Requirements

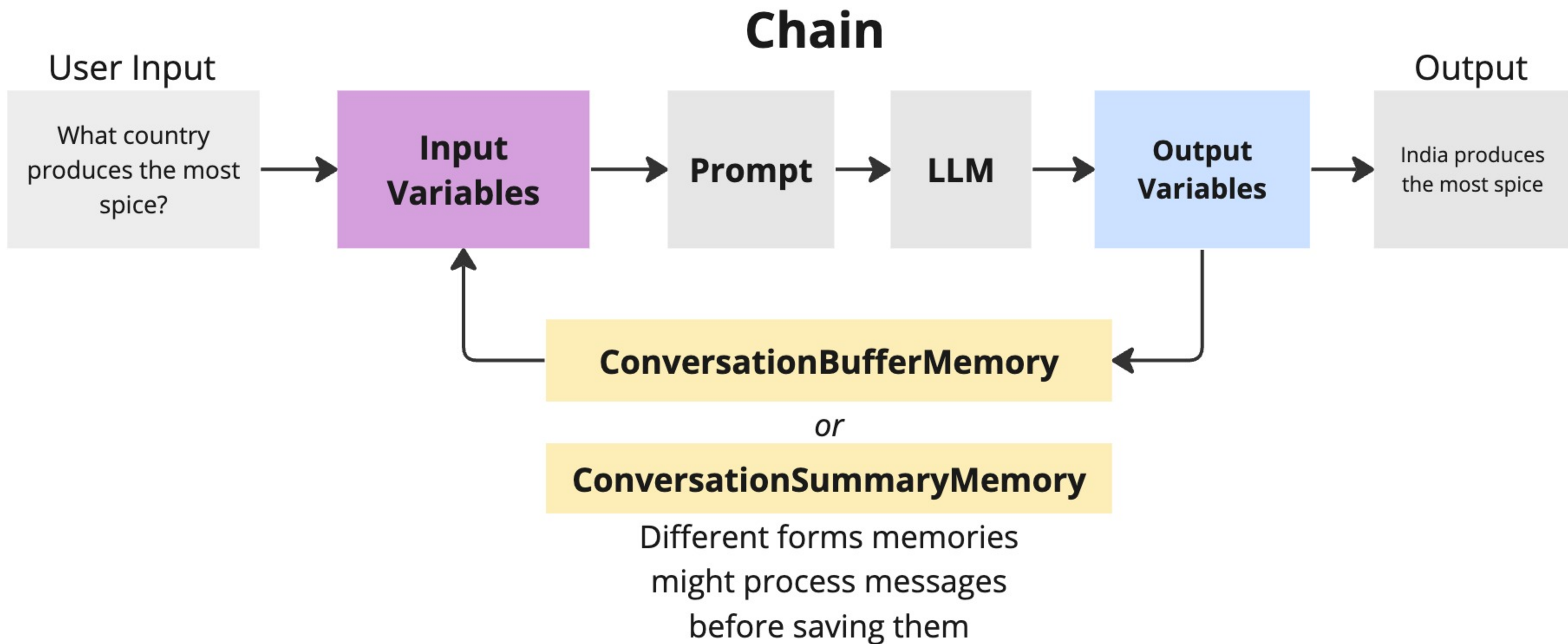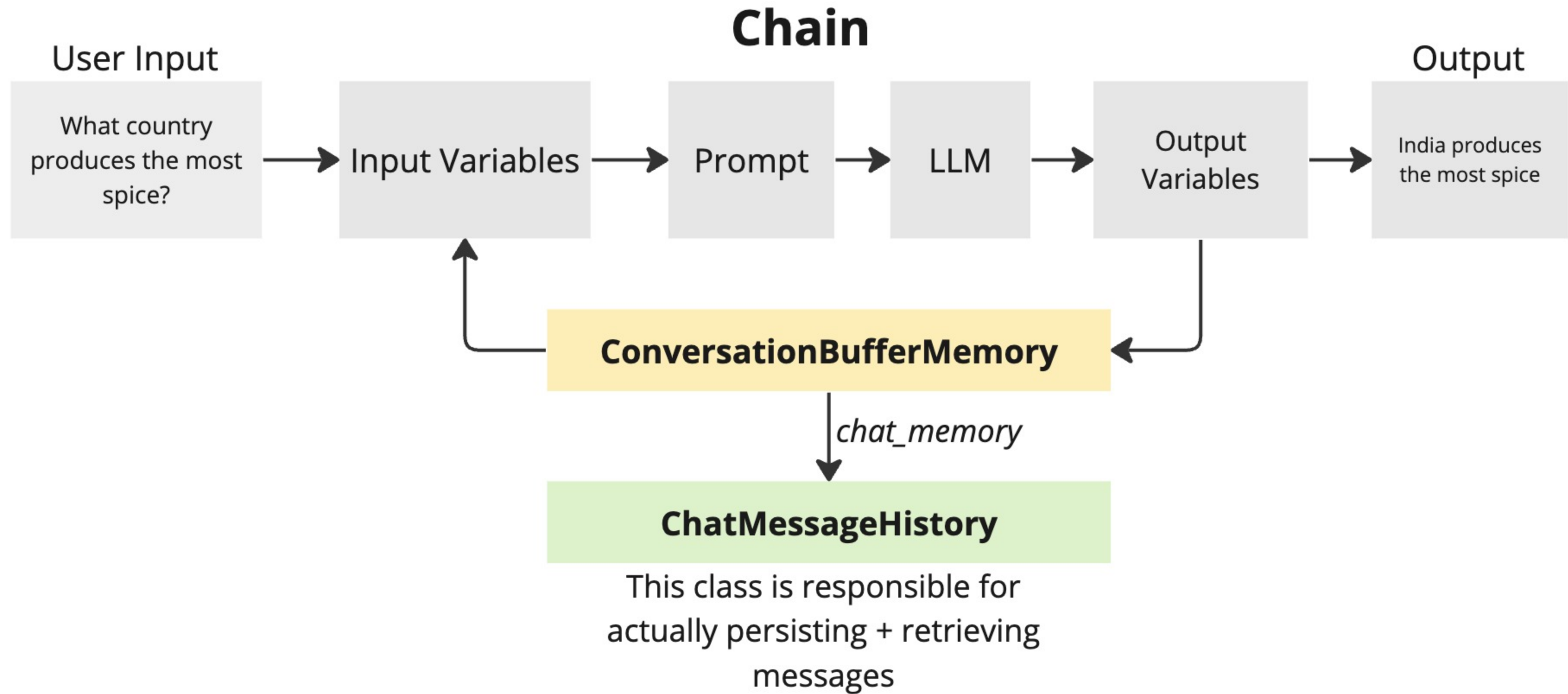|	
Document retrieval needs to be scoped to a particular PDF	→ *Make a scoped retriever*
Need to organize and persist messages + conversations so they can used by both web + chat	→ *Make custom memory to store messages in SQLite*
Need to handle vague user messages	→ Make a Conversational Retrieval Chain

Chain

User Input

What country produces the most spice?

→ **Input Variables** → **Prompt** → **LLM** → **Output Variables** →

Output

India produces the most spice

ConversationBufferMemory

or

ConversationSummaryMemory

Different forms memories might process messages before saving them

Chain

User Input

What country produces the most spice?

Input Variables → Prompt → LLM → Output Variables

Output

India produces the most spice

ConversationBufferMemory

chat_memory

ChatMessageHistory

This class is responsible for actually persisting + retrieving messages

```python
class ChatMessageHistory(BaseChatMessageHistory, BaseModel):
    """In memory implementation of chat message history.

    Stores messages in an in memory list.
    """

    messages: List[BaseMessage] = []

    def add_message(self, message: BaseMessage) -> None:
        """Add a self-created message to the store"""
        self.messages.append(message)

    def clear(self) -> None:
        self.messages = []
```
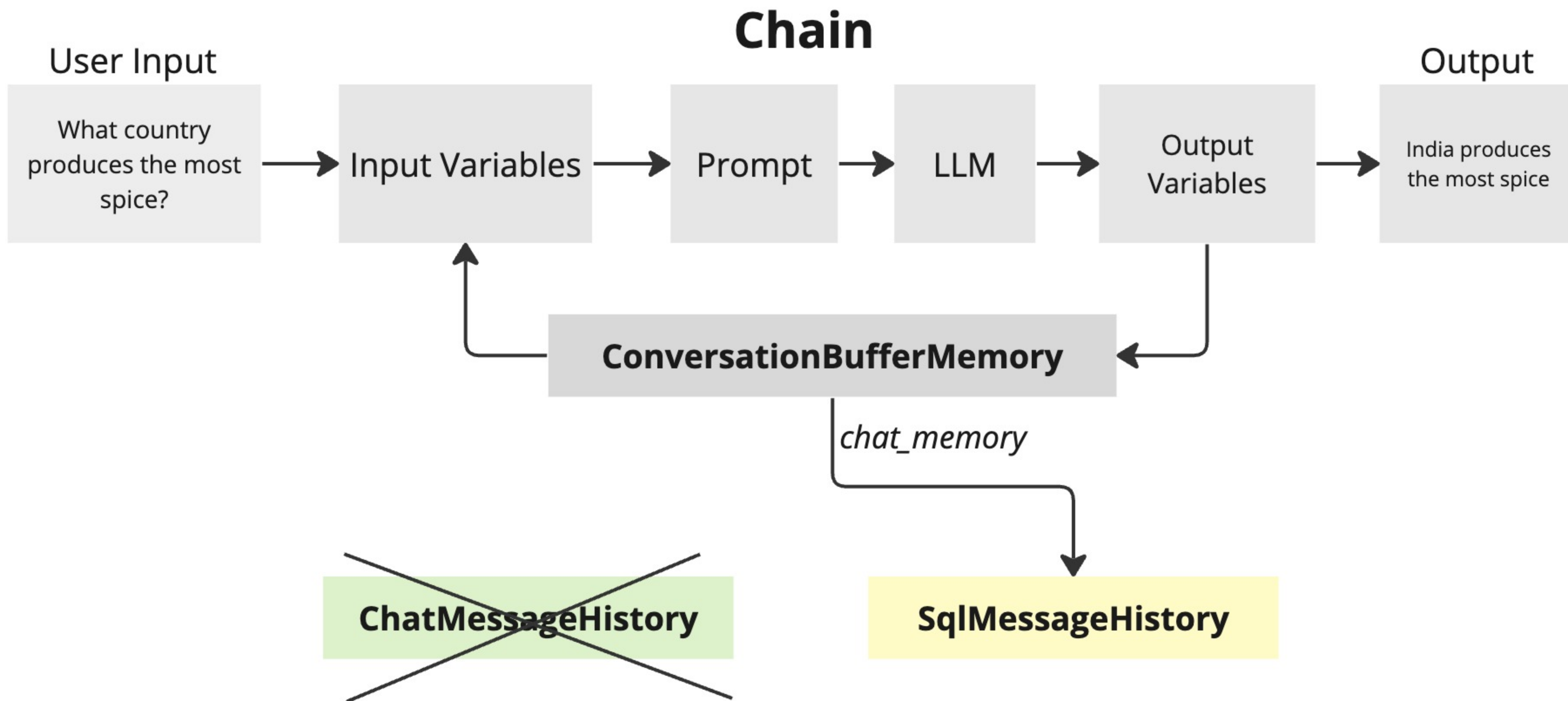
ChatMessageHistory

```python
class ChatMessageHistory(BaseChatMessageHistory, BaseModel):
    """In memory implementation of chat message history.

    Stores messages in an in memory list.
    """

    messages: List[BaseMessage] = []

    def add_message(self, message: BaseMessage) -> None:
        """Add a self-created message to the store"""
        self.messages.append(message)

    def clear(self) -> None:
        self.messages = []
```
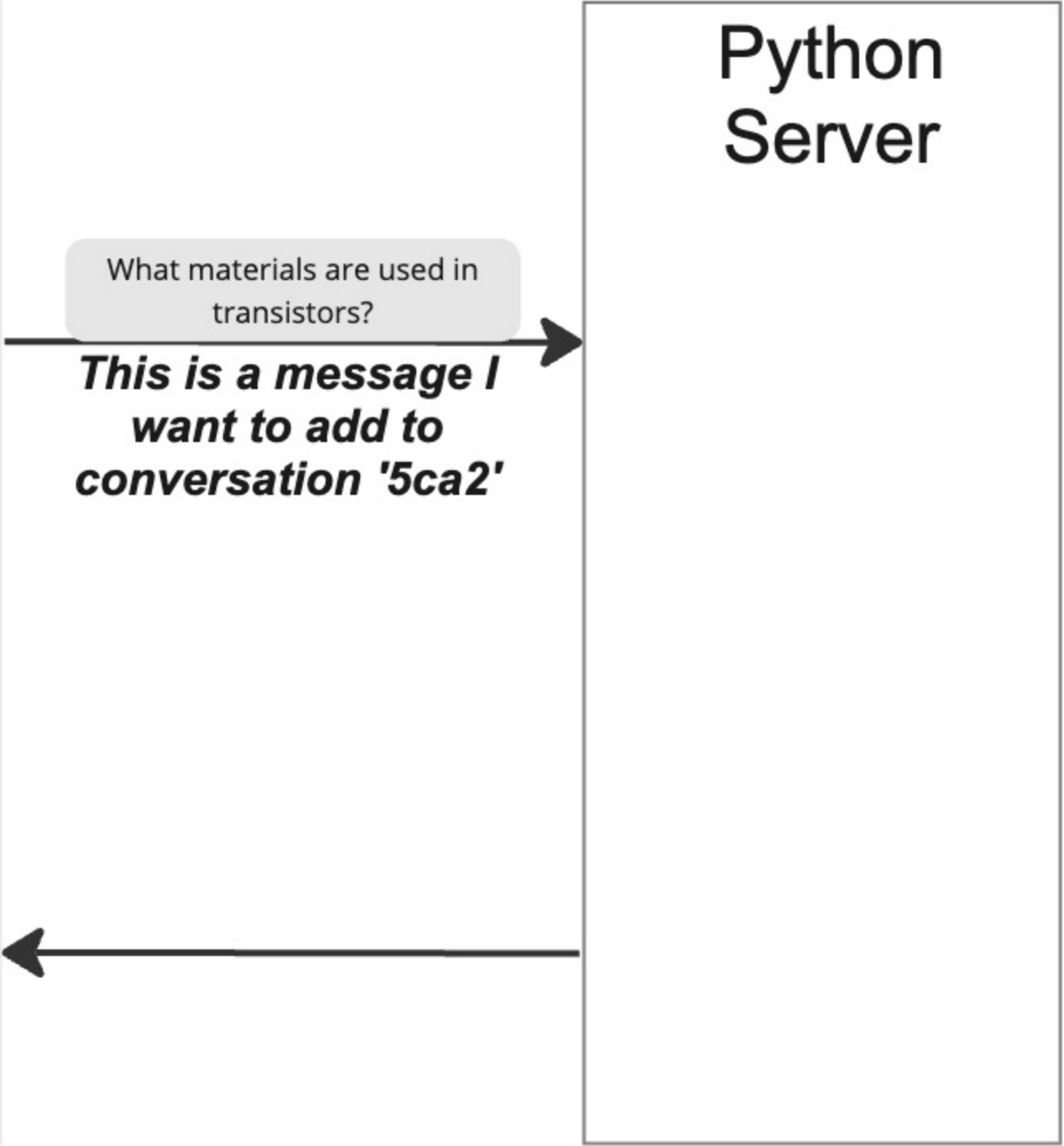
SqlMessageHistory

```python
class SqlMessageHistory(BaseChatMessageHistory, BaseModel):
    conversation_id: str

    @property
    def messages(self):
        """Find all messages in database for conversation_id"""

    def add_message(self, message):
        """Add a message to the database"""

    def clear(self):
        """Do nothing"""
```

We are going to make this
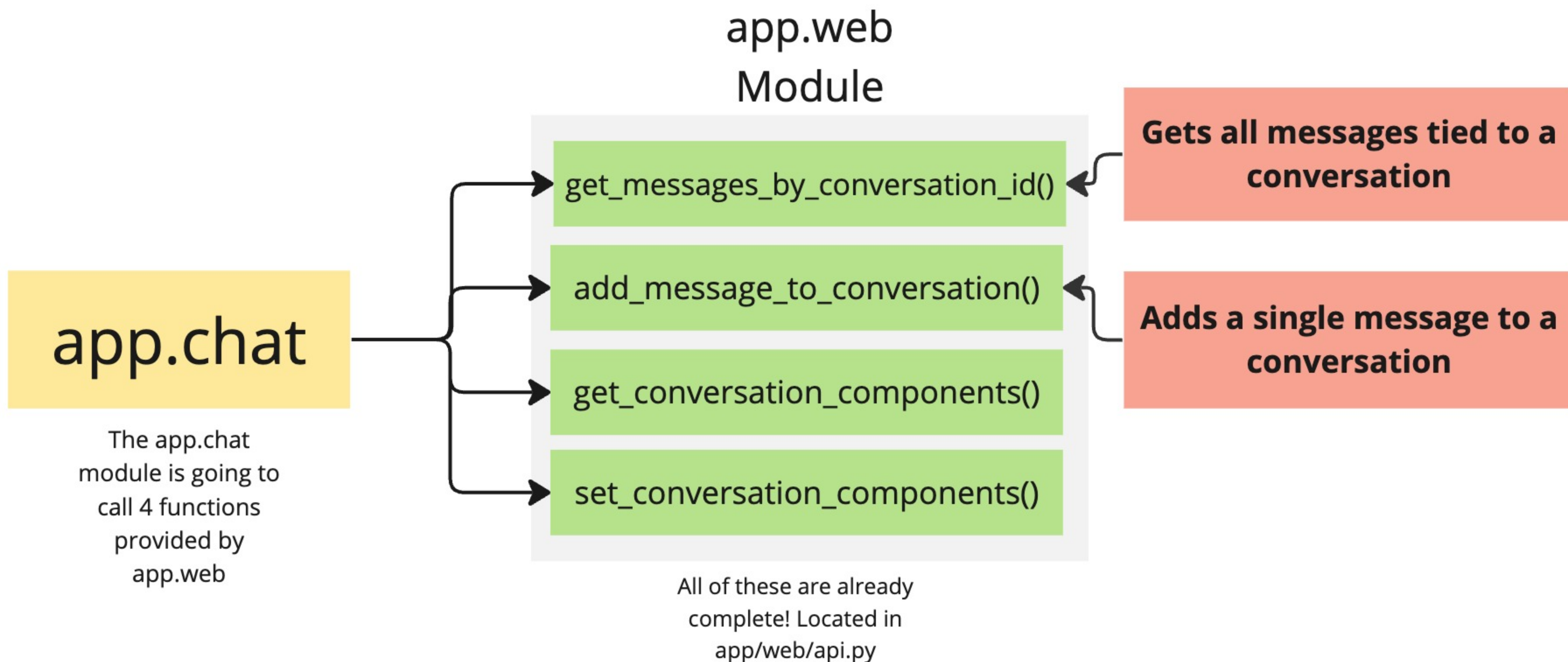
Chain

User Input

What country produces the most spice?

Output

India produces the most spice

Input Variables → Prompt → LLM → Output Variables

ConversationBufferMemory

chat_memory

ChatMessageHistory

SqlMessageHistory

Browser

What materials are used in transistors?

This is a message I want to add to conversation '5ca2'

Python Server

SQLite Database
Table of Messages

ID	conversation_id	role	content
be17bac7	8e26	human	What country produces...
ccb58e26	8e26	ai	India produces the most...
3a495ca2	8e26	human	How much?
9f78186e	5ca2	human	How are transistors made?
cb58e26	5ca2	ai	Transistors are made by..
wk4jlkj45	5ca2	human	What materials are used in transistors?

app.web
Module

app.chat

The app.chat module is going to call 4 functions provided by app.web

get_messages_by_conversation_id()

add_message_to_conversation()

get_conversation_components()

set_conversation_components()

All of these are already complete! Located in app/web/api.py

Gets all messages tied to a conversation

Adds a single message to a conversation
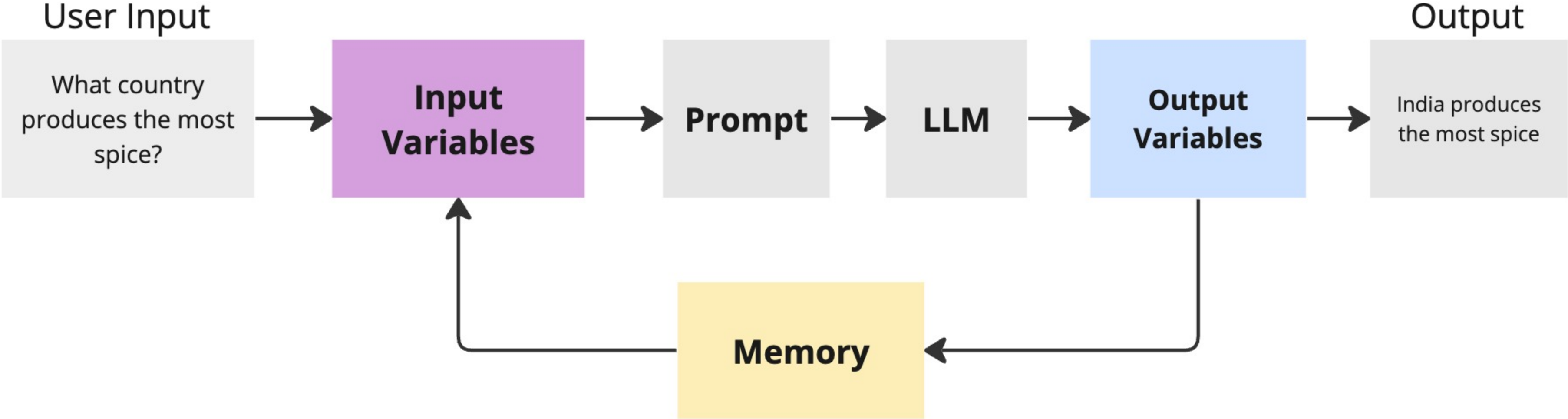
Chain

Chain

Input Variables

User Input

What country produces the most spice?

key	value

Prompt → **LLM**

Output Variables

key	value

Memory

ConversationalQA Chain

User Input

What country produces the most spice?

Input Variables

key	value
question	
chat_history	

Chain 1 → **Chain 2**

Output Variables

key	value
answer	

Memory

ConversationalQA Chain

Input Variables

User Input

What country produces the most spice?

key	value
question	
chat_history	

Chain 1 → **Chain 2** →

Output Variables

key	value
answer	

Custom Memory