build_chat()

Primary Goal

Builds and returns a Retrieval Chain to answer a user's question

The chain should support streaming text generation

The chain should use random variations of component parts

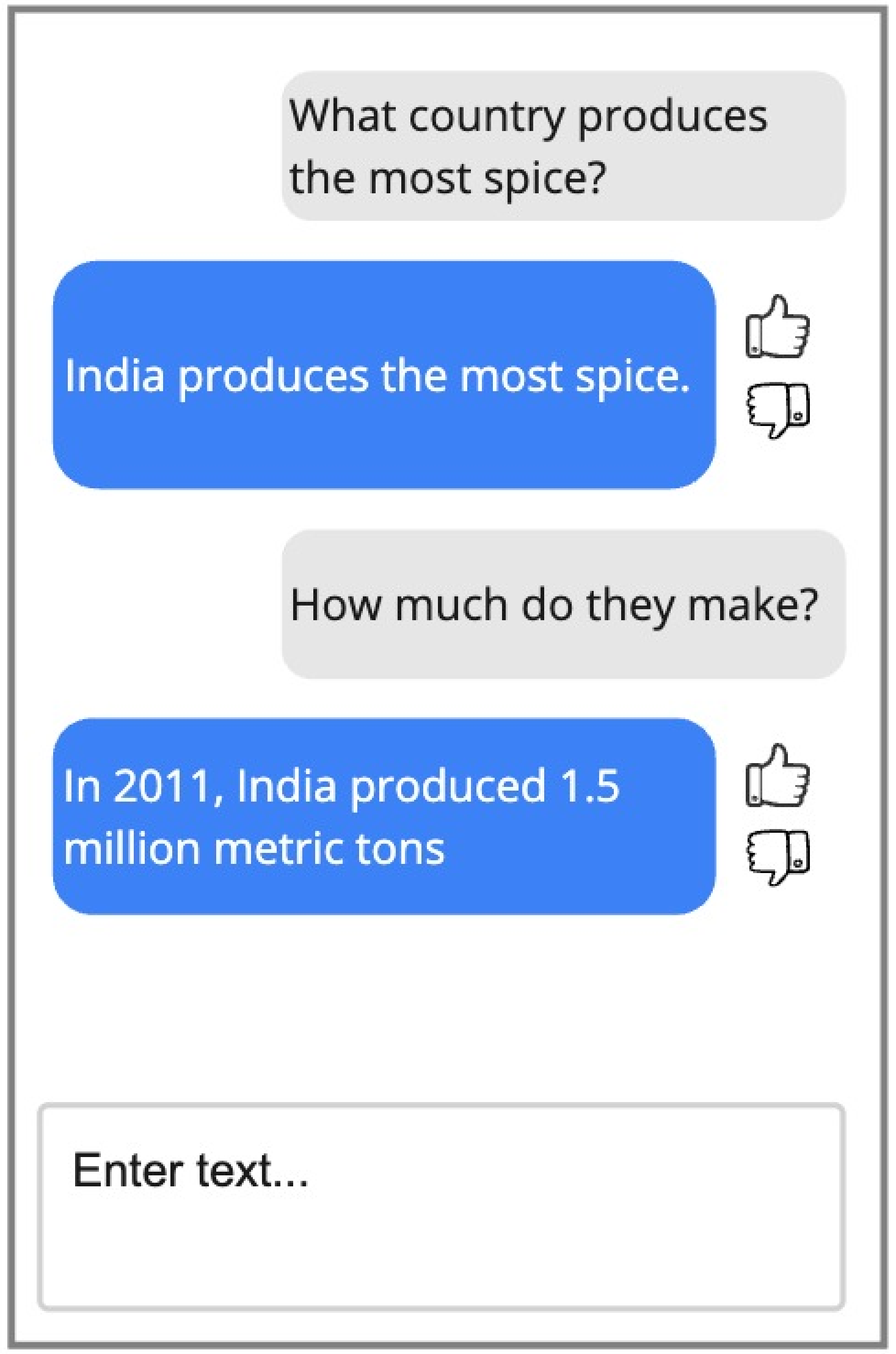
RetrievalChain

The chain we made on our previous app

ConversationalRetrievalChain

The chain we're going to make for this PDF chat app

Chat Panel



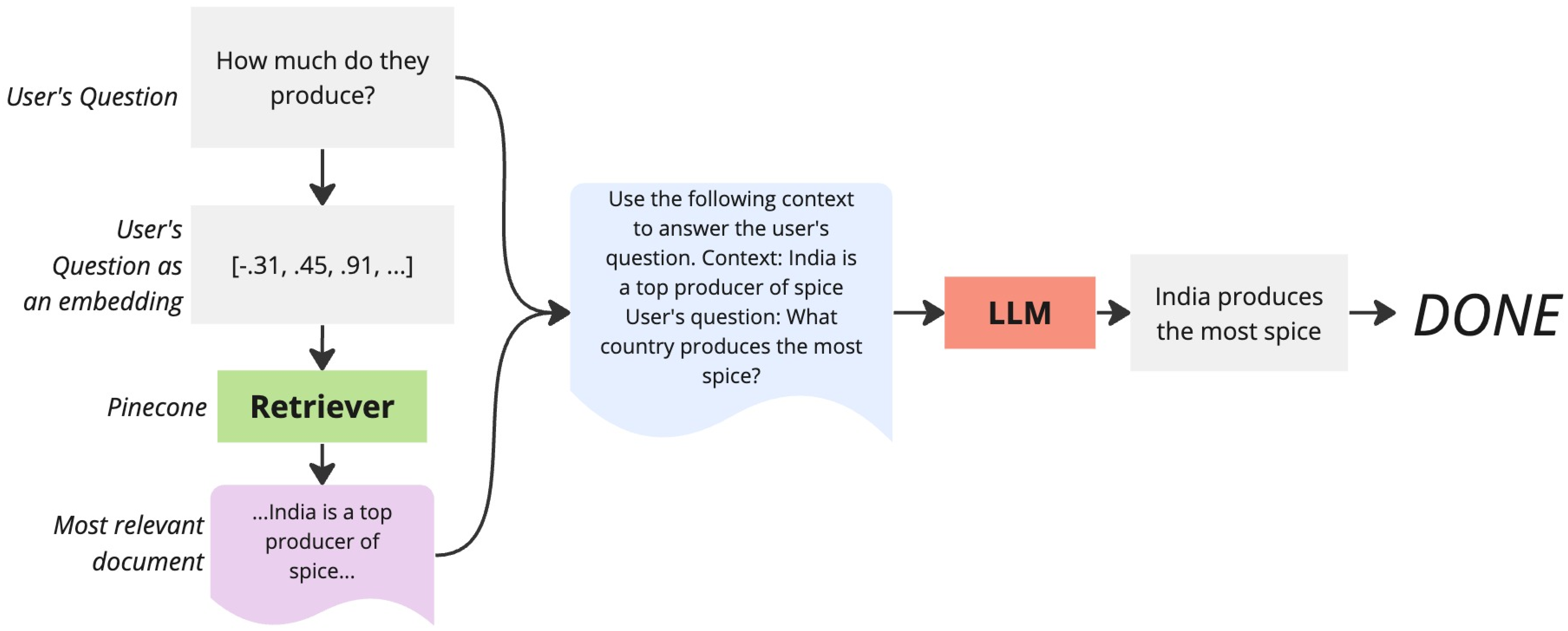
What country produces the most spice?

India produces the most spice.

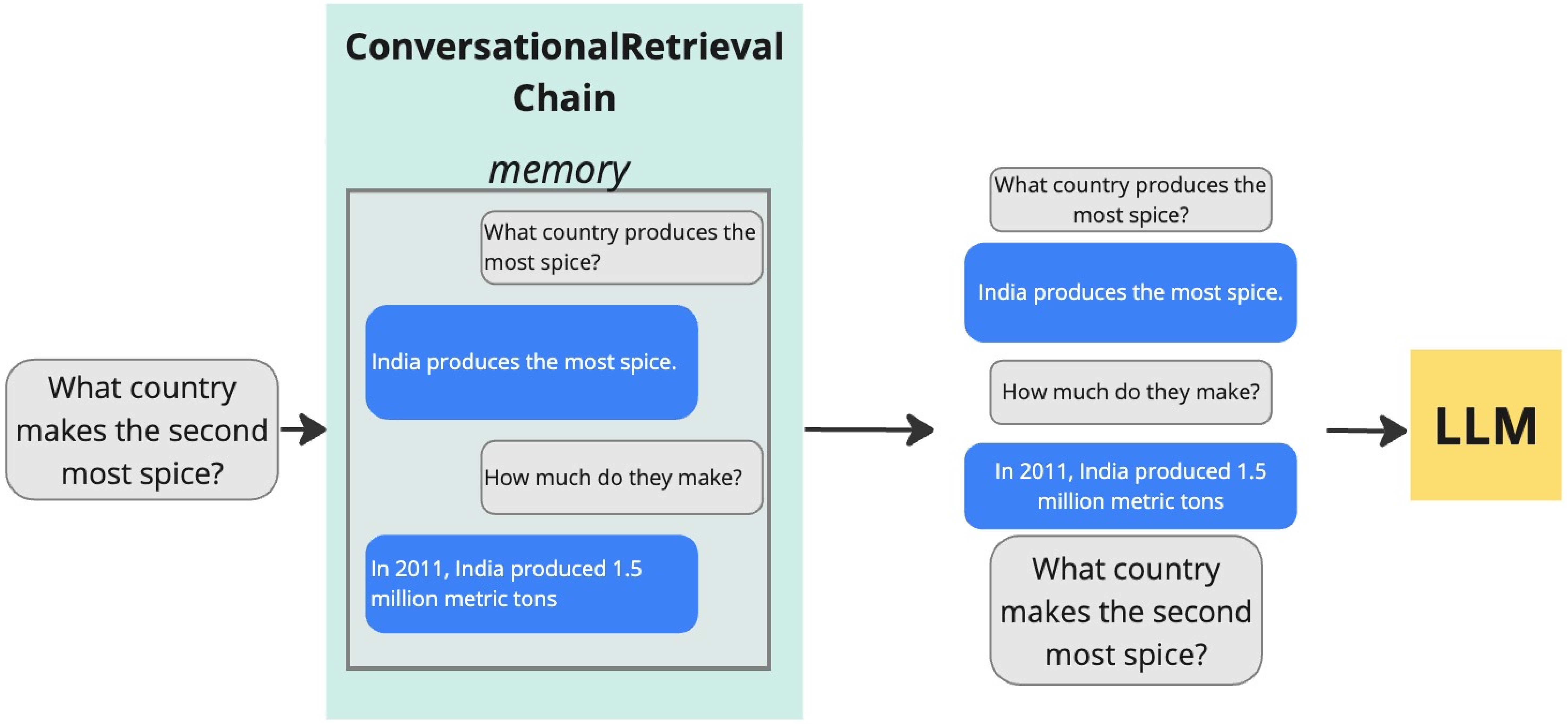
How much do they make?

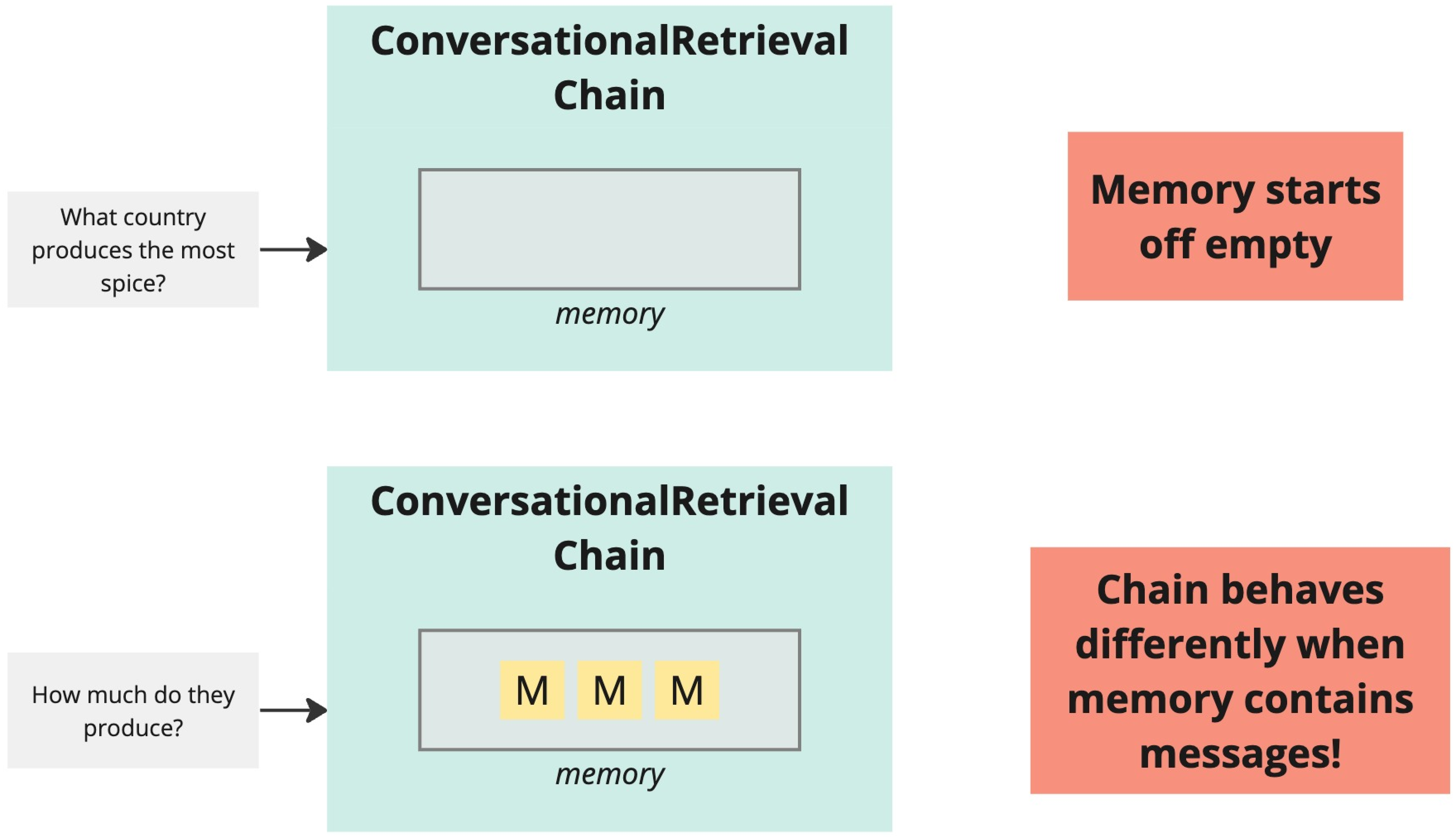
In 2011, India produced 1.5 million metric tons

Retrieval Chain

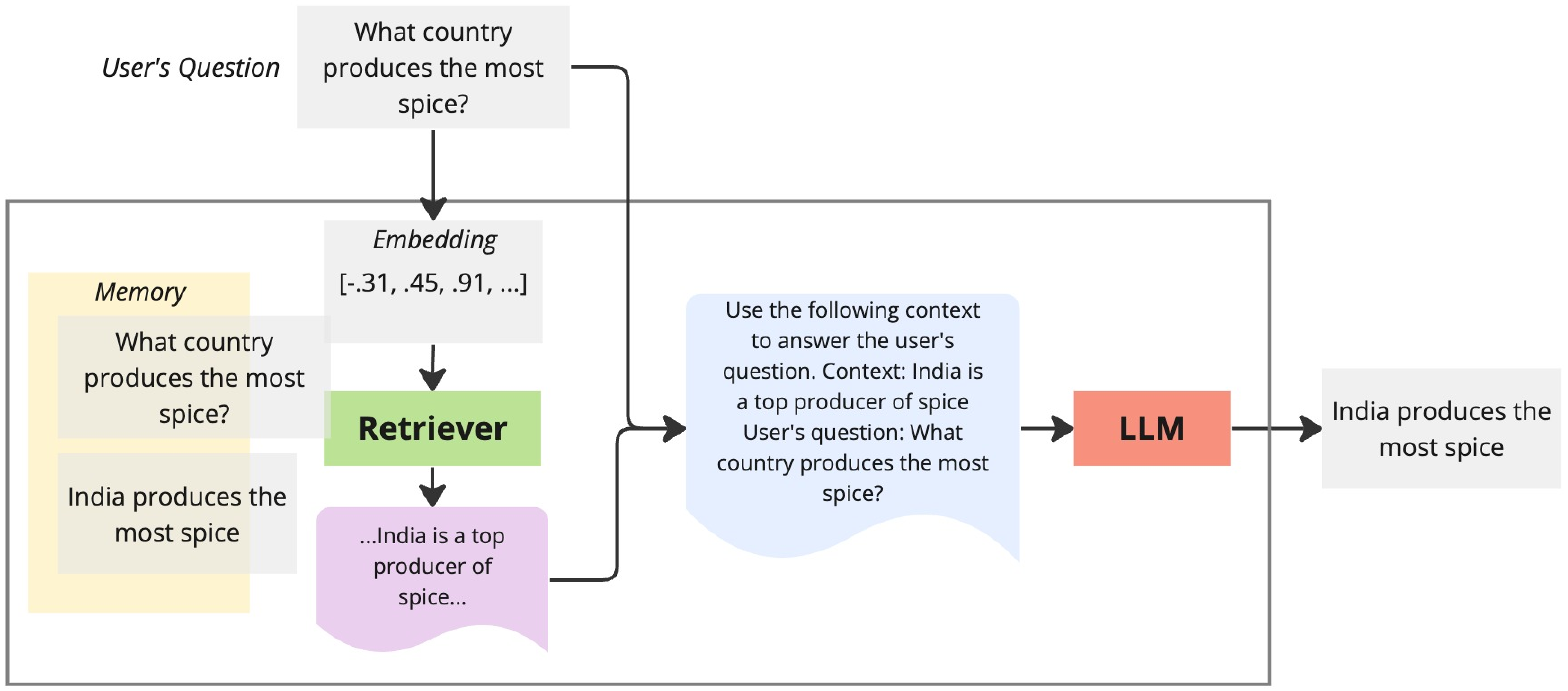

User's Question
How much do they produce?
User's Question as an embedding
[-.31, .45, .91, ...]
Pinecone
Retriever
Most relevant document
...India is a top producer of spice...
Use the following context to answer the user's question. Context: India is a top producer of spice User's question: What country produces the most spice?
LLM
India produces the most spice
DONE

This is *not* how memory works with a ConversationalRetrieval Chain!

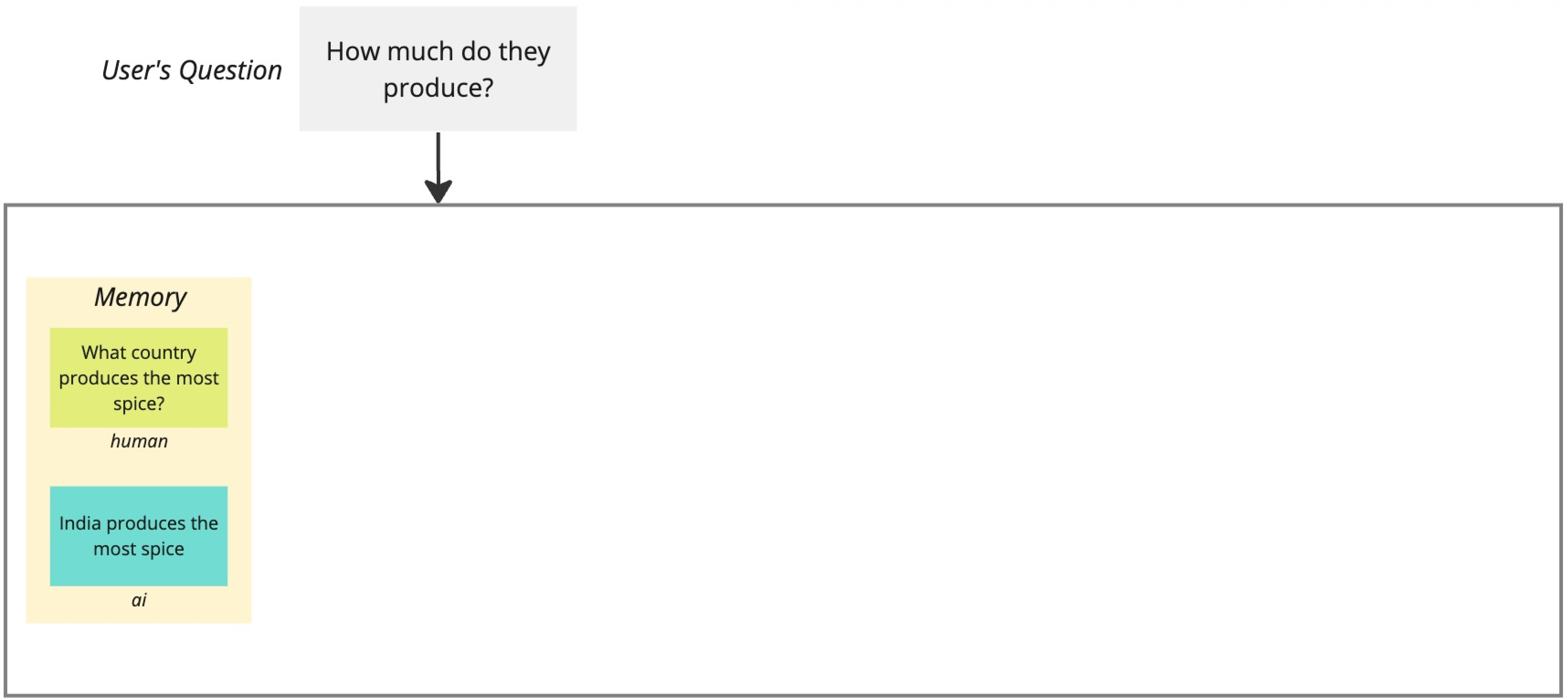

What country makes the second most spice?
ConversationalRetrieval Chain
memory
What country produces the most spice?
India produces the most spice.
How much do they make?
In 2011, India produced 1.5 million metric tons
What country produces the most spice?
India produces the most spice.
How much do they make?
In 2011, India produced 1.5 million metric tons
What country makes the second most spice?
LLM


What country produces the most spice?
ConversationalRetrieval Chain
memory
Memory starts off empty
How much do they produce?
ConversationalRetrieval Chain
M
M
M
memory
Chain behaves differently when memory contains messages!

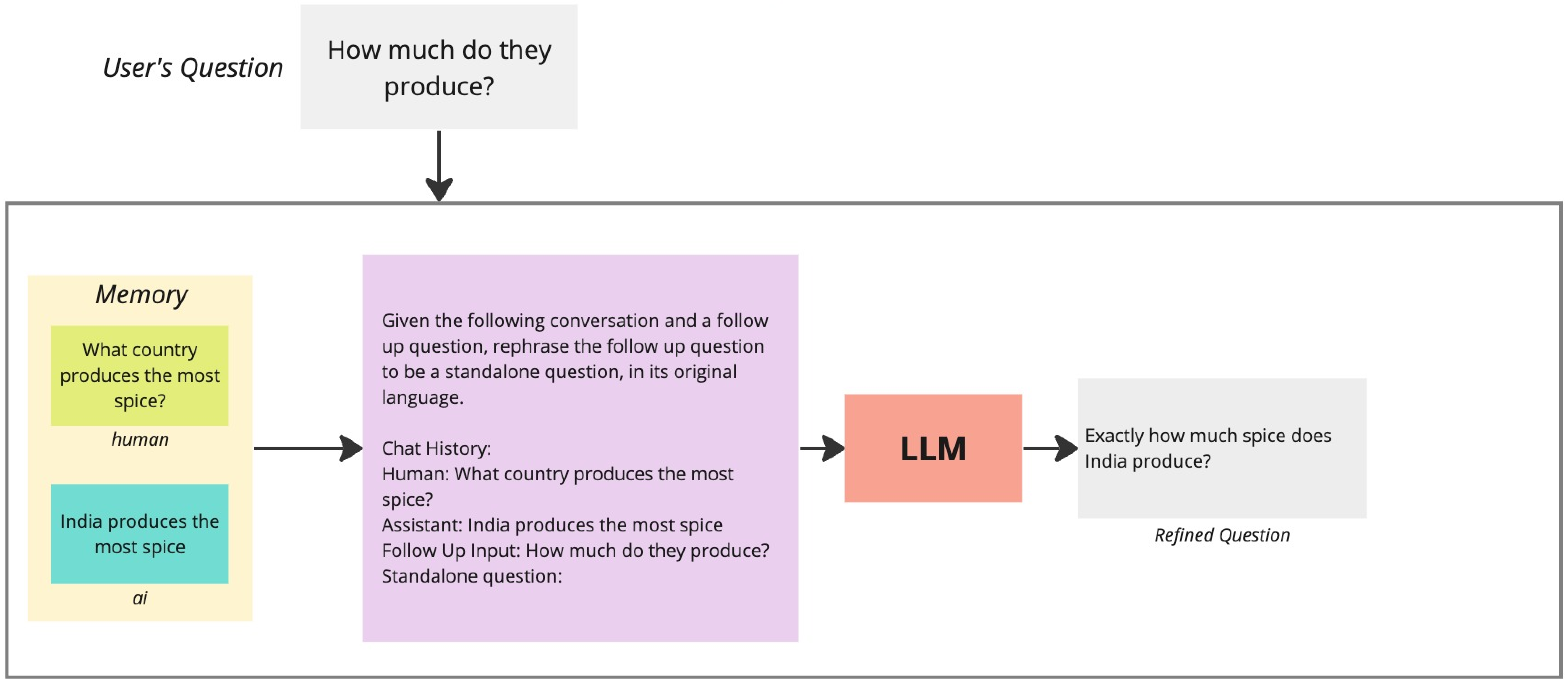
ConversationalRetrieval Chain


User's Question
What country produces the most spice?
Embedding
[-.31, .45, .91, ...]
Memory
What country produces the most spice?
India produces the most spice
Retriever
...India is a top producer of spice...
Use the following context to answer the user's question. Context: India is a top producer of spice User's question: What country produces the most spice?
LLM
India produces the most spice

ConversationalRetrieval Chain


User's Question
How much do they produce?
Memory
What country produces the most spice?
human
India produces the most spice
ai

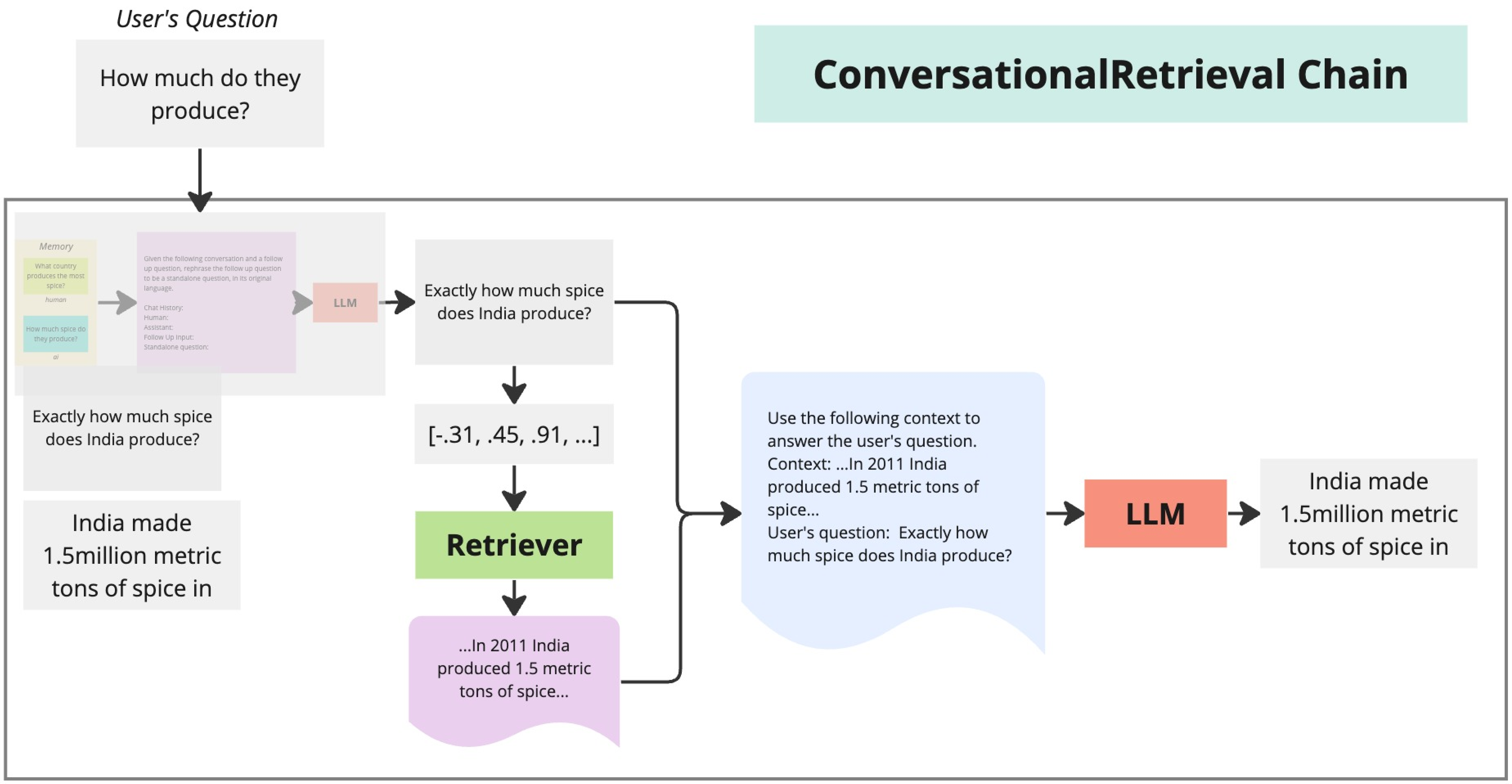
ConversationalRetrieval Chain


User's Question
How much do they produce?
Memory
What country produces the most spice?
human
India produces the most spice
ai
LLM
Exactly how much spice does India produce?
Refined Question

Given the following conversation and a follow up question, rephrase the follow up question to be a standalone question, in its original language.

Chat History:
Human: What country produces the most spice?
Assistant: India produces the most spice
Follow Up Input: How much do they produce?
Standalone question:

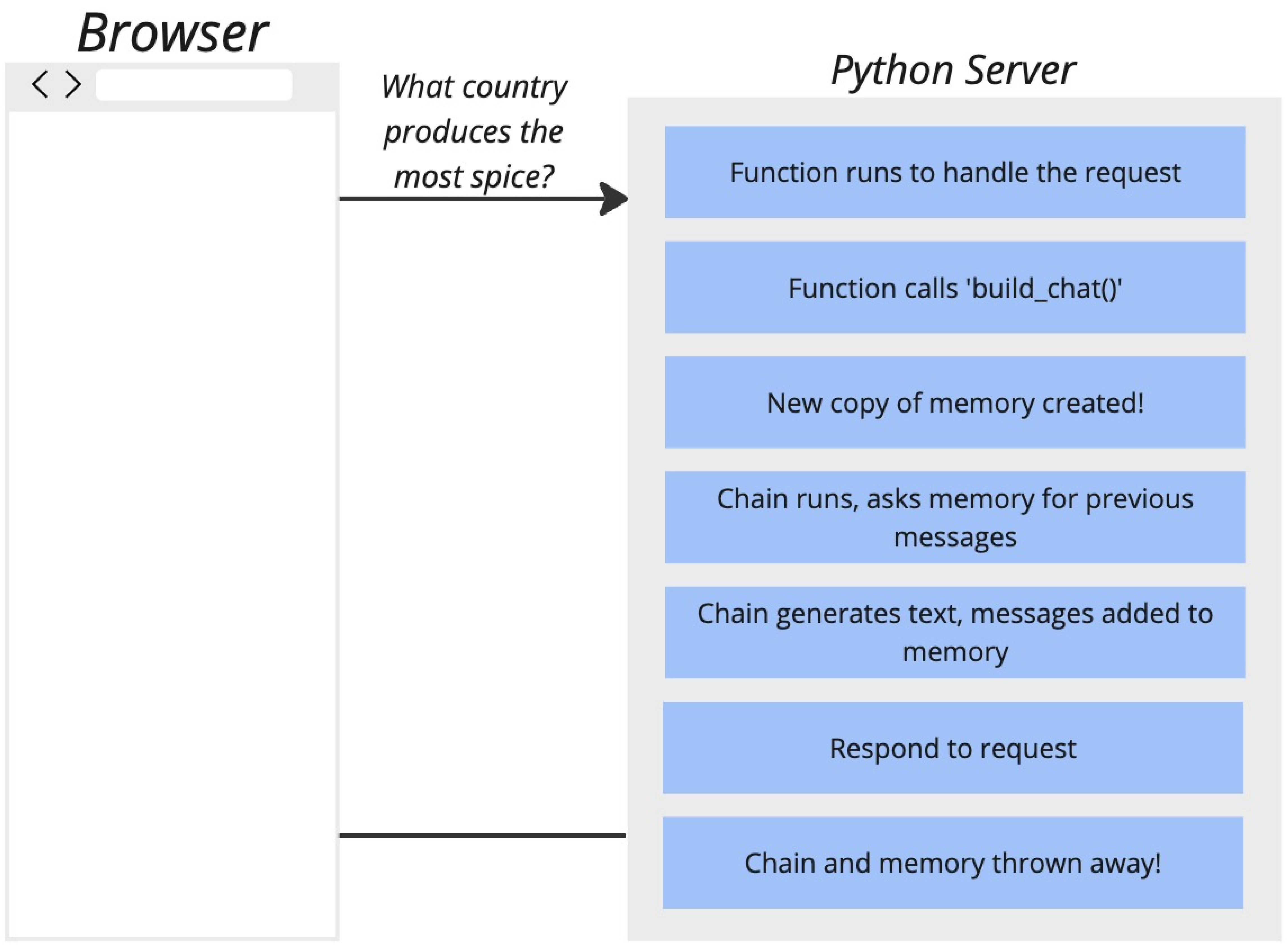
ConversationalRetrieval Chain

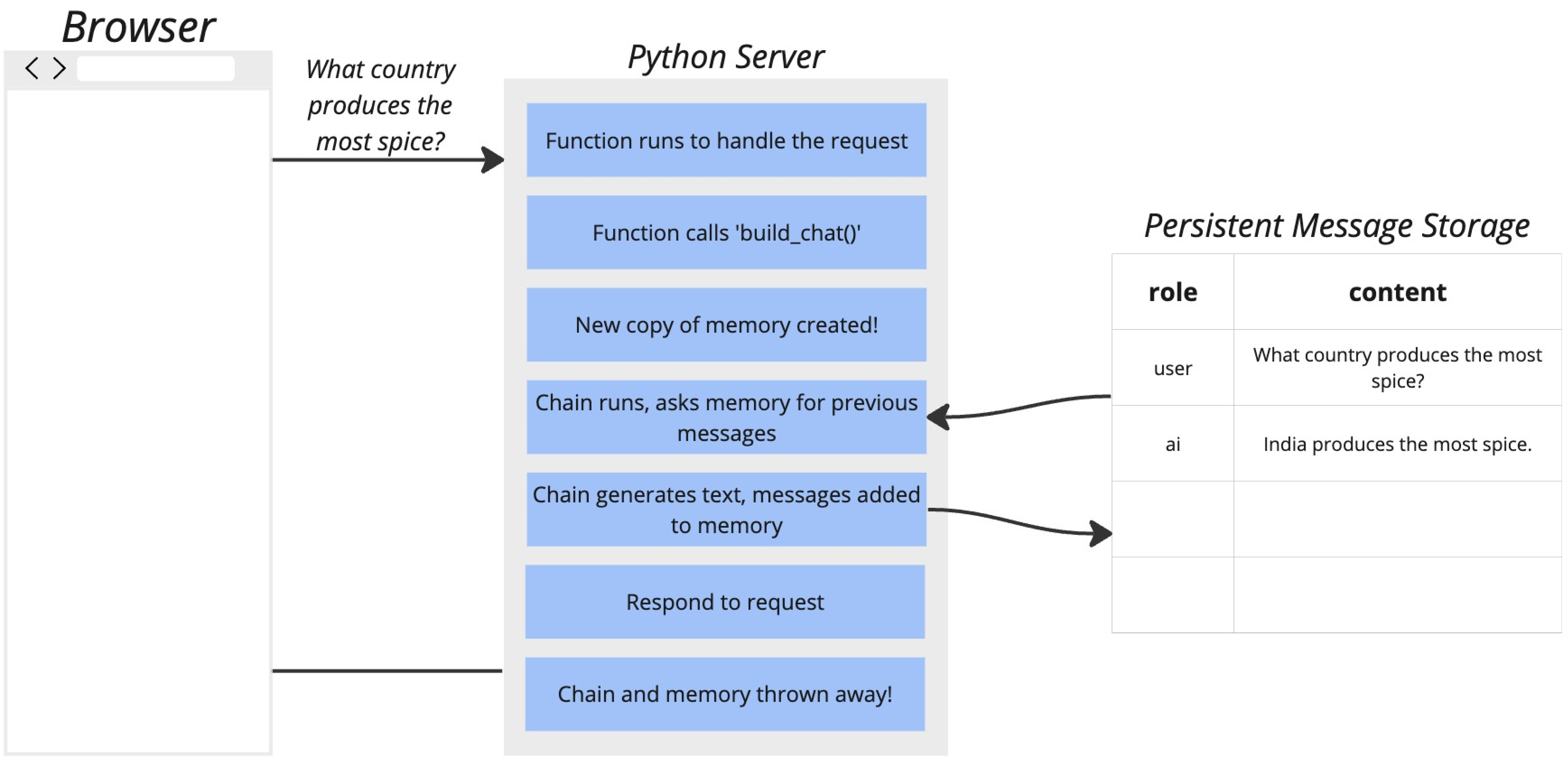

User's Question
How much do they produce?
Memory
What country produces the most spice?
human
How much spice do they produce?
ai
LLM
Exactly how much spice does India produce?
Exactly how much spice does India produce?
[-.31, .45, .91, ...]
Retriever
...In 2011 India produced 1.5 metric tons of spice...
LLM
India made 1.5million metric tons of spice in
India made 1.5million metric tons of spice in

Given the following conversation and a follow up question, rephrase the follow up question to be a standalone question, in its original language.

Chat History:
Human:
Assistant:
Follow Up Input:
Standalone question:

Use the following context to answer the user's question.
Context: ...In 2011 India produced 1.5 metric tons of spice...
User's question: Exactly how much spice does India produce?


Browser
What country produces the most spice?
Python Server
Function runs to handle the request
Function calls 'build_chat()'
New copy of memory created!
Chain runs, asks memory for previous messages
Chain generates text, messages added to memory
Respond to request
Chain and memory thrown away!

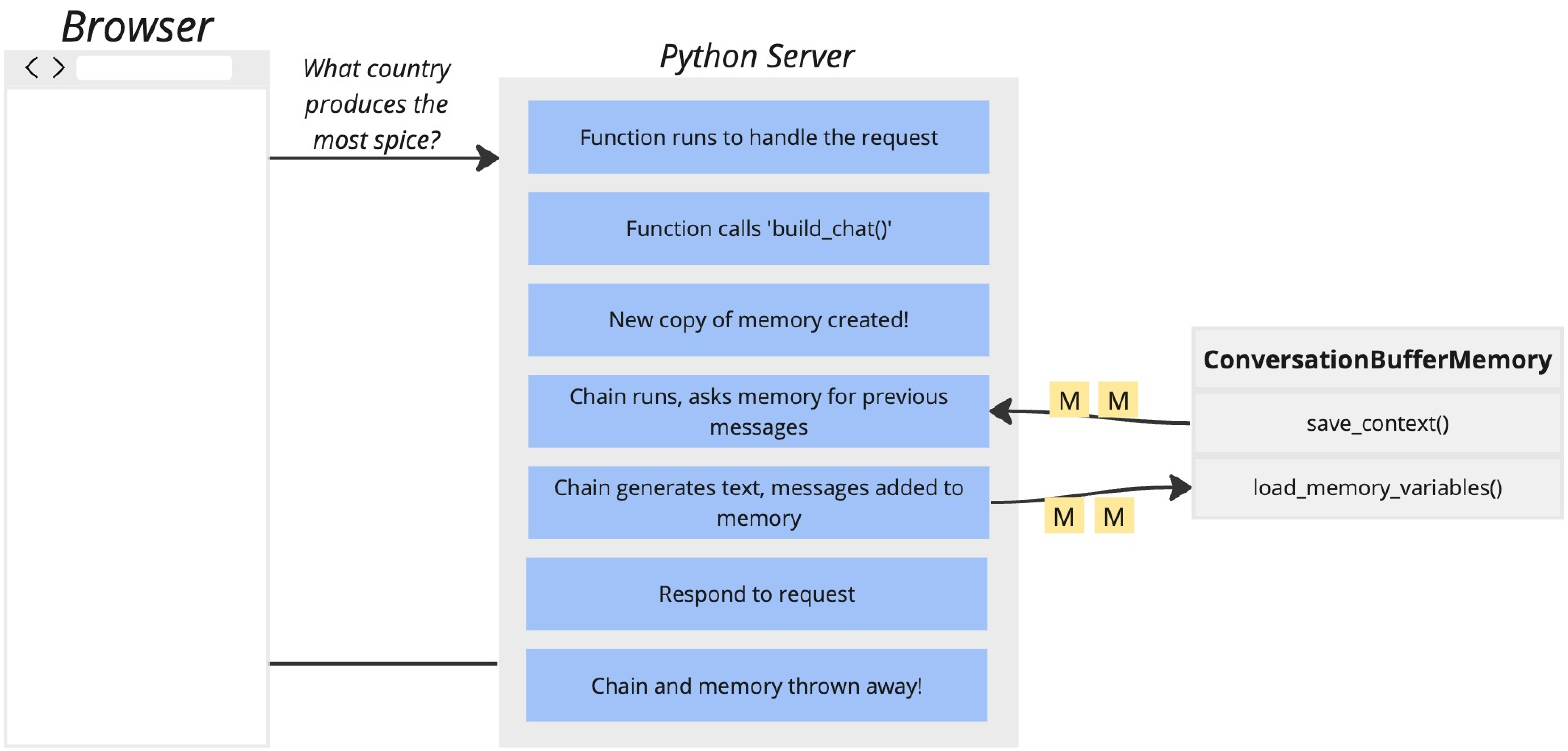

Browser
What country produces the most spice?
Python Server
Function runs to handle the request
Function calls 'build_chat()'
New copy of memory created!
Chain runs, asks memory for previous messages
Chain generates text, messages added to memory
Respond to request
Chain and memory thrown away!
Persistent Message Storage

role	content
user	What country produces the most spice?
ai	India produces the most spice.

```python
class ConversationBufferMemory(BaseChatMemory):
    chat_memory = ChatMessageHistory()

    def save_context(self, user_question, ai_response):
        """Add a user/ai message pair to history"""
        self.chat_memory.add_user_message(user_question)
        self.chat_memory.add_ai_message(ai_response)

    def load_memory_variables(self):
        """Called when the chain needs to access saved messages"""
        return self.chat_memory.messages
```


Browser
What country produces the most spice?
Python Server
Function runs to handle the request
Function calls 'build_chat()'
New copy of memory created!
Chain runs, asks memory for previous messages
Chain generates text, messages added to memory
Respond to request
Chain and memory thrown away!
ConversationBufferMemory
save_context()
load_memory_variables()
M
M
M
M