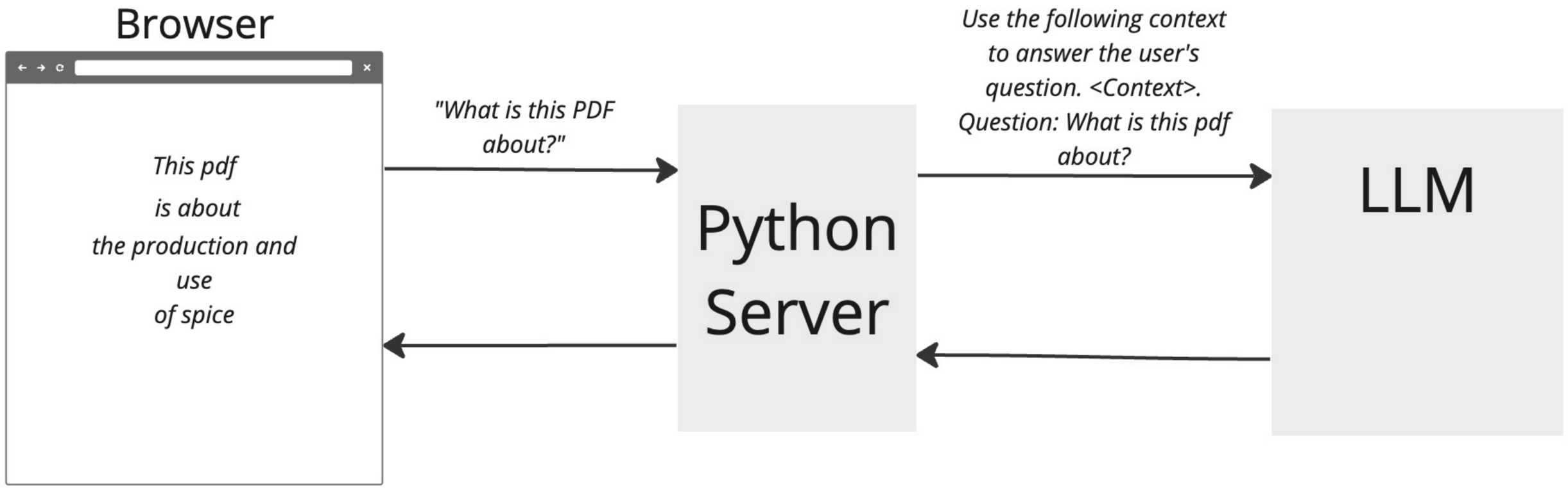
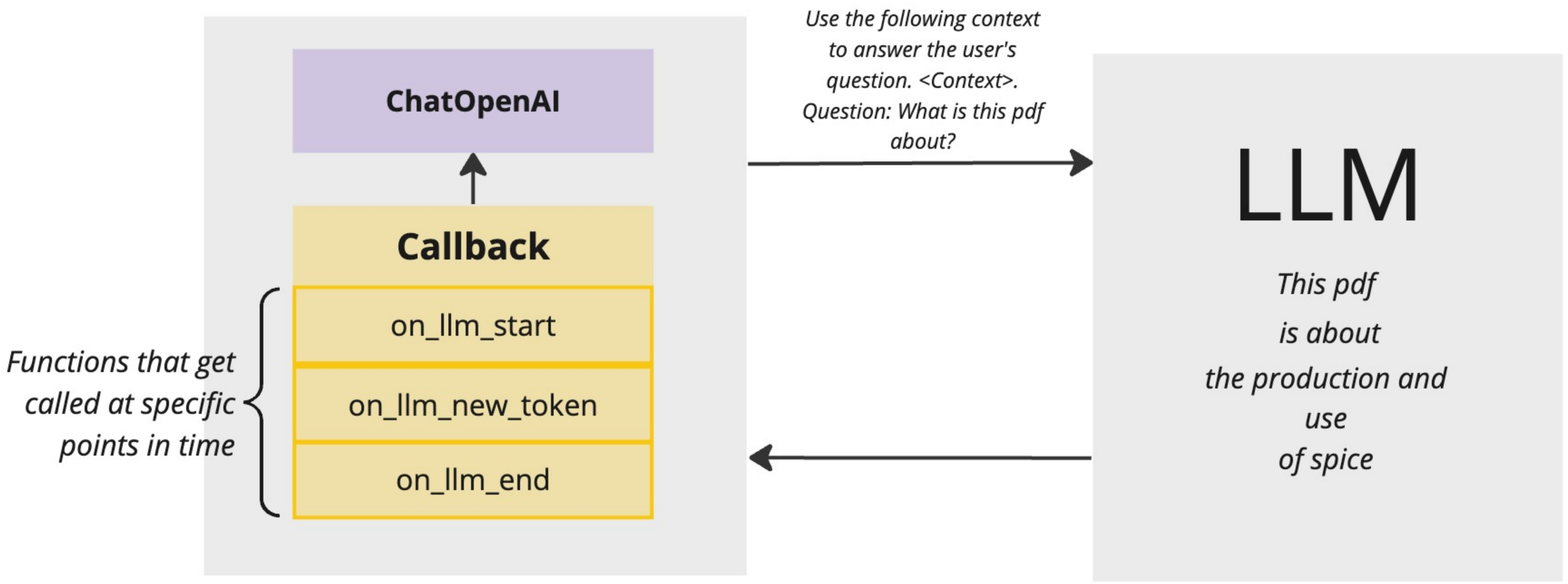
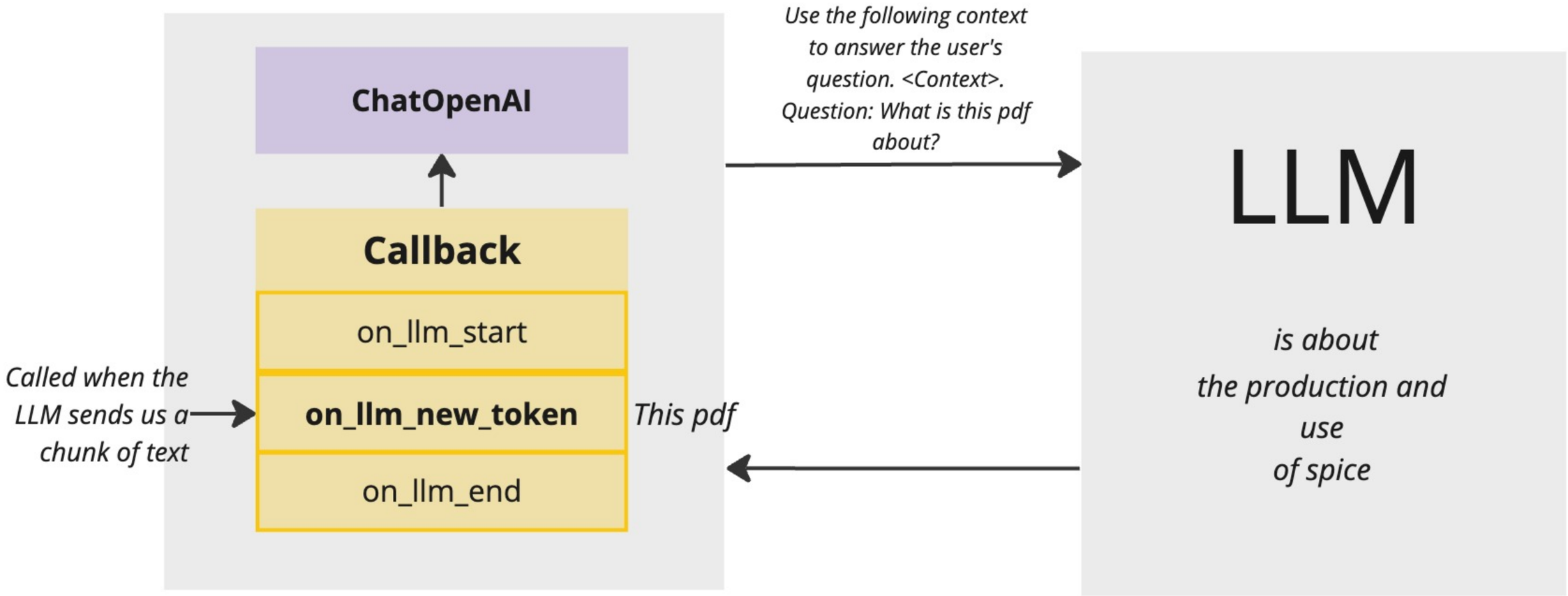

Browser
This pdf
is about
the production and
use
of spice
"What is this PDF about?"
Python Server
Use the following context to answer the user's question. <Context>. Question: What is this pdf about?
LLM

**Our app will support both sync + streaming text generation (just for demo purposes)**

Need to figure out how to tell LangChain if we want to stream generation or not


ChatOpenAI
Callback
on_llm_start
on_llm_new_token
on_llm_end
Functions that get called at specific points in time
Use the following context to answer the user's question. <Context>. Question: What is this pdf about?
LLM
This pdf is about the production and use of spice


ChatOpenAI
Callback
on_llm_start
on_llm_new_token
on_llm_end
Called when the LLM sends us a chunk of text
This pdf
Use the following context to answer the user's question. <Context>. Question: What is this pdf about?
LLM
is about the production and use of spice

## Result of **first** question

*Printed text appears to be identical to the output*

**Good!**

```
India
 produces
 the
 most
 spices
,
 contributing
 to

75
%
 of
 global
 spice
 production
.
```

## Result of **second** question

*The question????*

**???**

```
How
 much
 spice
 does
 India
 produce
?
```

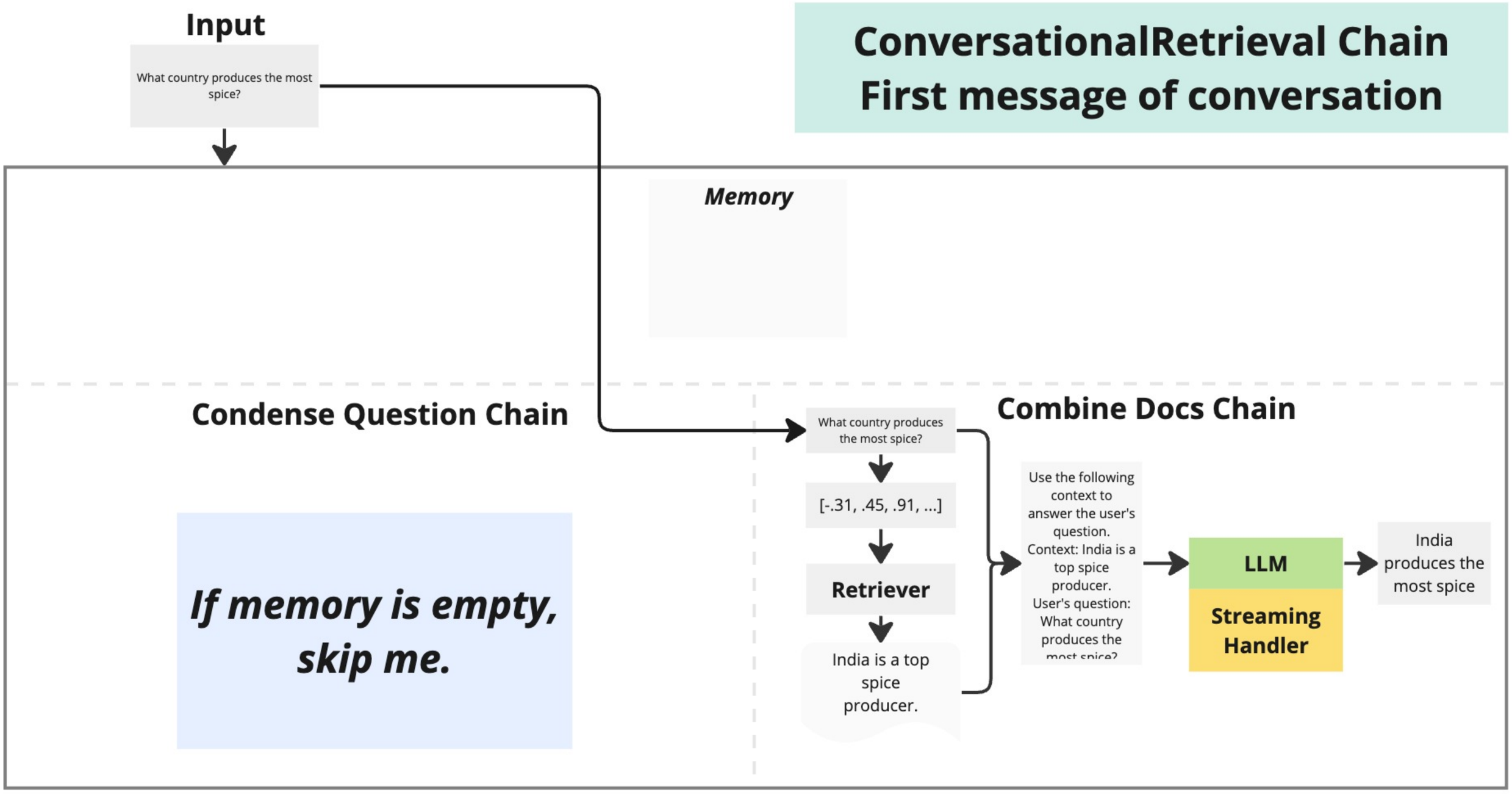
*Second group of prints appear to be the answer*

**Good!**

```
India
 produces

1
,
525
,
000
 metric
 tonnes
 of
 spices
 in

201
1
,
 according
 to
 the
 given
 information
.
```


Input
What country produces the most spice?
ConversationalRetrieval Chain
First message of conversation
Memory
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
LLM
Streaming Handler
India produces the most spice

# ConversationalRetrieval Chain
# After first message of conversation

## Input

How much?

***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**

**Streaming Handler**

Exactly how much spice does India produce?

*Refined Question*

Exactly how much spice does India produce?

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

Retriever

in 2011 india produced 1.5m tons of spice

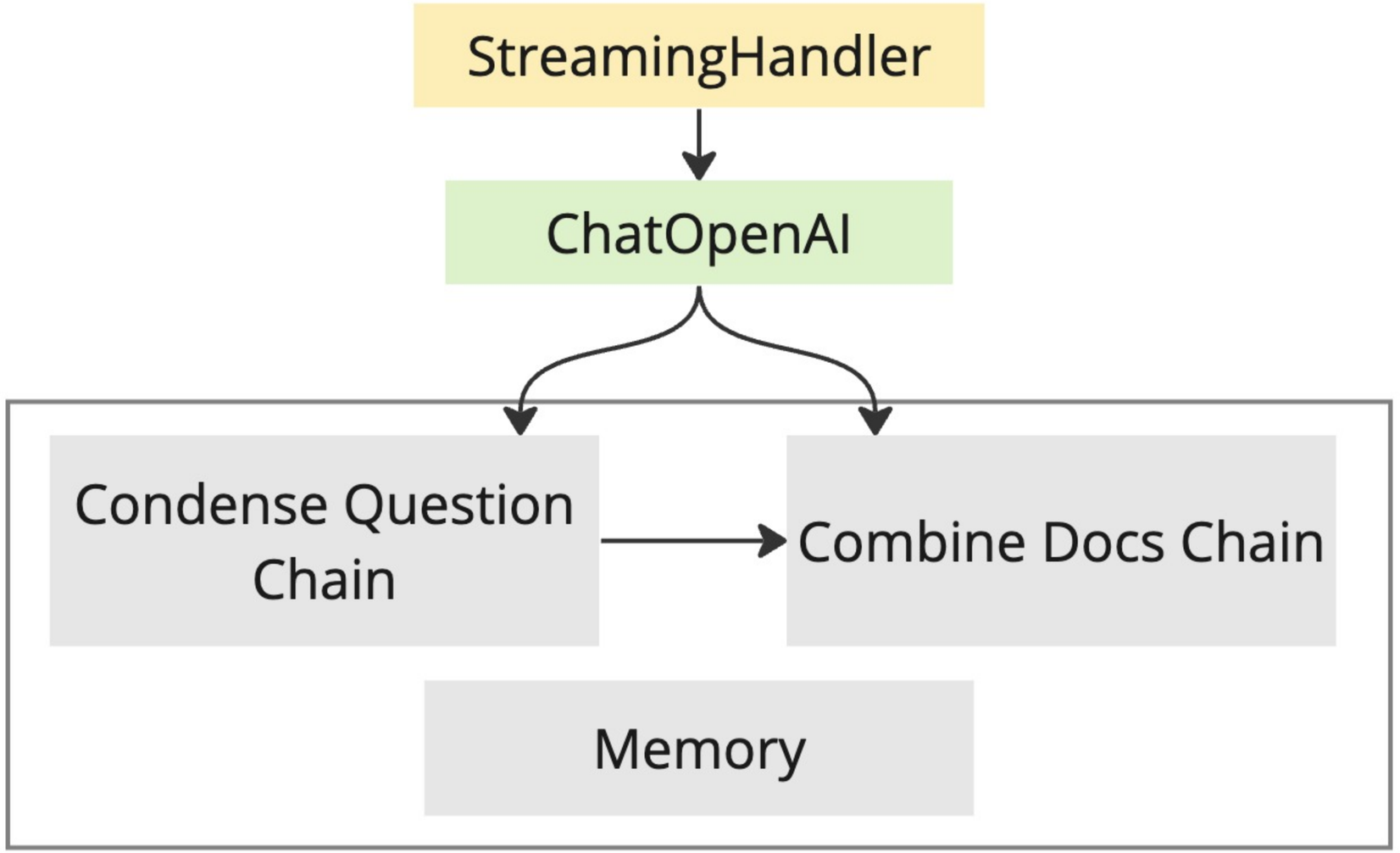
## Combine Docs Chain

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

**LLM**

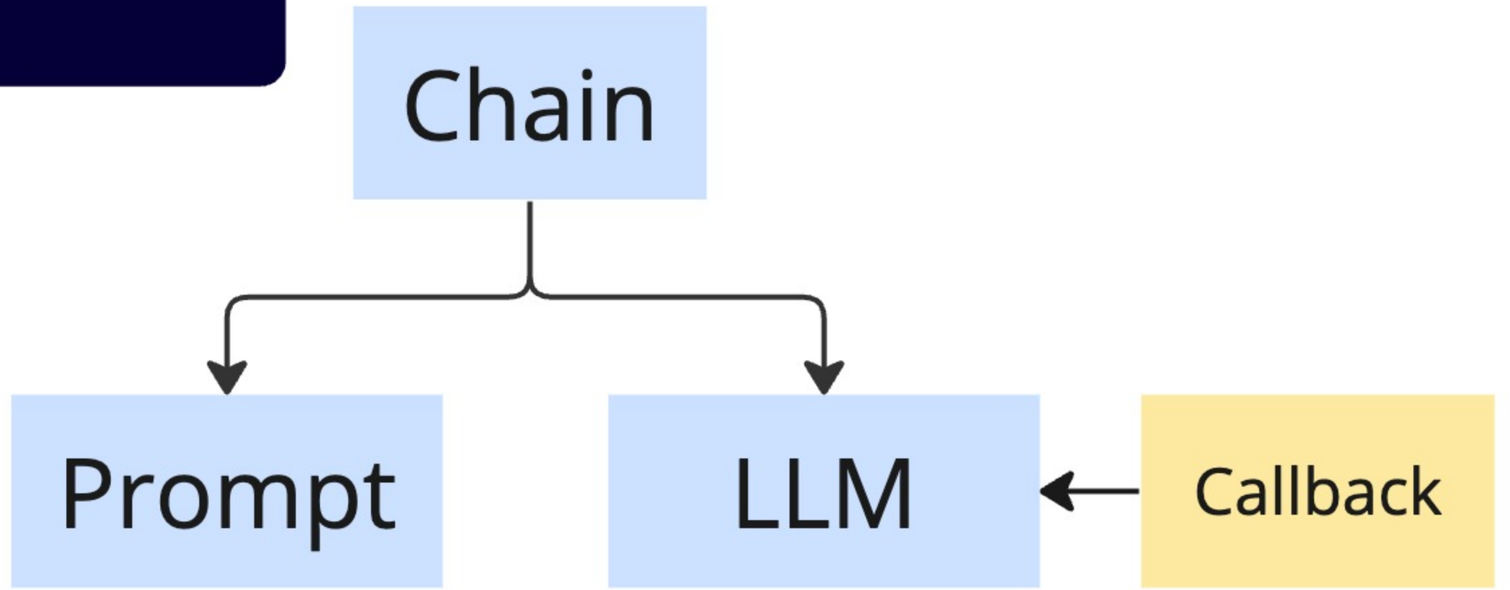
**Streaming Handler**

India produced 1.5m tons in


StreamingHandler
ChatOpenAI
Condense Question Chain
Combine Docs Chain
Memory

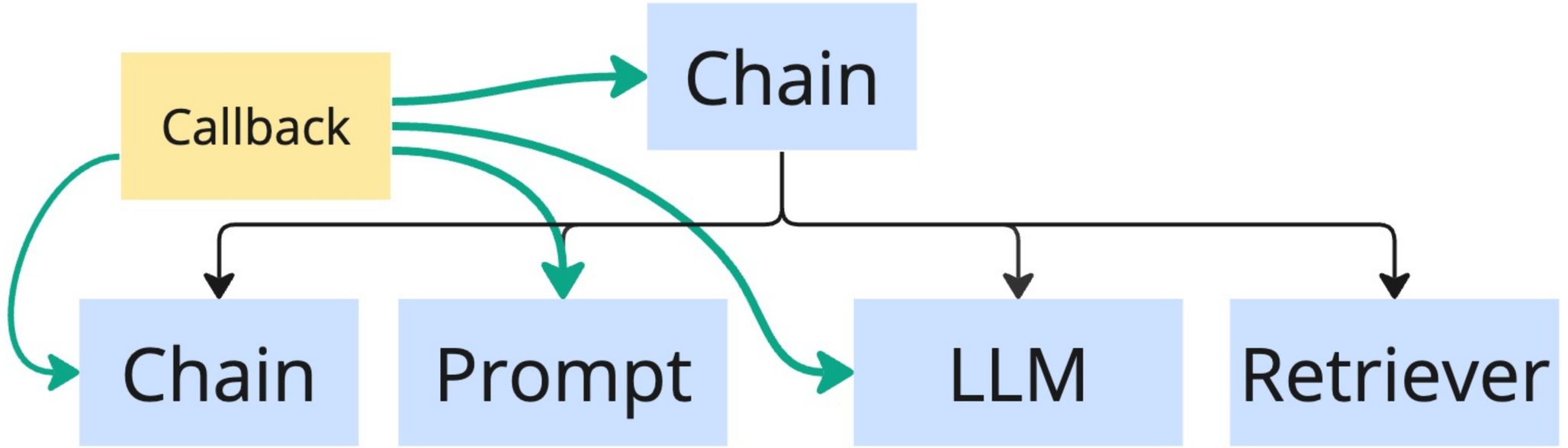
## Callback Applied at Creation

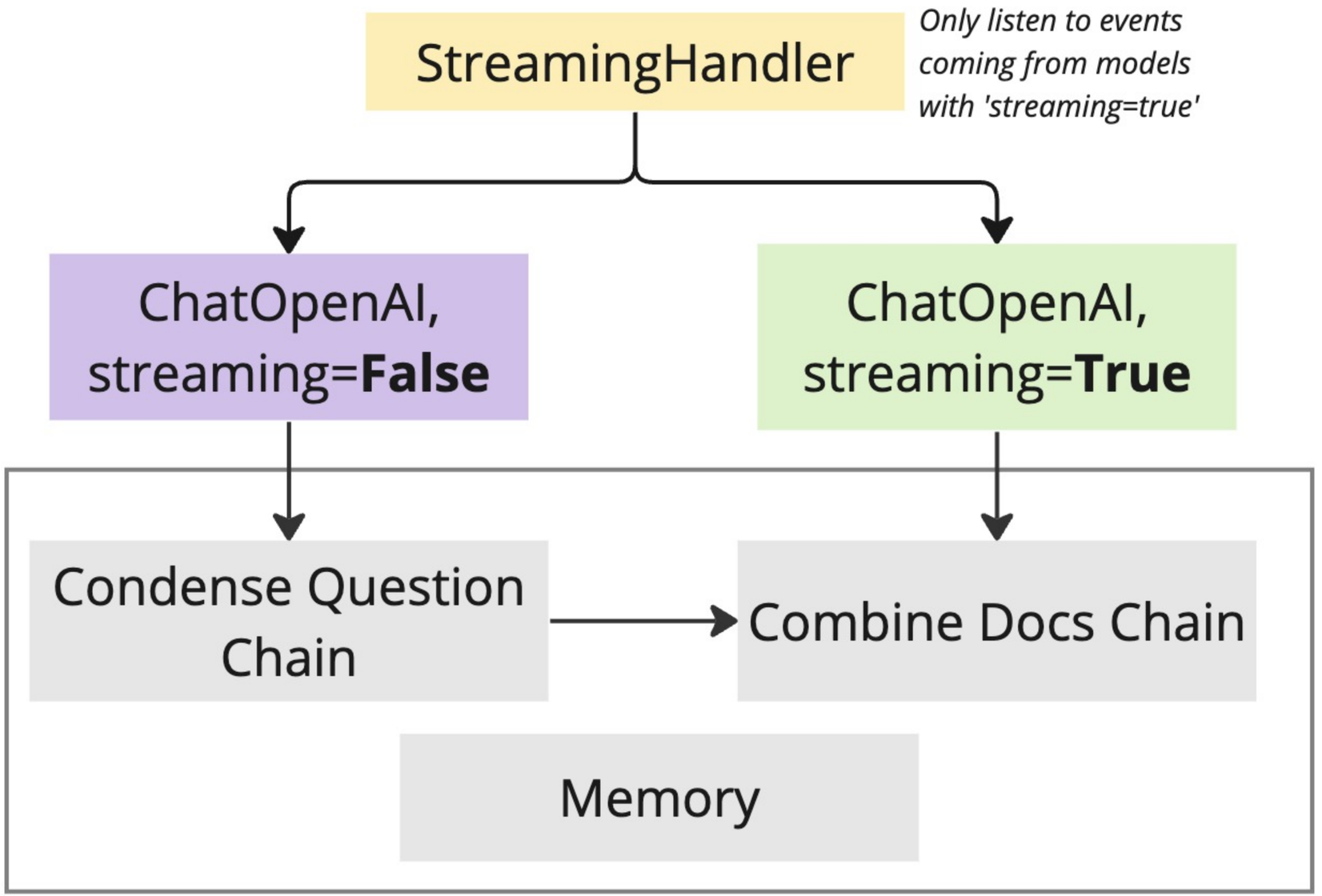
```
1
2     llm = ChatOpenAI(callbacks=[
3         MyCallback()
4     ])
5     chain = LLMChain(
6         llm=llm,
7         prompt=prompt
8     )
```

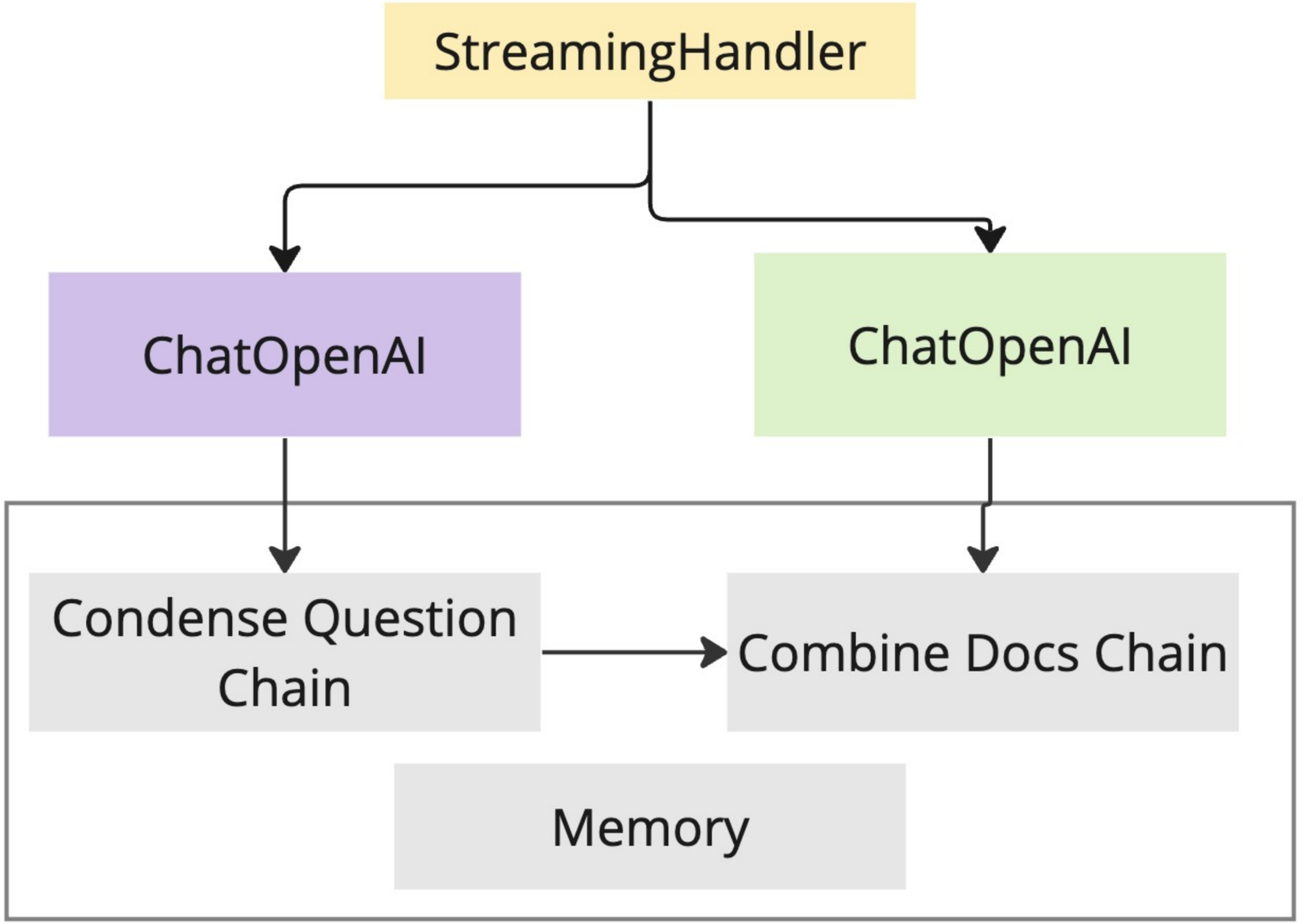


## Callback Applied at Call

```
result = chain(
    "What country...",
    callbacks=[MyCallback()]
)
```




StreamingHandler
Only listen to events coming from models with 'streaming=true'
ChatOpenAI, streaming=False
ChatOpenAI, streaming=True
Condense Question Chain
Combine Docs Chain
Memory


StreamingHandler
ChatOpenAI
ChatOpenAI
Condense Question Chain
Combine Docs Chain
Memory