code_interpreter

The code has been executed successfully and it printed "Hello World". This is one of the most basic and traditional first programs that people write when learning a new programming language. The print() function in Python is used to display text or variables on the screen, and the text we want to print is enclosed in quotation marks as it's a string.

How many 'r's are in the word 'strawberry'?

code_interpreter

I'll help you write Python code to count the number of 'r's in the word 'strawberry'.

Type your question...

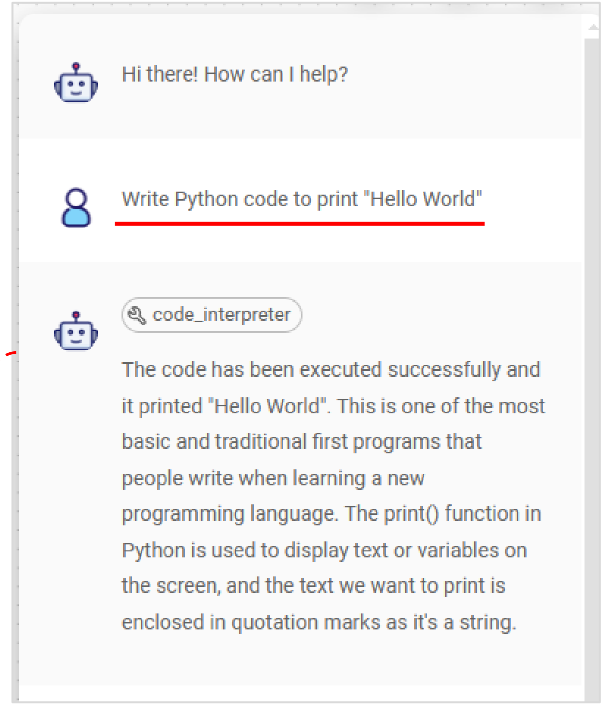
System Message

You are a helpful AI assistant. Generate Python code and then use the Interpreter by E2B to execute Python code written by yourself and return the response to the user.

* Anthropic model is used in this screen shot. Use OpenAI model unless you have the API key from Anthropic.

Tool Agent Observations

❶

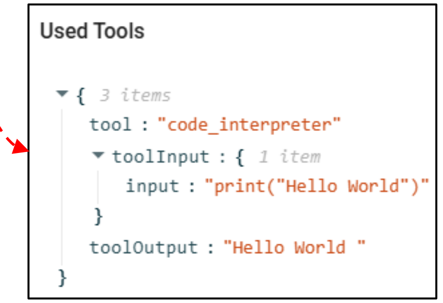


Hi there! How can I help?

Write Python code to print "Hello World"

code_interpreter

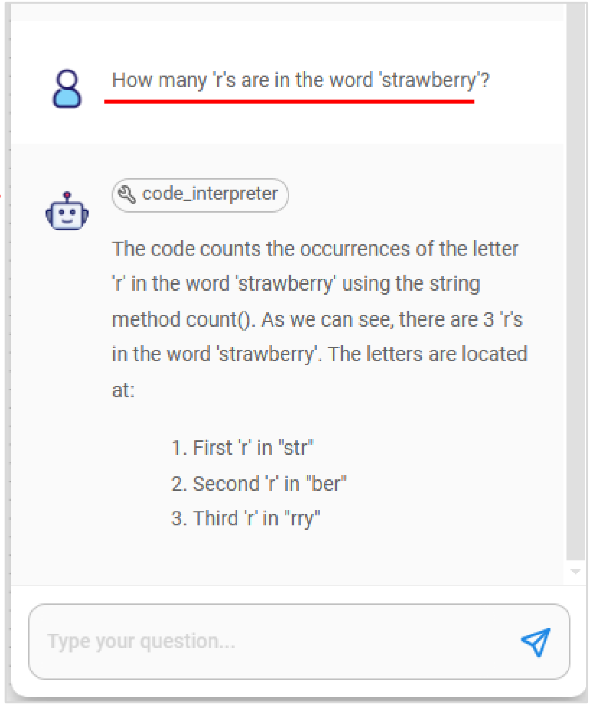
The code has been executed successfully and it printed "Hello World". This is one of the most basic and traditional first programs that people write when learning a new programming language. The print() function in Python is used to display text or variables on the screen, and the text we want to print is enclosed in quotation marks as it's a string.



Used Tools

```
▾ { 3 items
   tool : "code_interpreter"
  ▾ toolInput : { 1 item
     input : "print("Hello World")"
   }
   toolOutput : "Hello World "
}
```

❷



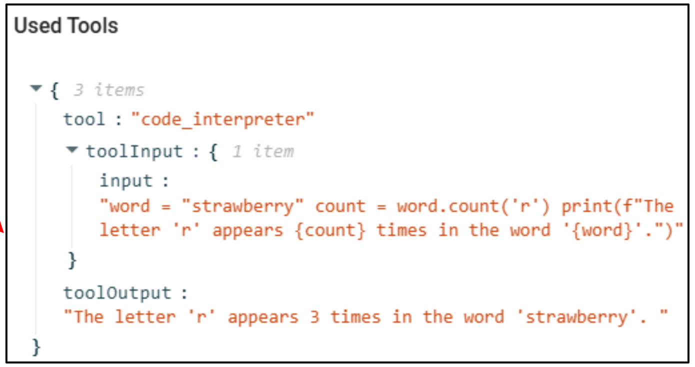
How many 'r's are in the word 'strawberry'?

code_interpreter

The code counts the occurrences of the letter 'r' in the word 'strawberry' using the string method count(). As we can see, there are 3 'r's in the word 'strawberry'. The letters are located at:

1. First 'r' in "str"
2. Second 'r' in "ber"
3. Third 'r' in "rry"

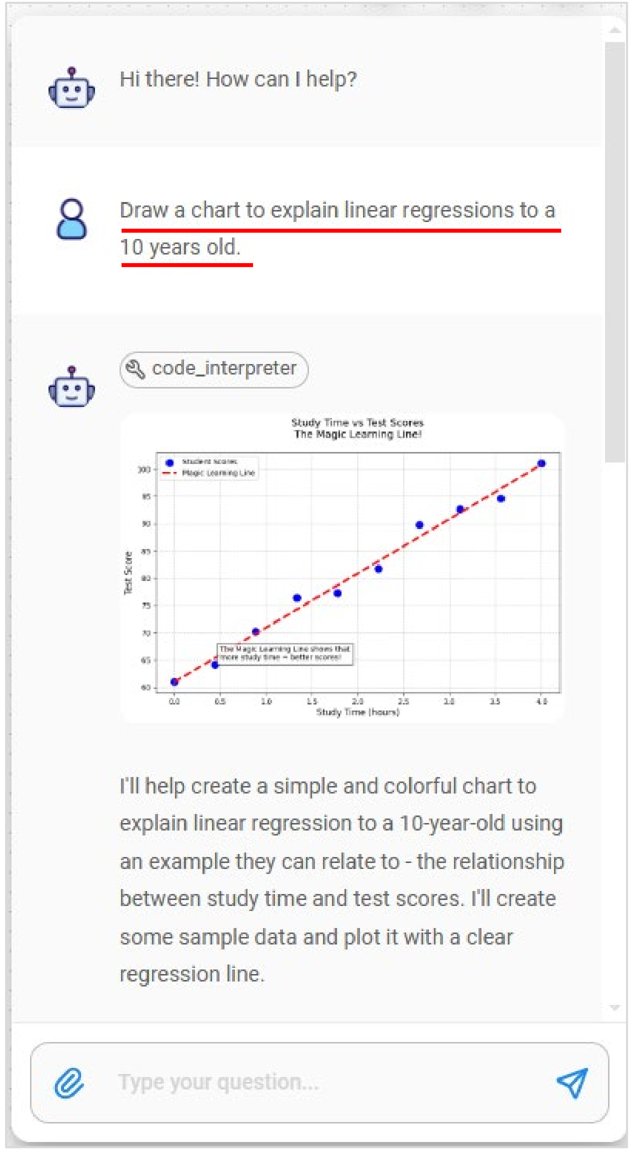
Type your question...



Used Tools

```
▾ { 3 items
   tool : "code_interpreter"
  ▾ toolInput : { 1 item
     input :
     "word = "strawberry" count = word.count('r') print(f"The
     letter 'r' appears {count} times in the word '{word}'.")"
   }
   toolOutput :
   "The letter 'r' appears 3 times in the word 'strawberry'. "
}
```

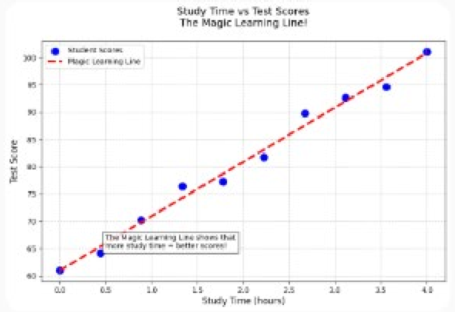
❸



Hi there! How can I help?

Draw a chart to explain linear regressions to a 10 years old.

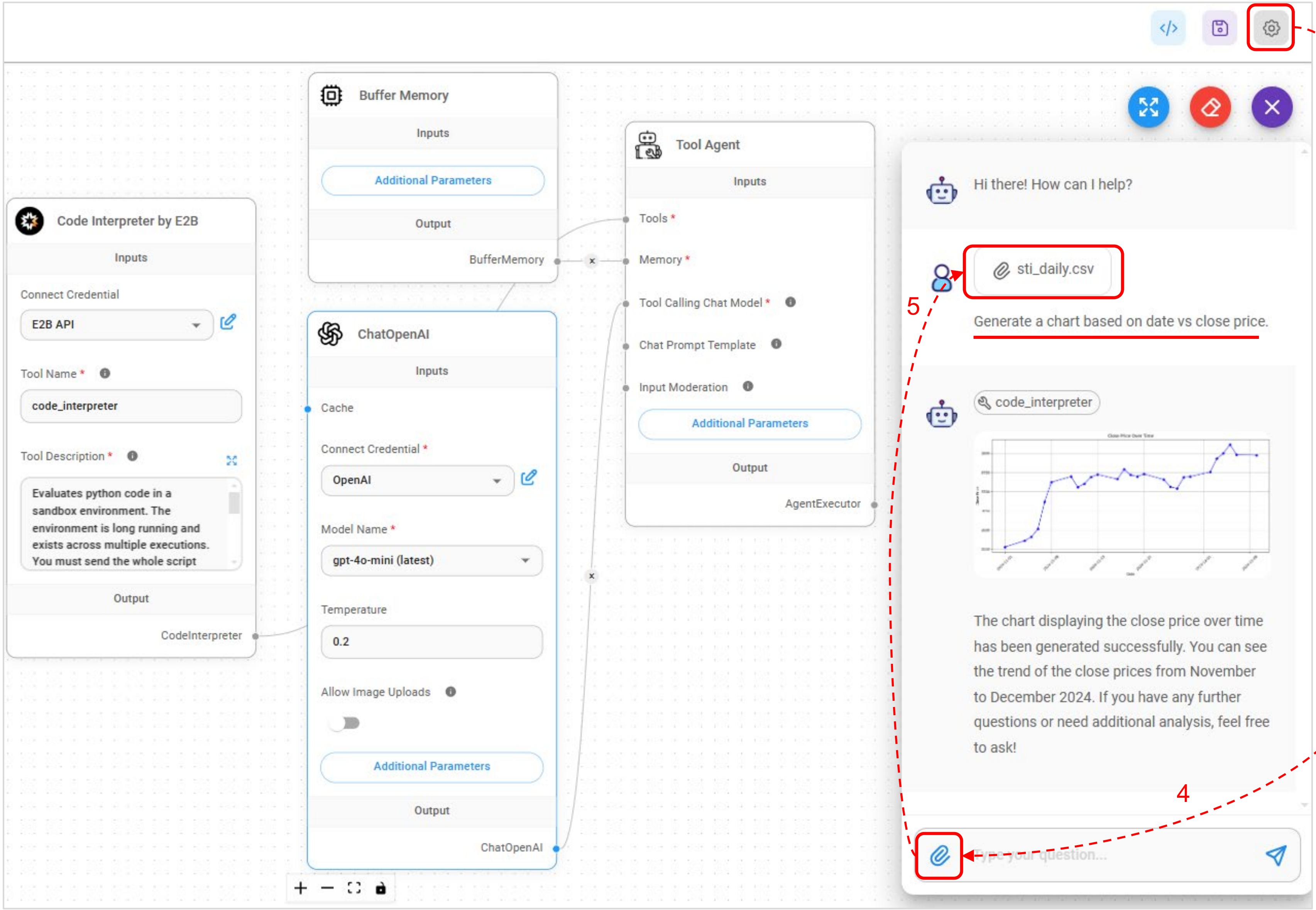
code_interpreter



I'll help create a simple and colorful chart to explain linear regression to a 10-year-old using an example they can relate to - the relationship between study time and test scores. I'll create some sample data and plot it with a clear regression line.

Type your question...

Tool Agent Observation

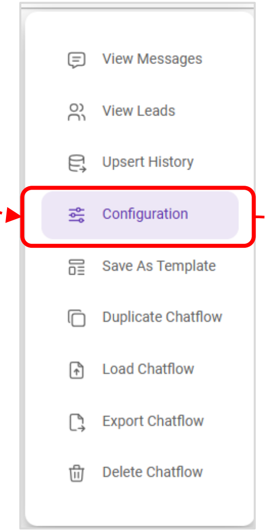


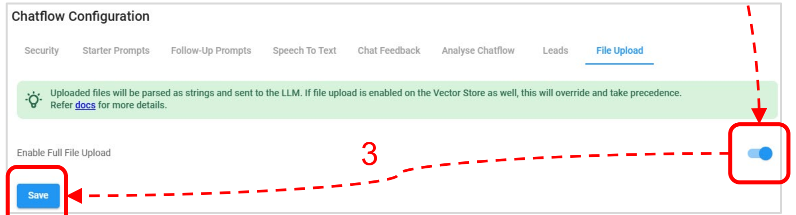
Evaluates python code in a sandbox environment. The environment is long running and exists across multiple executions. You must send the whole script

Hi there! How can I help?

Generate a chart based on date vs close price.

The chart displaying the close price over time has been generated successfully. You can see the trend of the close prices from November to December 2024. If you have any further questions or need additional analysis, feel free to ask!





Uploaded files will be parsed as strings and sent to the LLM. If file upload is enabled on the Vector Store as well, this will override and take precedence. Refer docs for more details.

Note: To enable “paper clip”, go to Configuration under Settings

Reference

- Flowise
  https://docs.flowiseai.com/

Thank you!