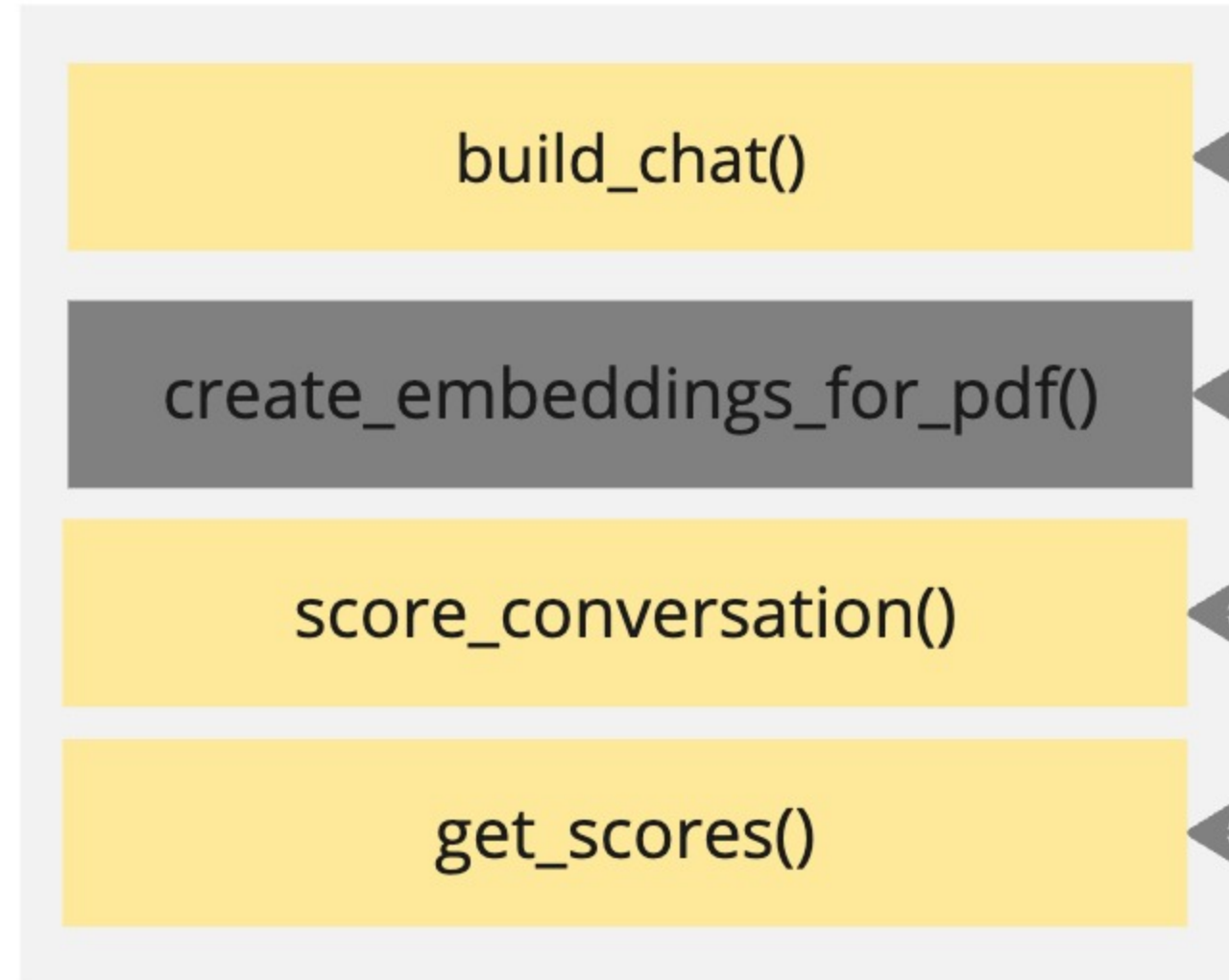


app.chat Module



*We need to implement
these! (plus tons of
supporting code)*

app.web

*The app.web modules
needs us to implement
these 4 functions*

Browser

< >

*What country
produces the
most spice?*

Python Server

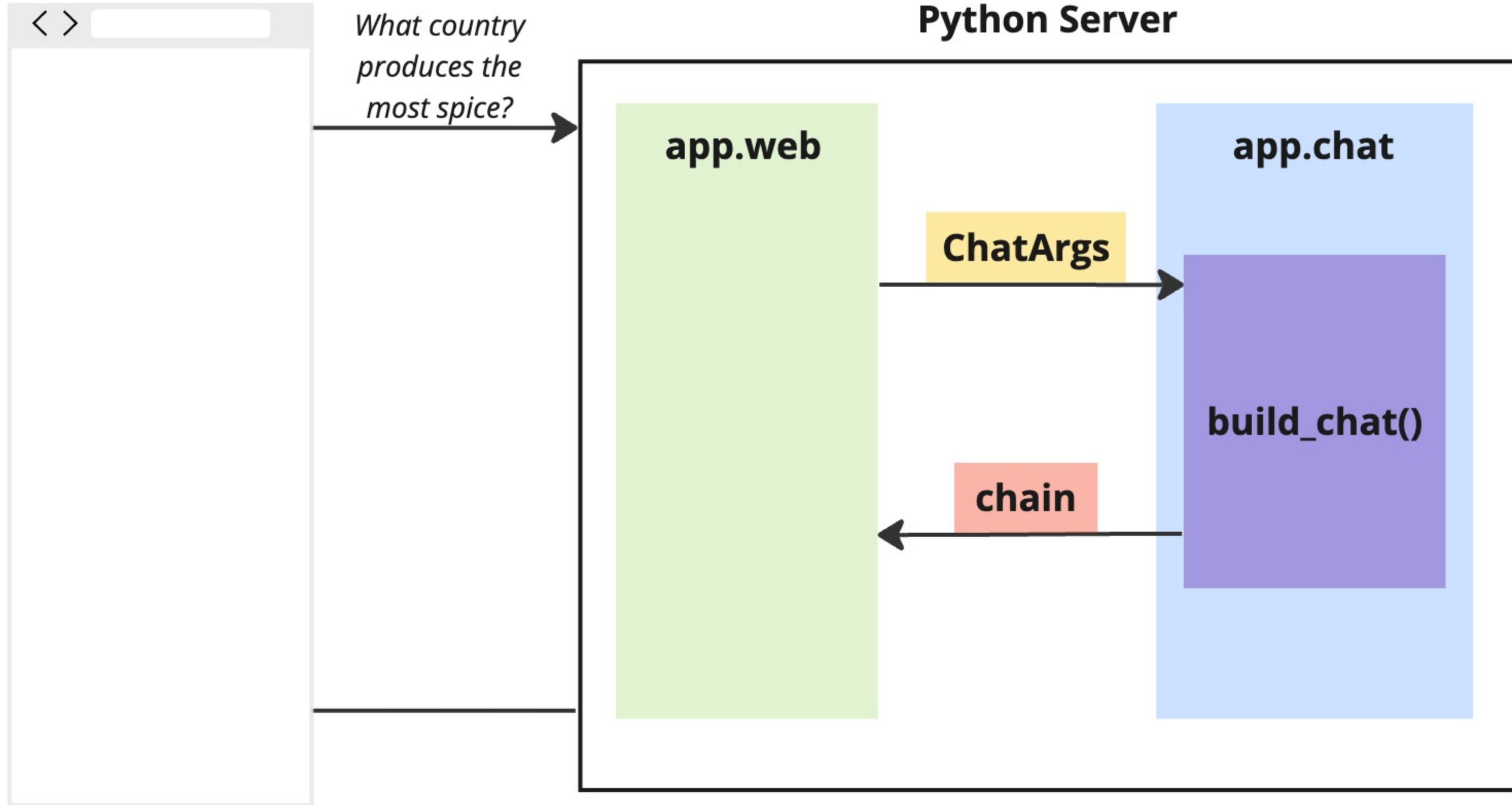
app.web

ChatArgs

app.chat

build_chat()

chain



We know that we'll need some kind of
RetrievalQA chain



Let's review our requirements and make sure
'RetrievalQA' chain is actually the right choice

It isn't

←

→

↺

×

Home

Documents

Scores

Sign Out

Your Documents

New

| Name | PDF ID | Action |
|-----------------|--------|----------------------|
| my_document.pdf | ca366 | View |
| business.pdf | b7c9 | View |
| spice.pdf | e763 | View |

HomeDocumentsScoresSign Out

HistoryNew Chat

What country produces the most spice?

India produces the most spice.

How much?

In 2011, India produced 1.5 million metric tons of spice

alskdfjlaksjdfllkasjdf

WIKIPEDIA

The Free Encyclopedia

Spice

A **spice** is a seed, fruit, root, bark, or other plant substance primarily used for flavoring or coloring food. Spices are distinguished from herbs, which are the leaves, flowers, or stems of plants used for flavoring or as a garnish. Spices are sometimes used in medicine, religious rituals, cosmetics, or perfume production. For example, vanilla is commonly used as an ingredient in fragrance manufacturing.^[1]

A spice may be available in several forms: fresh, whole-dried, or pre-ground dried. Generally, spices are dried. Spices may be ground into a powder for convenience. A whole dried spice has the longest shelf life, so it can be purchased and stored in larger amounts, making it cheaper on a per-serving basis. A fresh spice, such as ginger, is usually more flavorful than its dried form, but fresh spices are more expensive and have a much shorter shelf life. Some spices are not always available either fresh or whole, for example, turmeric, and often must be purchased in ground form. Small seeds, such as fennel and mustard seeds, are often used both whole and in powder form.


As of 2019, there is not enough clinical evidence to indicate that consuming spices affects human health.^[2]

India contributes to 75% of global spice production. This is reflected culturally through their cuisine; historically, the spice trade developed throughout the Indian subcontinent as well as in East Asia and the Middle East. Europe's demand for spices was among the economic and cultural factors that encouraged exploration in the early modern period.


Etymology

The word *spice* originated in Middle English^[3] which came from the Old French words *espece*, *espis(c)e*, and *espis(c)e*.^[4] According to the *Middle English Dictionary*, the Old French words came from Anglo-French *spice*;^[4] according to Merriam Webster,


Spices



Spices at a central market in Agadir, Morocco



A group of Indian herbs and spices in bowls



Spices of São Paulo flea market, São Paulo, Brazil

← → ↻

×

HomeDocumentsScoresSign Out

HistoryNew Chat

What country produces the most spice?

India produces the most spice.

How much?

In 2011, India produced 1.5 million metric tons of spice

Add text

WIKIPEDIA
The Free Encyclopedia

Spice


A **spice** is a seed, fruit, root, bark, or other plant substance primarily used for flavoring or coloring food. Spices are distinguished from herbs, which are the leaves, flowers, or stems of plants used for flavoring or as a garnish. Spices are sometimes used in medicine, religious rituals, cosmetics, or perfume production. For example, vanilla is commonly used as an ingredient in fragrance manufacturing.^[1]

A spice may be available in several forms: fresh, whole-dried, or pre-ground dried. Generally, spices are dried. Spices may be ground into a powder for convenience. A whole dried spice has the longest shelf life, so it can be purchased and stored in larger amounts, making it cheaper on a per-serving basis. A fresh spice, such as ginger, is usually more flavorful than its dried form, but fresh spices are more expensive and have a much shorter shelf life. Some spices are not always available either fresh or whole, for example, turmeric, and often must be purchased in ground form. Small seeds, such as fennel and mustard seeds, are often used both whole and in powder form.


As of 2019, there is not enough clinical evidence to indicate that consuming spices affects human health.^[2]

India contributes to 75% of global spice production. This is reflected culturally through their cuisine; historically, the spice trade developed throughout the Indian subcontinent as well as in East Asia and the Middle East. Europe's demand for spices was among the economic and cultural factors that encouraged exploration in the early modern period.


Spices



Spices at a central market in Agadir, Morocco



A group of Indian herbs and spices in bowls



Spices of Saúde flea market, São Paulo, Brazil

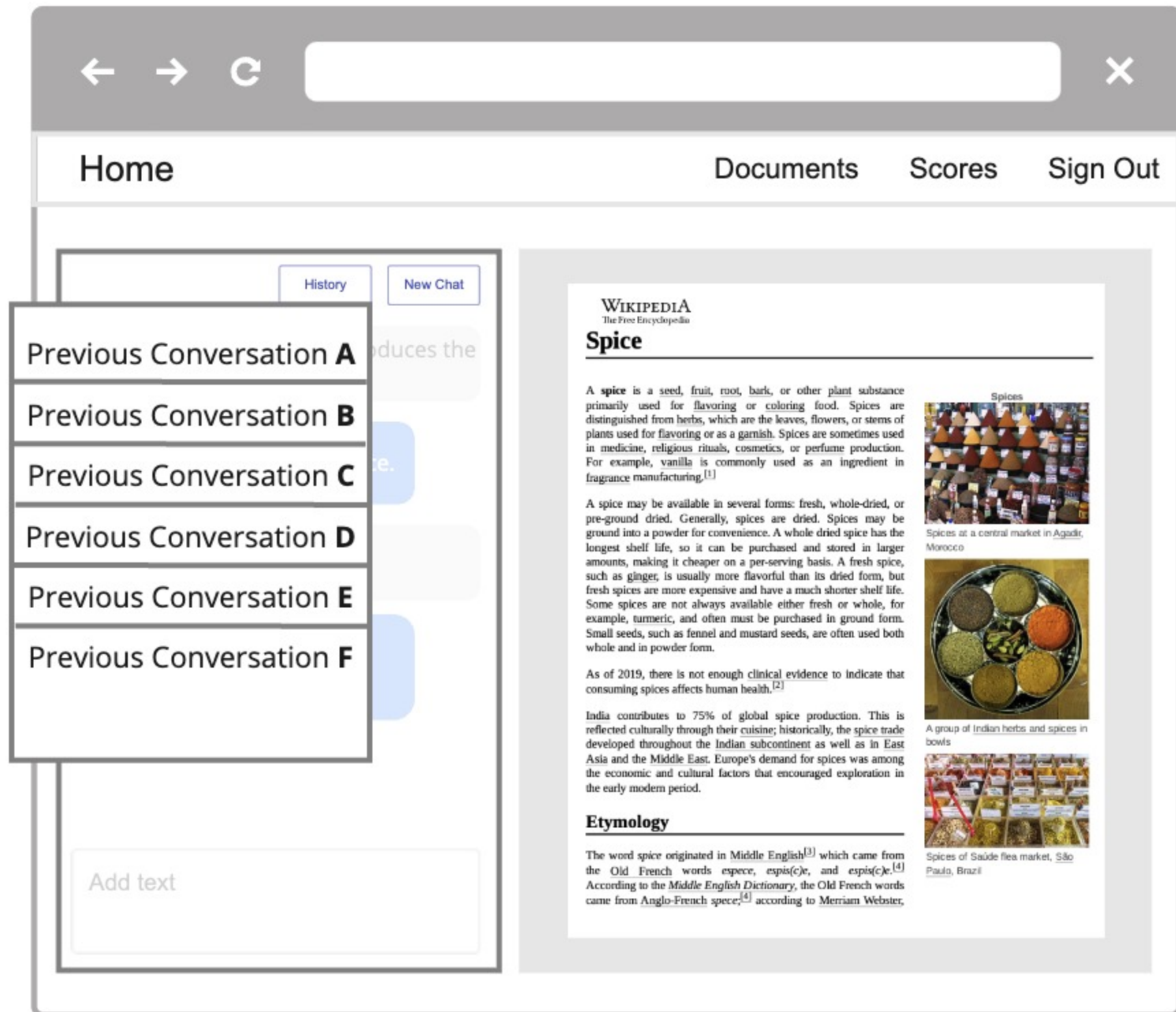
Etymology

The word *spice* originated in Middle English^[3] which came from the Old French words *espece*, *espis(c)e*, and *espis(c)ie*.^[4] According to the *Middle English Dictionary*, the Old French words came from Anglo-French *spece*;^[4] according to Merriam Webster,

When a user views a PDF,
they want to chat with
that PDF



Any document retrieval we
do should find documents
that came from only the
PDF the user is viewing



A user can have many
separate persisted
conversations

Within a conversation, there
may be many messages

←

→

↺

×

Home

Documents

Scores

Sign Out

History

New Chat

What country produces the most spice?

India produces the most spice.

How much?

In 2011, India produced 1.5 million metric tons of spice

Add text

WIKIPEDIA

The Free Encyclopedia

Spice

A **spice** is a seed, fruit, root, bark, or other plant substance primarily used for flavoring or coloring food. Spices are distinguished from herbs, which are the leaves, flowers, or stems of plants used for flavoring or as a garnish. Spices are sometimes used in medicine, religious rituals, cosmetics, or perfume production. For example, vanilla is commonly used as an ingredient in fragrance manufacturing.^[1]

A spice may be available in several forms: fresh, whole-dried, or pre-ground dried. Generally, spices are dried. Spices may be ground into a powder for convenience. A whole dried spice has the longest shelf life, so it can be purchased and stored in larger amounts, making it cheaper on a per-serving basis. A fresh spice, such as ginger, is usually more flavorful than its dried form, but fresh spices are more expensive and have a much shorter shelf life. Some spices are not always available either fresh or whole, for example, turmeric, and often must be purchased in ground form. Small seeds, such as fennel and mustard seeds, are often used both whole and in powder form.


As of 2019, there is not enough clinical evidence to indicate that consuming spices affects human health.^[2]

India contributes to 75% of global spice production. This is reflected culturally through their cuisine; historically, the spice trade developed throughout the Indian subcontinent as well as in East Asia and the Middle East. Europe's demand for spices was among the economic and cultural factors that encouraged exploration in the early modern period.


Etymology

The word *spice* originated in Middle English^[3] which came from the Old French words *espece*, *espis(c)e*, and *espis(c)e*.^[4] According to the *Middle English Dictionary*, the Old French words came from Anglo-French *spice*,^[4] according to Merriam Webster,


Spices



Spices at a central market in Agadir, Morocco



A group of Indian herbs and spices in bowls



Spices of Sãode flea market, São Paulo, Brazil

Conversations/messages
will persist even if a user
leaves the page

Conversations/messages
will be used ***outside of any
LLM!!***

←

→

↺

×

Home

Documents

Scores

Sign Out

History

New Chat

What country produces the most spice?

India produces the most spice.

How much?

In 2011, India produced 1.5 million metric tons of spice

Add text

WIKIPEDIA

The Free Encyclopedia

Spice


A **spice** is a seed, fruit, root, bark, or other plant substance primarily used for flavoring or coloring food. Spices are distinguished from herbs, which are the leaves, flowers, or stems of plants used for flavoring or as a garnish. Spices are sometimes used in medicine, religious rituals, cosmetics, or perfume production. For example, vanilla is commonly used as an ingredient in fragrance manufacturing.^[1]

A spice may be available in several forms: fresh, whole-dried, or pre-ground dried. Generally, spices are dried. Spices may be ground into a powder for convenience. A whole dried spice has the longest shelf life, so it can be purchased and stored in larger amounts, making it cheaper on a per-serving basis. A fresh spice, such as ginger, is usually more flavorful than its dried form, but fresh spices are more expensive and have a much shorter shelf life. Some spices are not always available either fresh or whole, for example, turmeric, and often must be purchased in ground form. Small seeds, such as fennel and mustard seeds, are often used both whole and in powder form.


As of 2019, there is not enough clinical evidence to indicate that consuming spices affects human health.^[2]

India contributes to 75% of global spice production. This is reflected culturally through their cuisine; historically, the spice trade developed throughout the Indian subcontinent as well as in East Asia and the Middle East. Europe's demand for spices was among the economic and cultural factors that encouraged exploration in the early modern period.


Spices



Spices at a central market in Agadir, Morocco



A group of Indian herbs and spices in bowls



Spices of Saúde flea market, São Paulo, Brazil

Etymology

The word spice originated in Middle English^[3] which came from the Old French words *espece*, *espis(c)te*, and *espis(c)te*.^[4] According to the *Middle English Dictionary*, the Old French words came from Anglo-French *spece*,^[4] according to Merriam Webster,

Some user messages will be very unclear, or refer to previous messages in the conversation

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 be used by both web + chat

→ *Make custom memory to store messages in SQLite*

Need to handle vague user messages

→ *Make a Conversational Retrieval Chain*


Document retrieval needs to be scoped to a particular PDF


Making Our Retriever

```
1 vector_store = Pinecone.from_existing_index(  
2     os.getenv("PINECONE_INDEX_NAME"), embeddings  
3 )  
4  
5 search_kwargs = {"filter": {"pdf_id": 123 }}  
6  
7 retriever = vector_store.as_retriever(  
8     search_kwargs=search_kwargs  
9 )
```

Pinecone

| ID | vector | metadata | | |
|----------|---------|----------|----------------------|--------|
| | | page | text | pdf_id |
| be17bac7 | [...] | 3 | "spice is..." | 123 |
| ccb58e26 | [...] | 5 | "spice is..." | 123 |
| 3a495ca2 | [...] | 1 | "spice is..." | 123 |
| 9f78186e | [...] | 2 | "spice is..." | 123 |
| ccb58e26 | [...] | 1 | "transistors are..." | 456 |
| 3a495ca2 | [...] | 2 | "transistors are..." | 456 |
| 9f78186e | [...] | 4 | "transistors are..." | 456 |

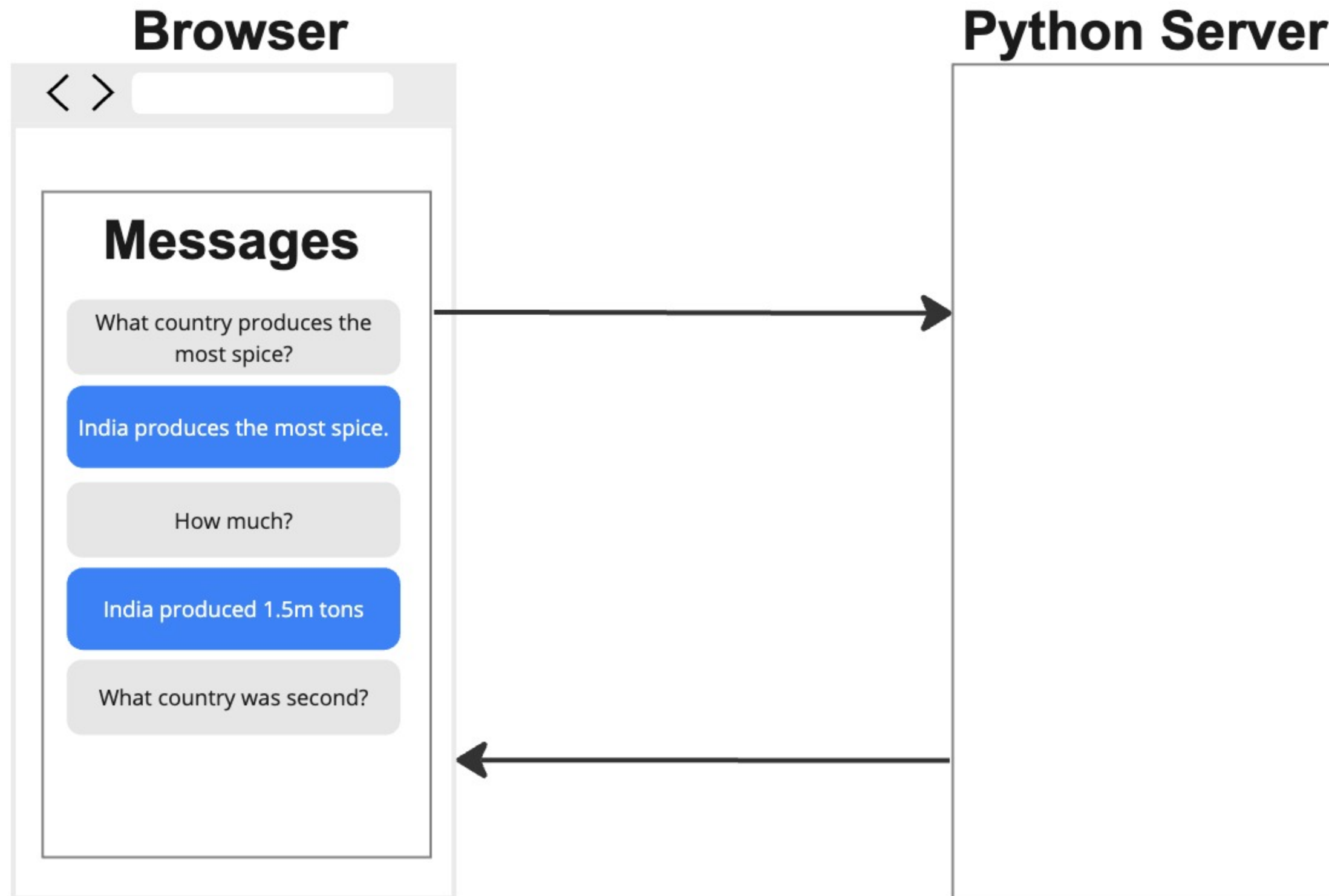

spice.pdf


transistors.pdf

Need to persist messages + conversations so they can be used by both web + chat

Option 1

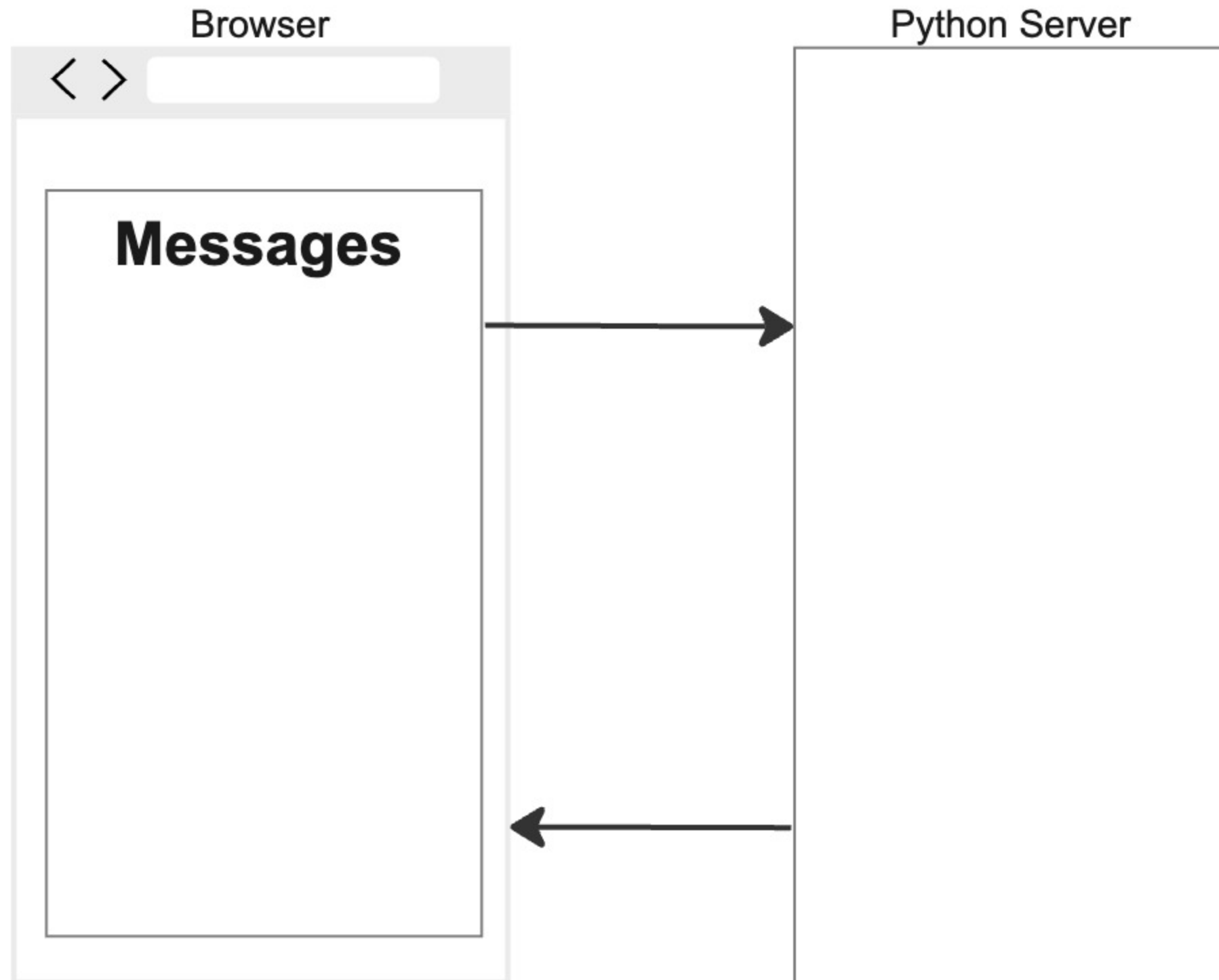
Store messages on the client. Whenever user submits a question, send all the messages over.



Need to persist messages + conversations so they can be used by both web + chat

Option 1

Store messages on the client. Whenever user submits a question, send all the messages over.



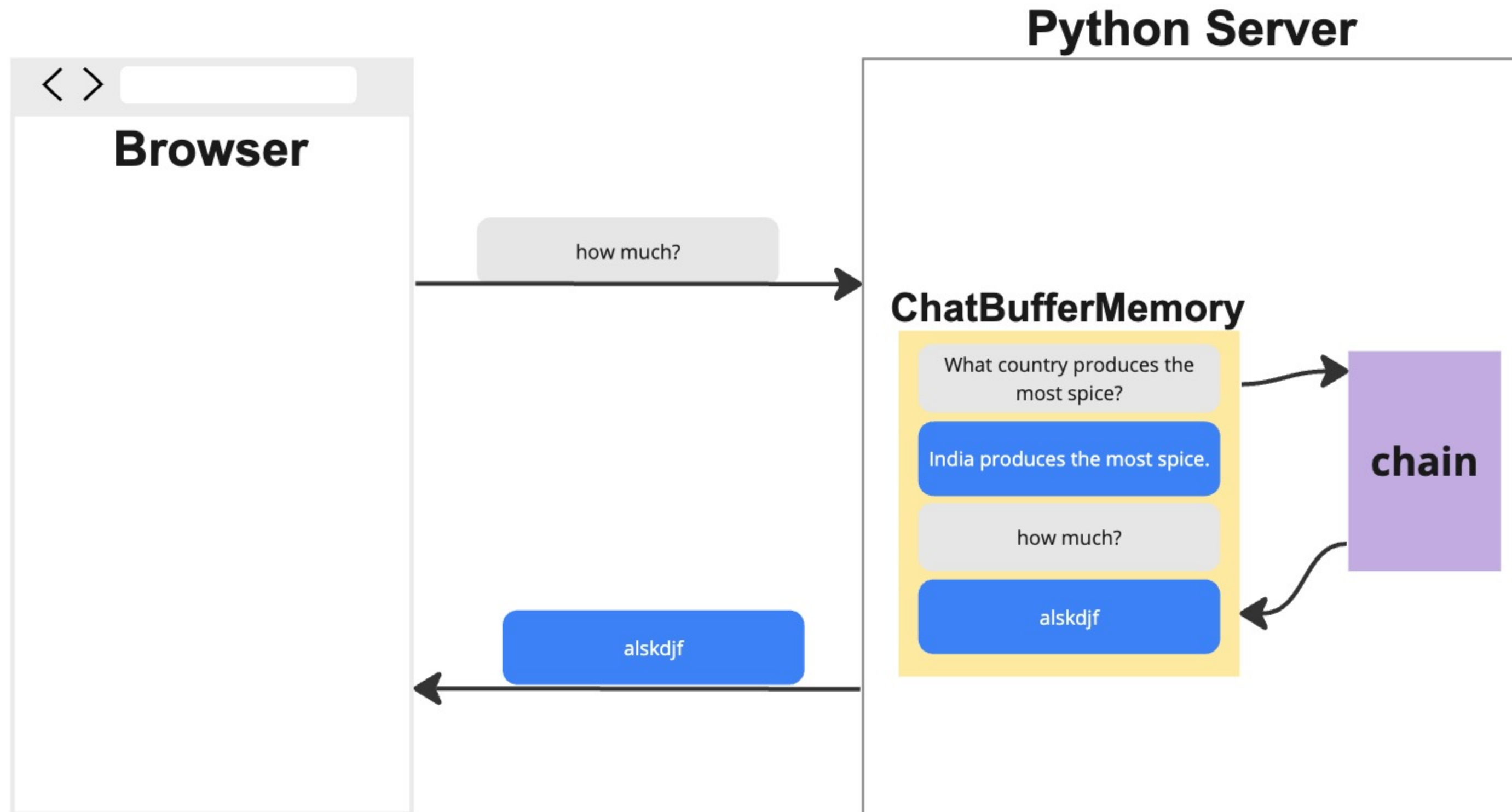
Easy to implement!

**No message persistence!
If user refreshes the
page, messages are lost**

Need to persist messages + conversations so they can be used by both web + chat

Option 2

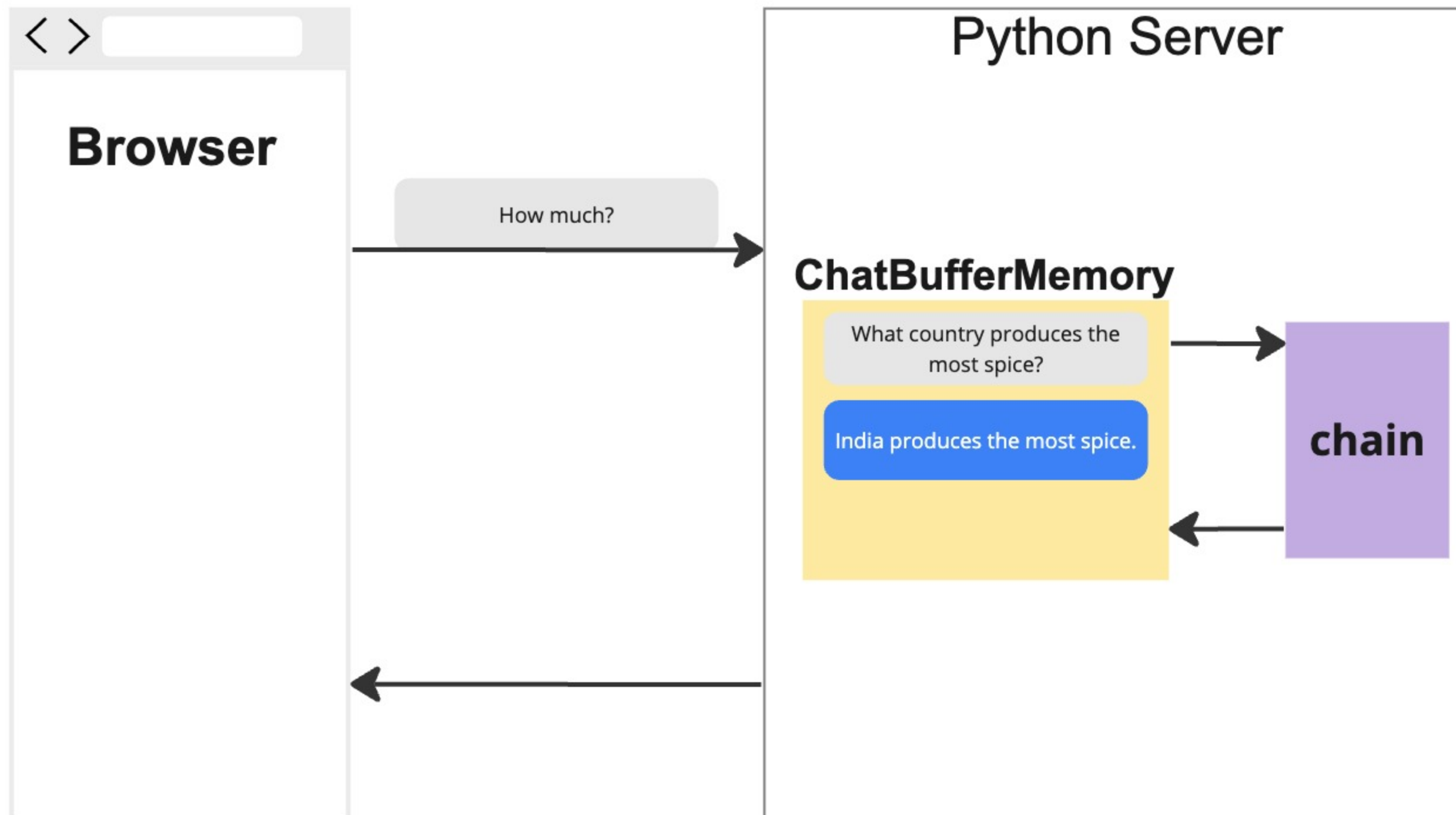
Store messages on Python server in a ChatBufferMemory



Need to persist messages + conversations so they can be used by both web + chat

Option 2

Store messages on Python server in a ChatBufferMemory



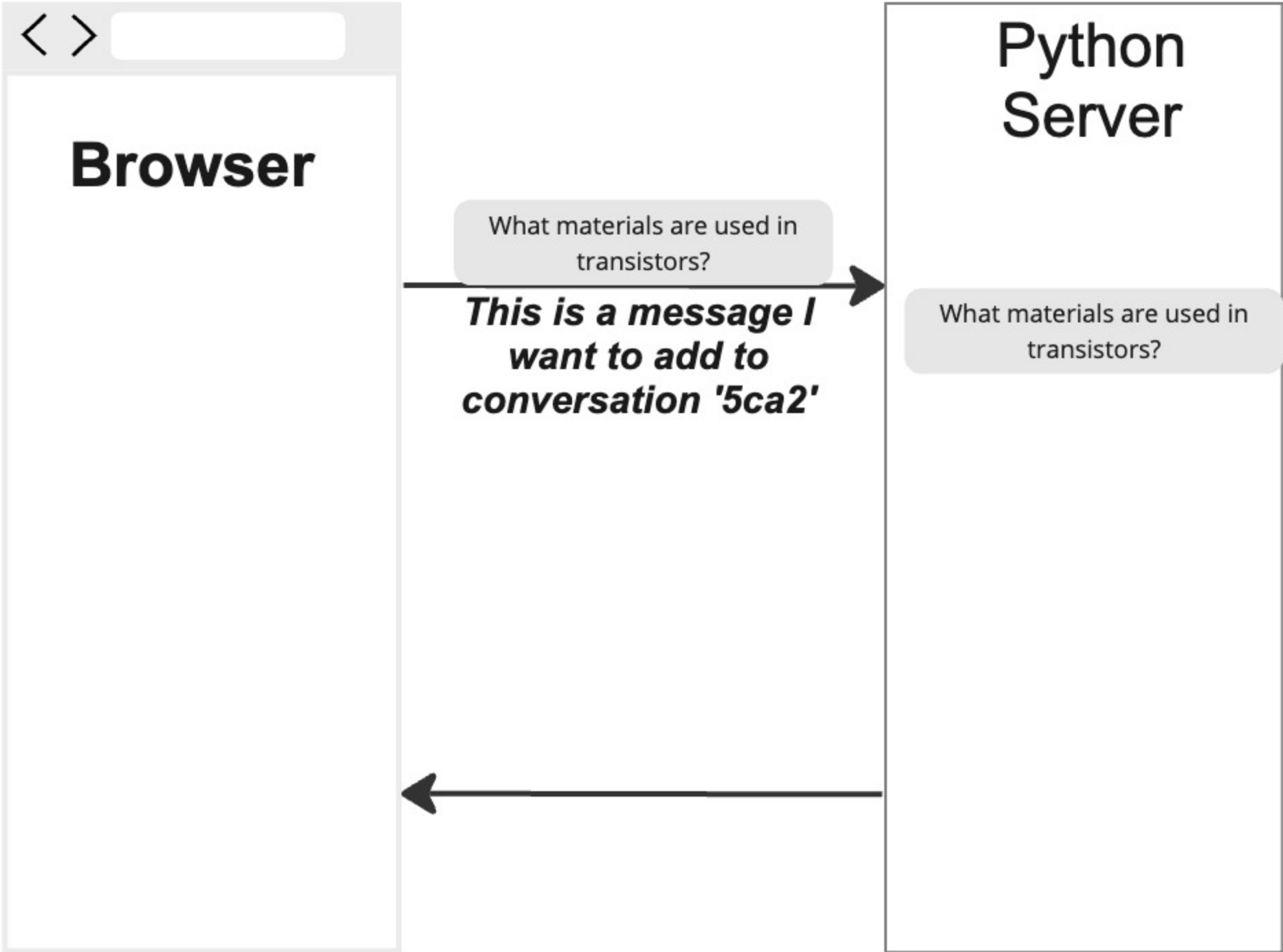
ChatBufferMemory stores messages in a list

If the server gets restarted, the messages are lost!

Need to persist messages + conversations so they can be used by both web + chat

Option 3

Store messages in a database



SQLite Database
Table of Messages

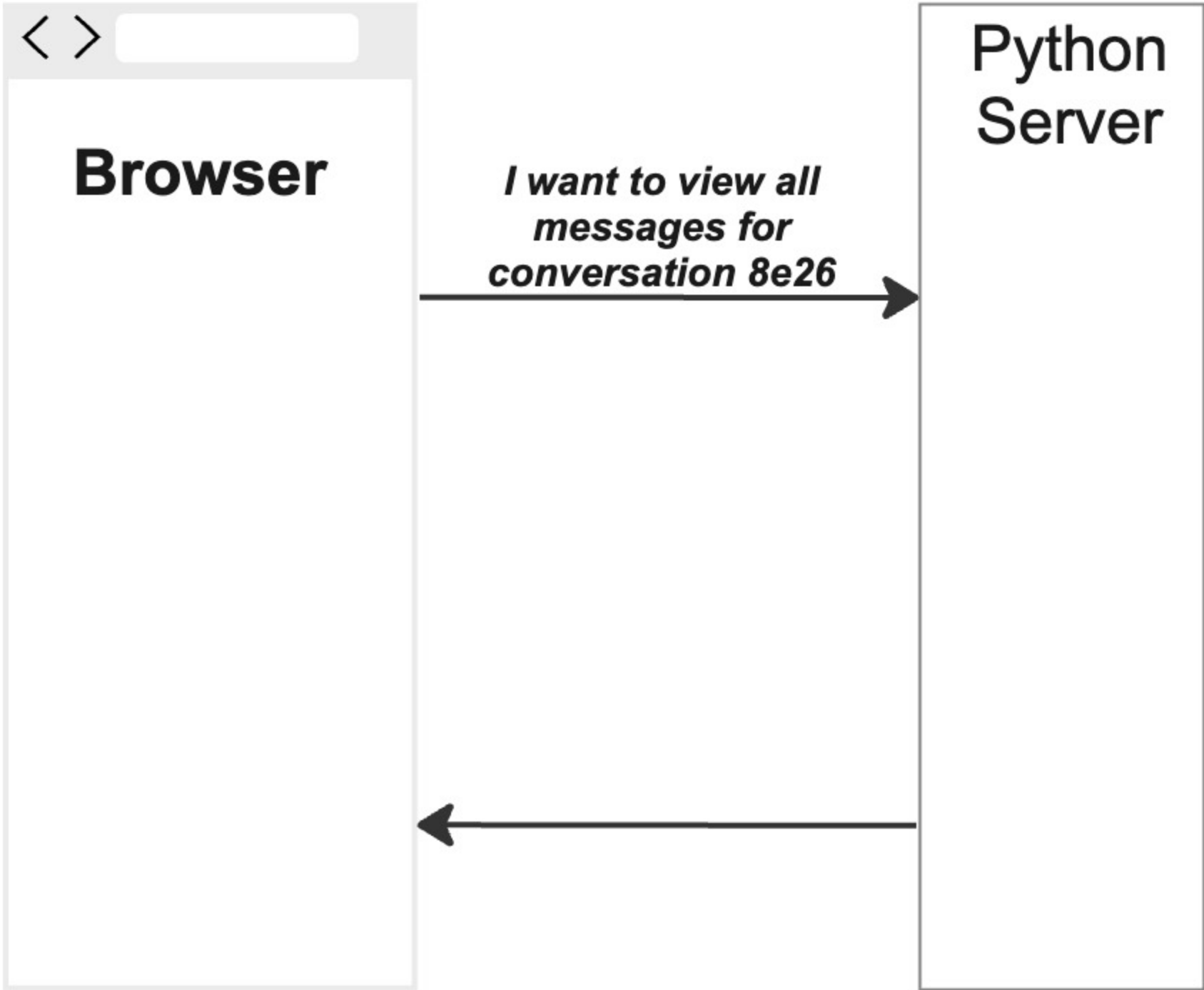
| ID | conversation_id | role | content |
|-----------|-----------------|-------|-----------------------------------------|
| be17bac7 | 8e26 | human | What country produces... |
| ccb58e26 | 8e26 | ai | India produces the most... |
| 3a495ca2 | 8e26 | human | How much? |
| 9f78186e | 5ca2 | human |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

Need to persist messages + conversations so they can be used by both web + chat

Option 3

Store messages in a database

Easy to show a list of messages without involving memory, chains, etc



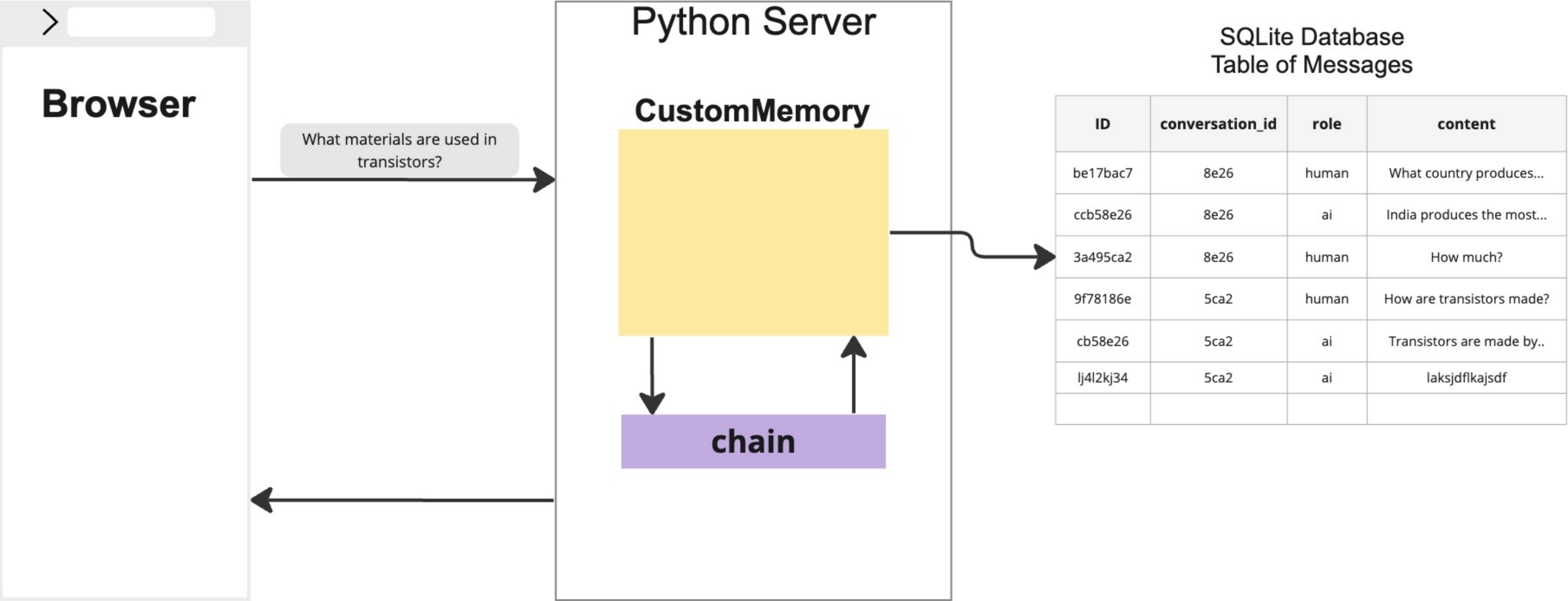
SQLite Database
Table of Messages

| 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.. |
| | | | |
| | | | |

Need to persist messages + conversations so they can be used by both web + chat

Option 3
Store messages in a database

We can make a custom memory to use the database



LangChain

Dynamodb

Chat Message History

Mongodb

Chat Message History

Cassandra

Chat Message History

Redis

Chat Message History

Postgres

Chat Message History

SQL

Chat Message History

Streamlit

Chat Message History

Zep

Chat Message History

Langchain has some tools to automatically store messages in a variety of different databases

Up to you, but I'd avoid this

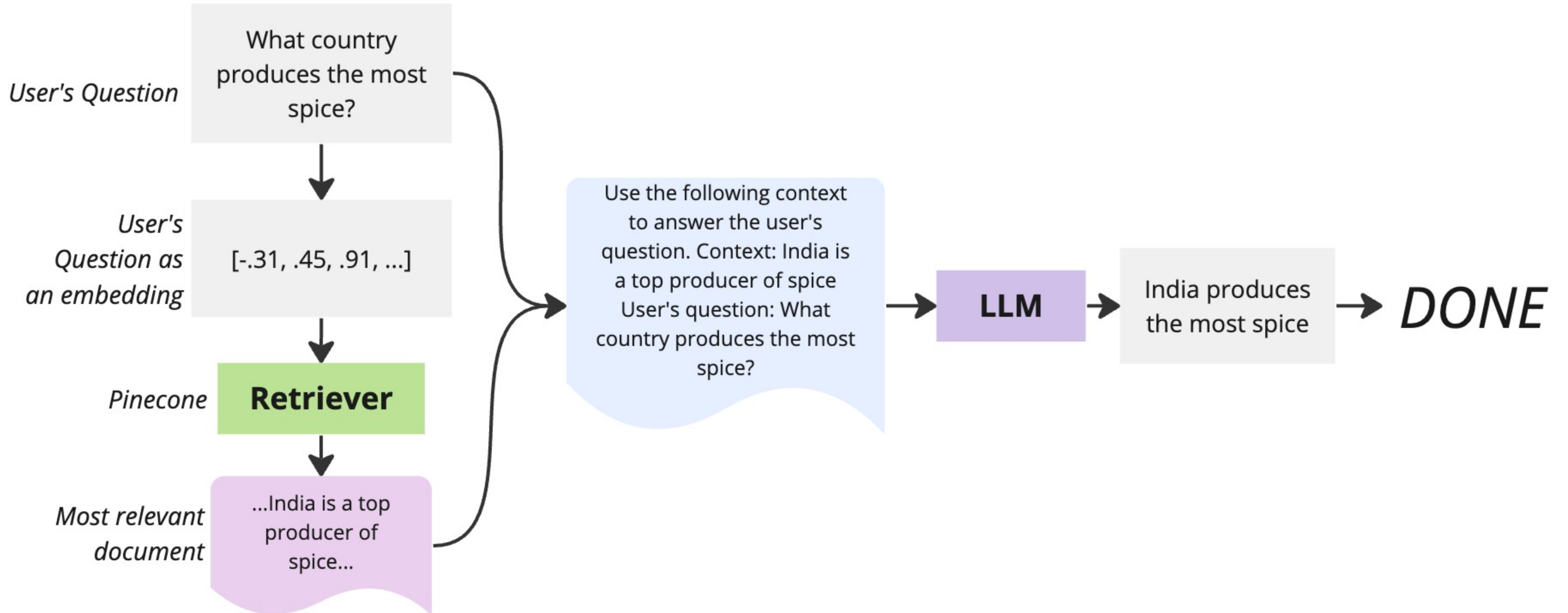
Exactly how are the messages being stored?

How do I access them in non-Langchain parts of my app?

Need to handle vague user messages

Option 1

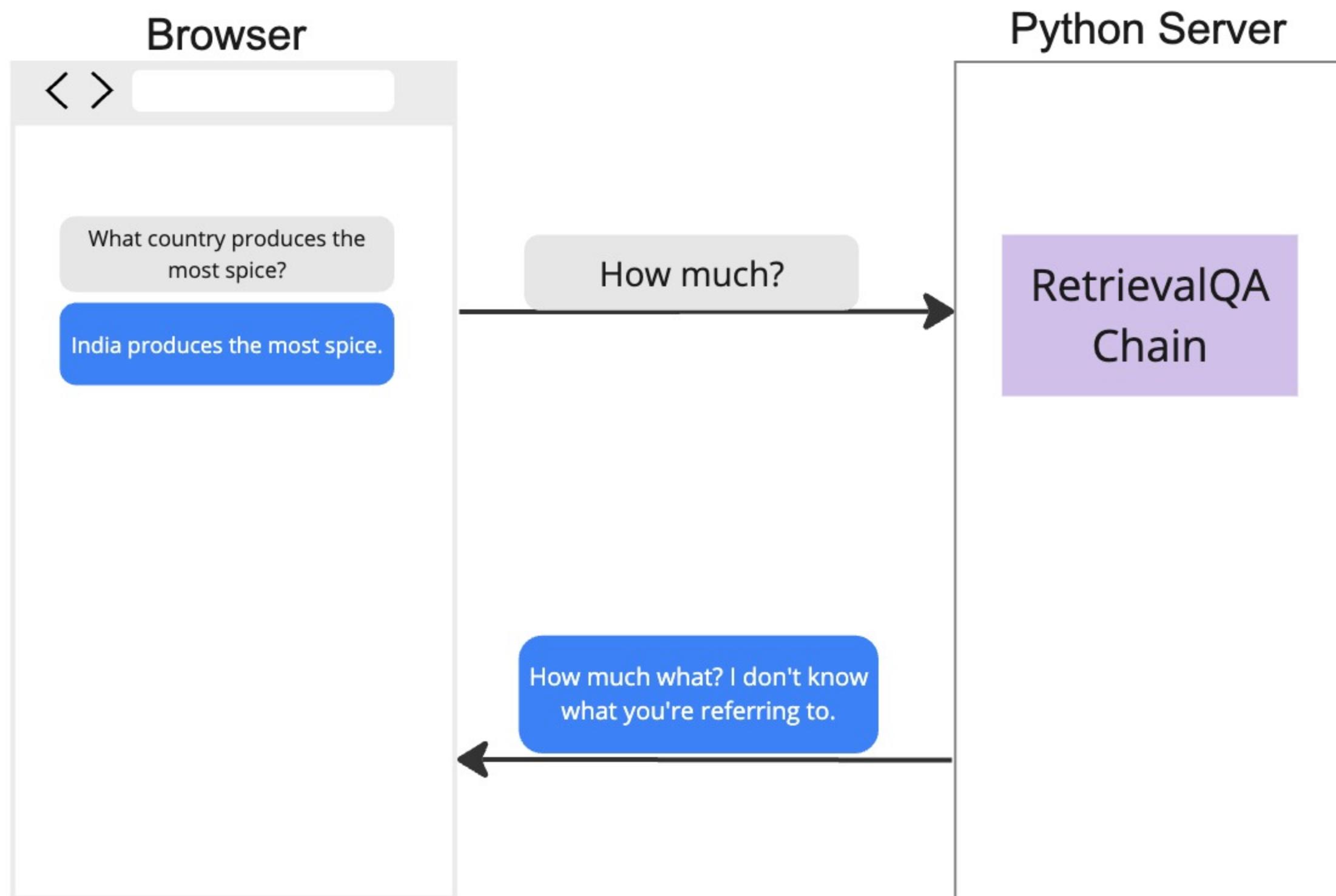
RetrievalQA Chain



Need to handle vague user messages

Option 1

RetrievalQA Chain

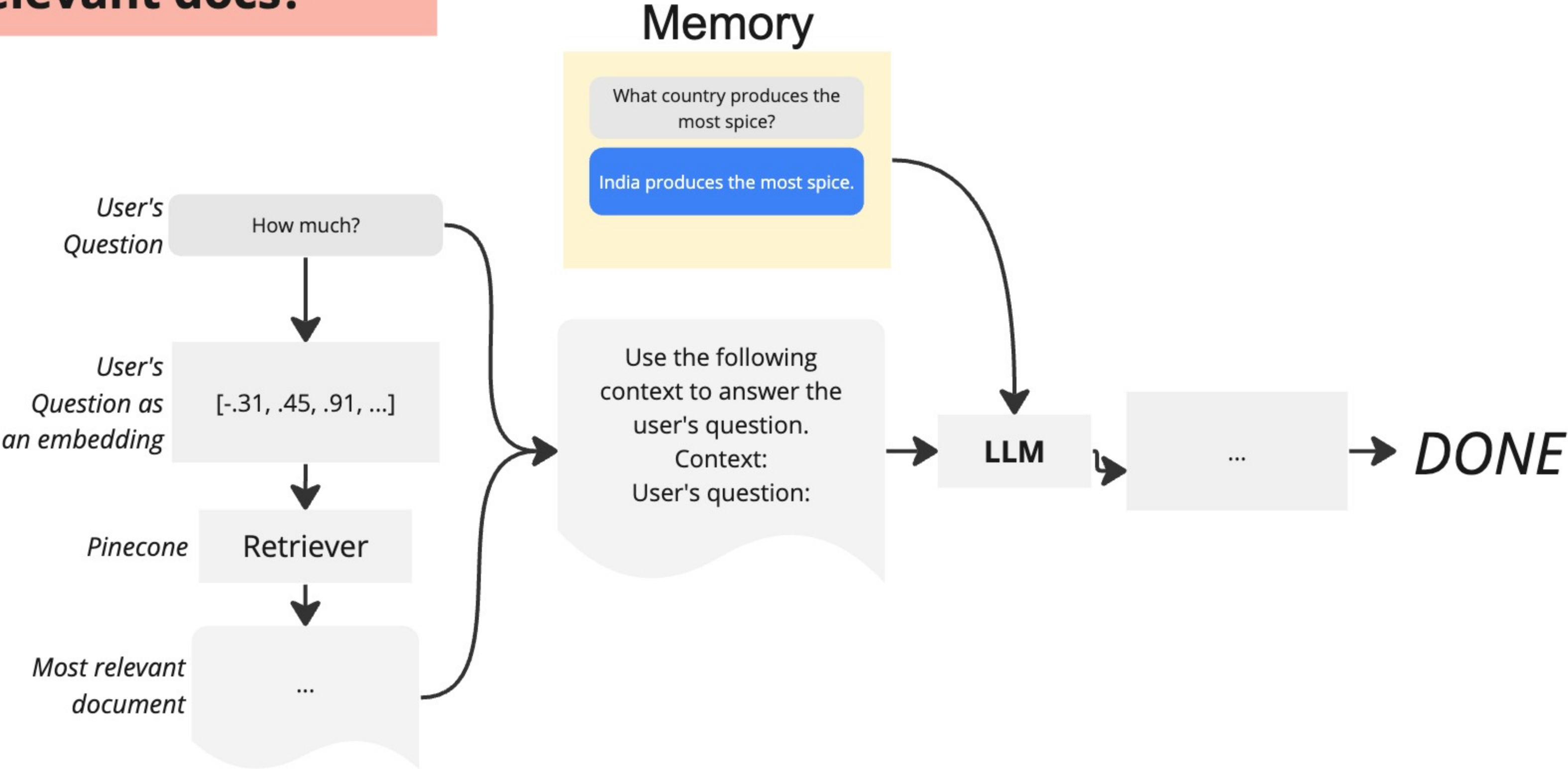


**RetrievalQA chain
doesn't use memory!**

Even if we did add in memory to RetrievalQA, how would we find relevant docs?

Need to handle vague user messages

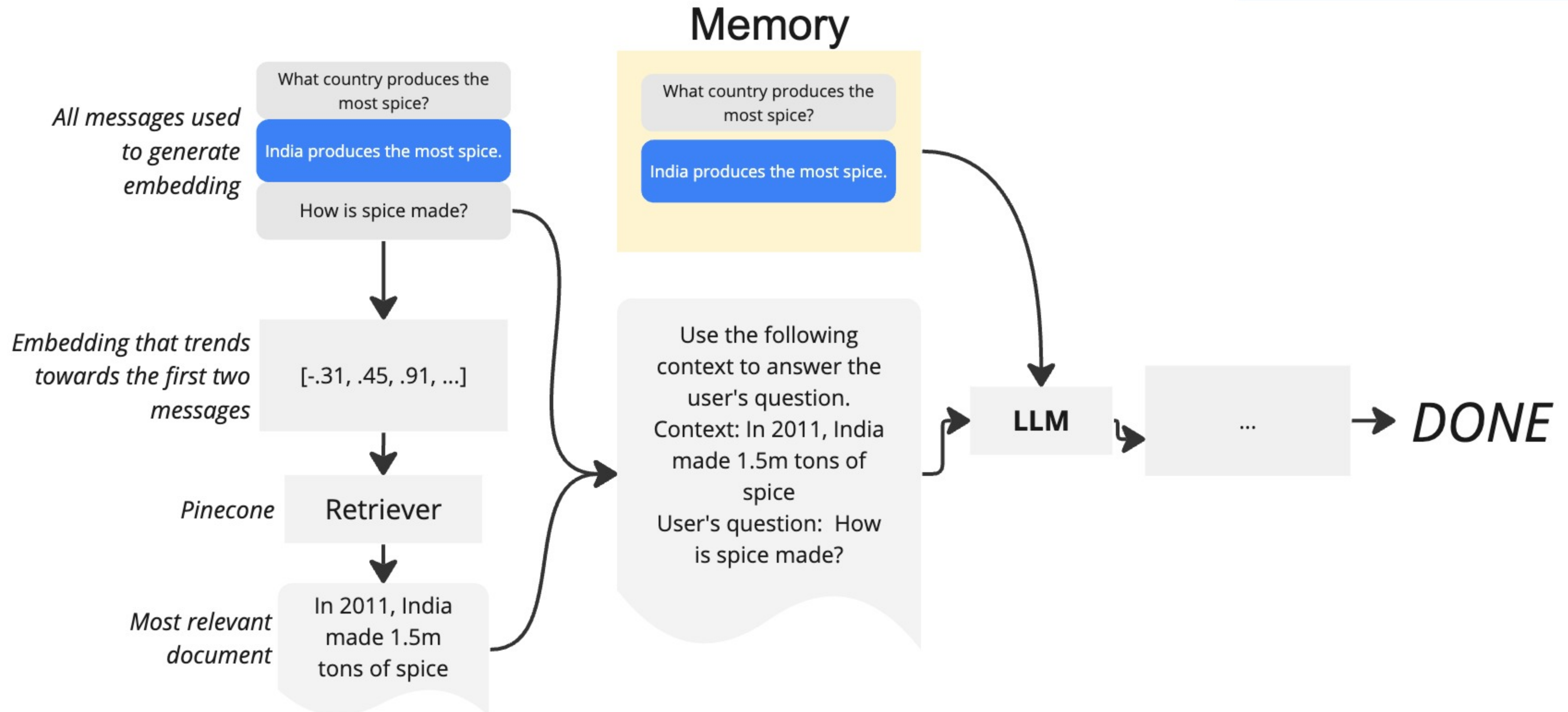
Option 1
RetrievalQA Chain



Combining all messages together to generate an embedding might not work

Need to handle vague user messages

Option 1 RetrievalQA Chain



Conversational Retrieval Chain

Memory

**Condense
Question Chain**



**Combine Docs
Chain**

Input

What country produces the most spice?

Conversational Retrieval Chain

First message of conversation

Memory

What country produces the most spice?

India produces the most spice

Condense Question Chain

If memory is empty, skip me.

What country produces the most spice?

[-.31, .45, .91, ...]

Retriever

India is a top spice producer.

Combine Docs Chain

Use the following context to answer the user's question.
Context: India is a top spice producer.
User's question: What country produces the most spice?

LLM

India produces the most spice

Input

How much?

Conversational Retrieval Chain After first message of conversation

Memory

What country produces the most spice?

India produces the most spice

Condense Question Chain

Prompt

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?
Standalone question:

LLM

Exactly how much spice does India produce?

Refined Question

Combine Docs Chain

Exactly how much spice does India produce?

[-.31, .45, .91, ...]

Retriever

in 2011 india produced 1.5m tons of spice

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

LLM

India produced 1.5m tons in 2011.

