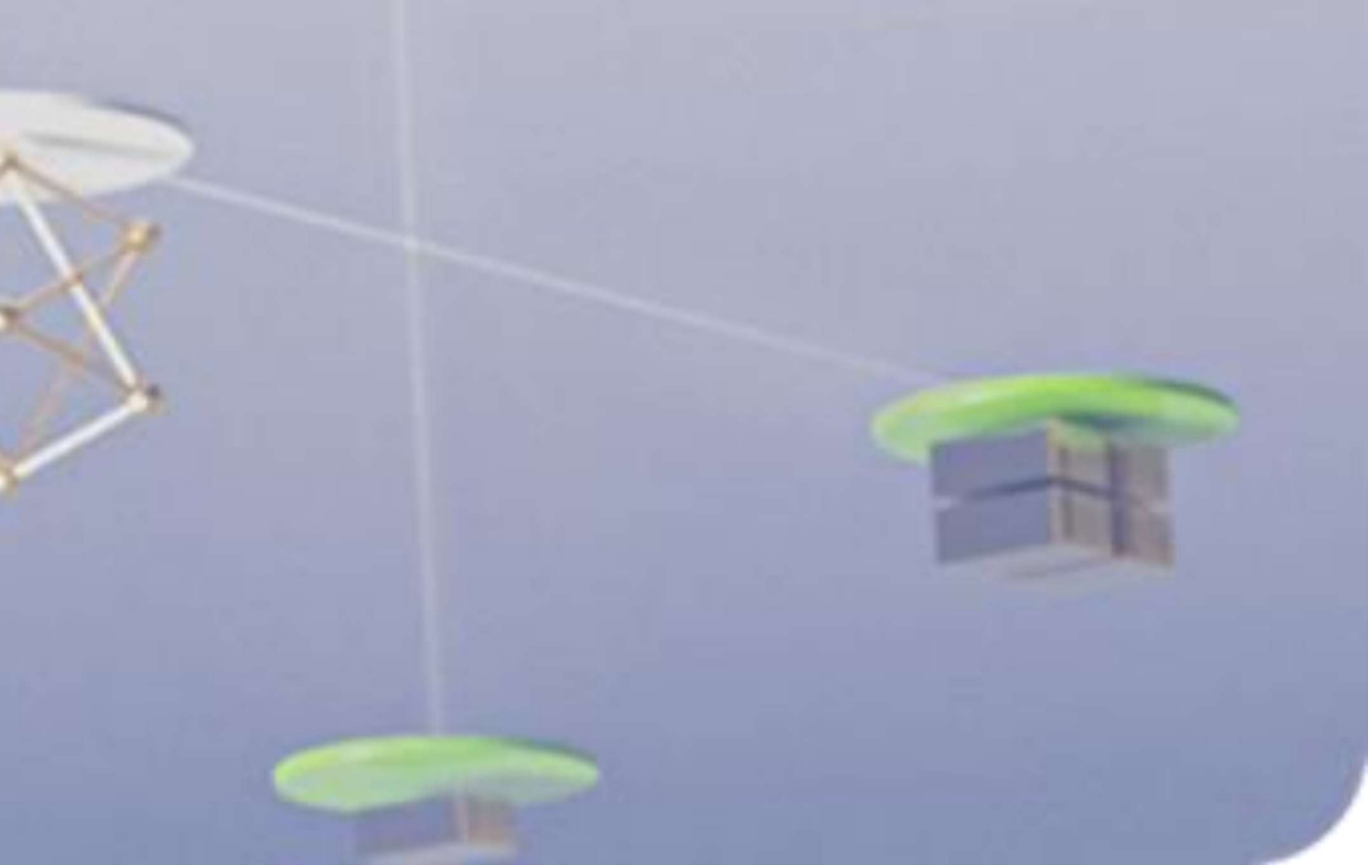
Building RAG A

Building RAG A



Retrieval Augmented Generation

Prerequisites

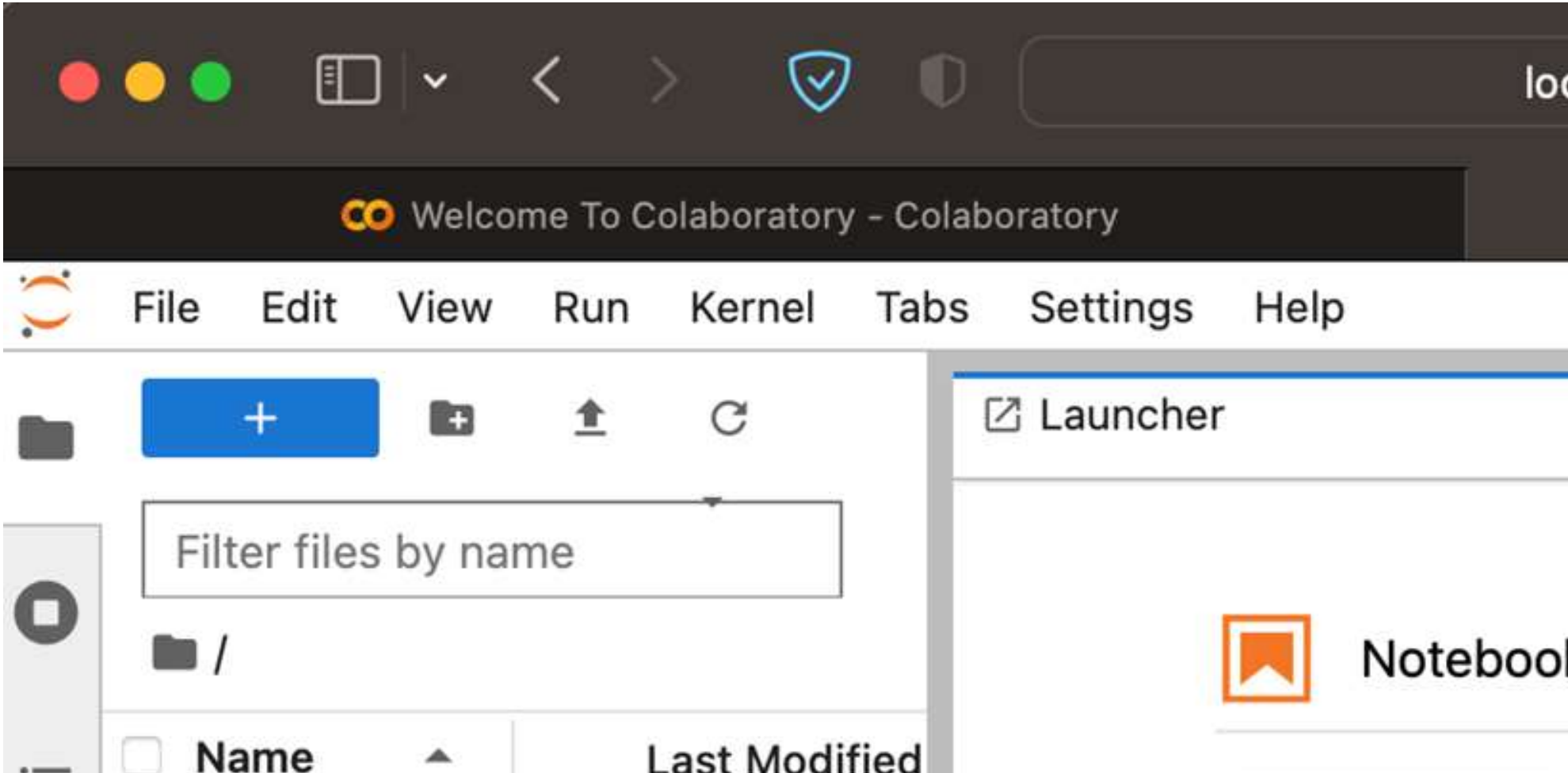
RAG Agents in Product

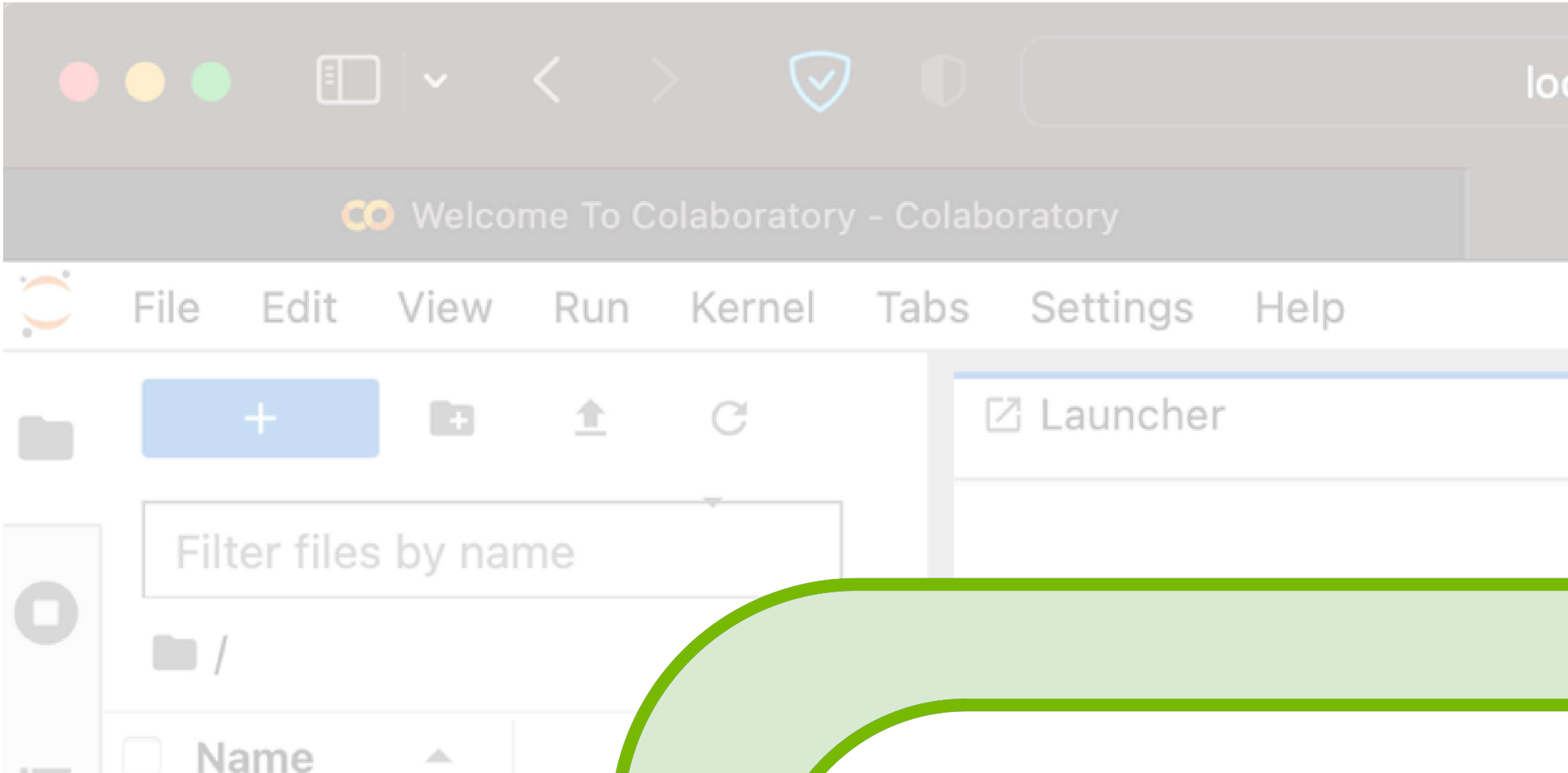
- Prior LLM/LangChain

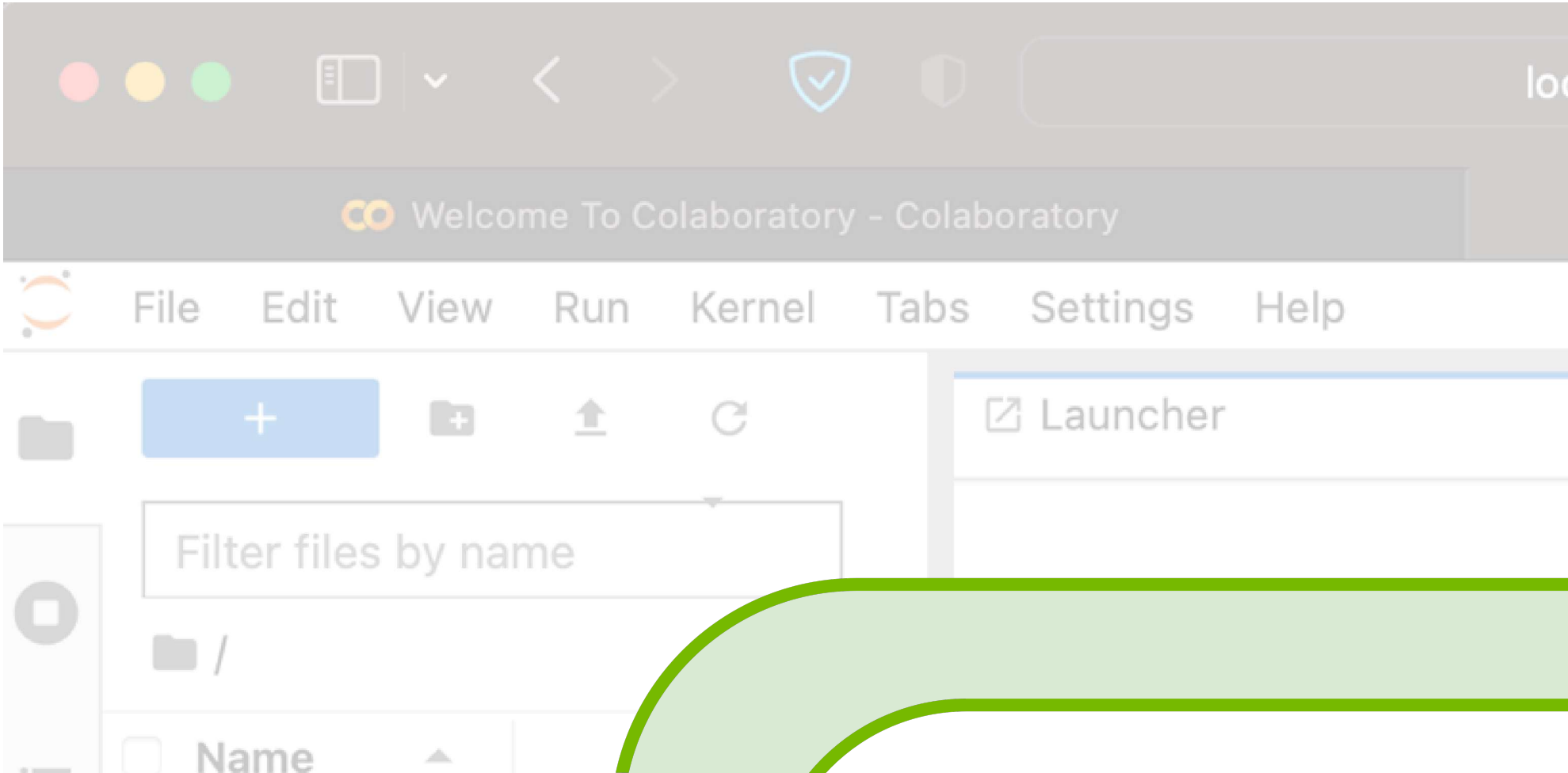
Course Objecti

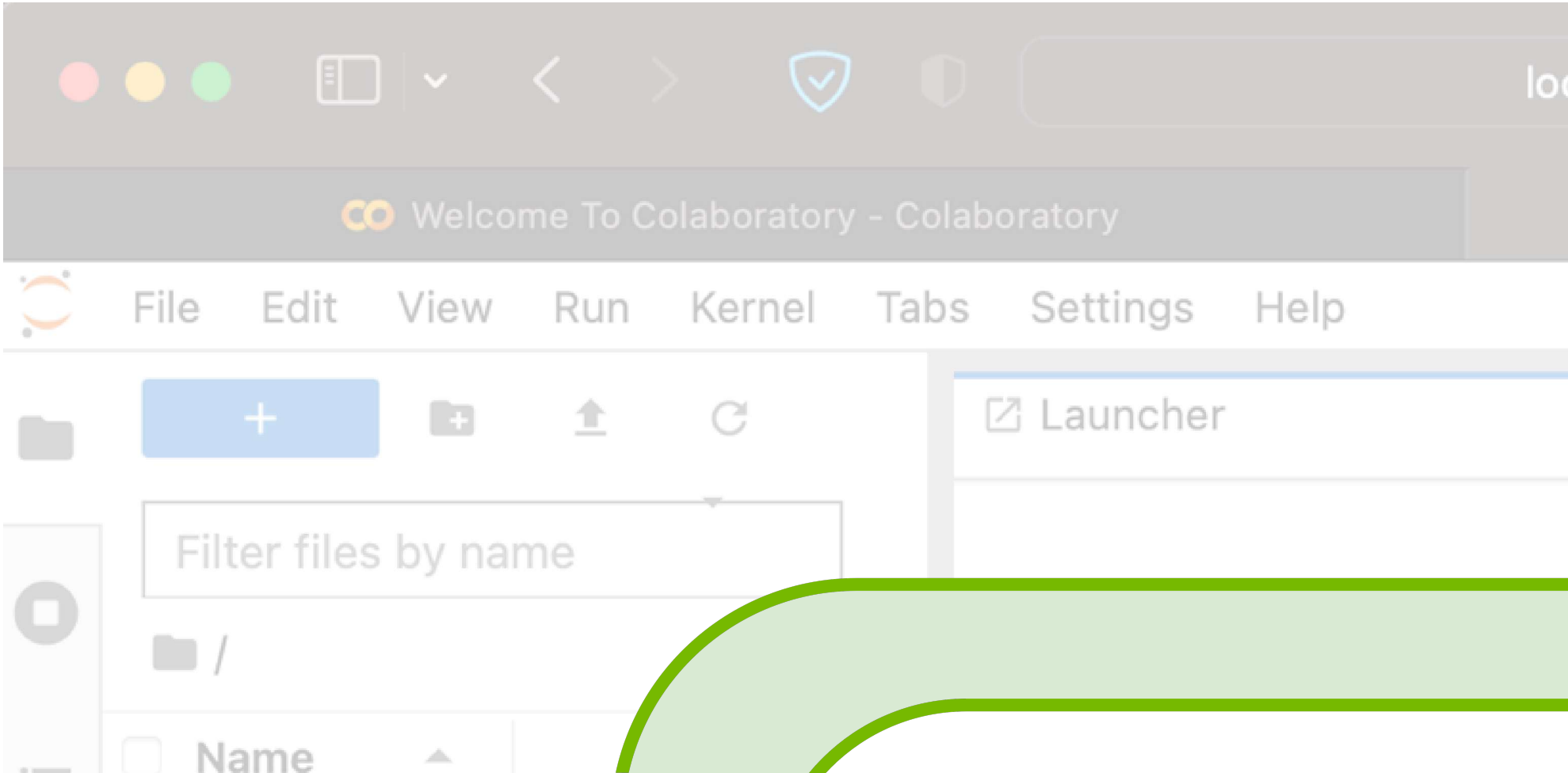
RAG Agents in Product

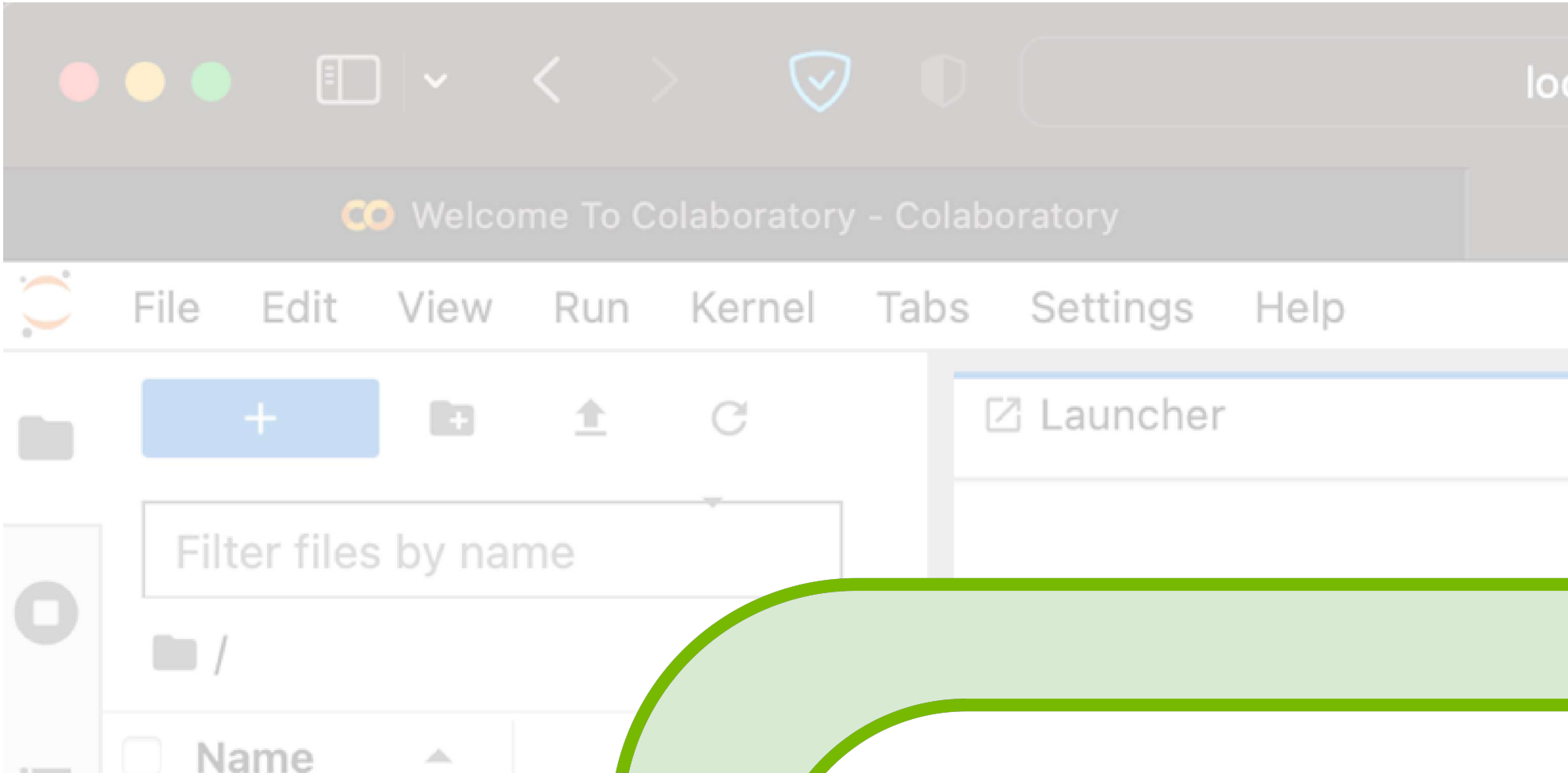
Building RAG A

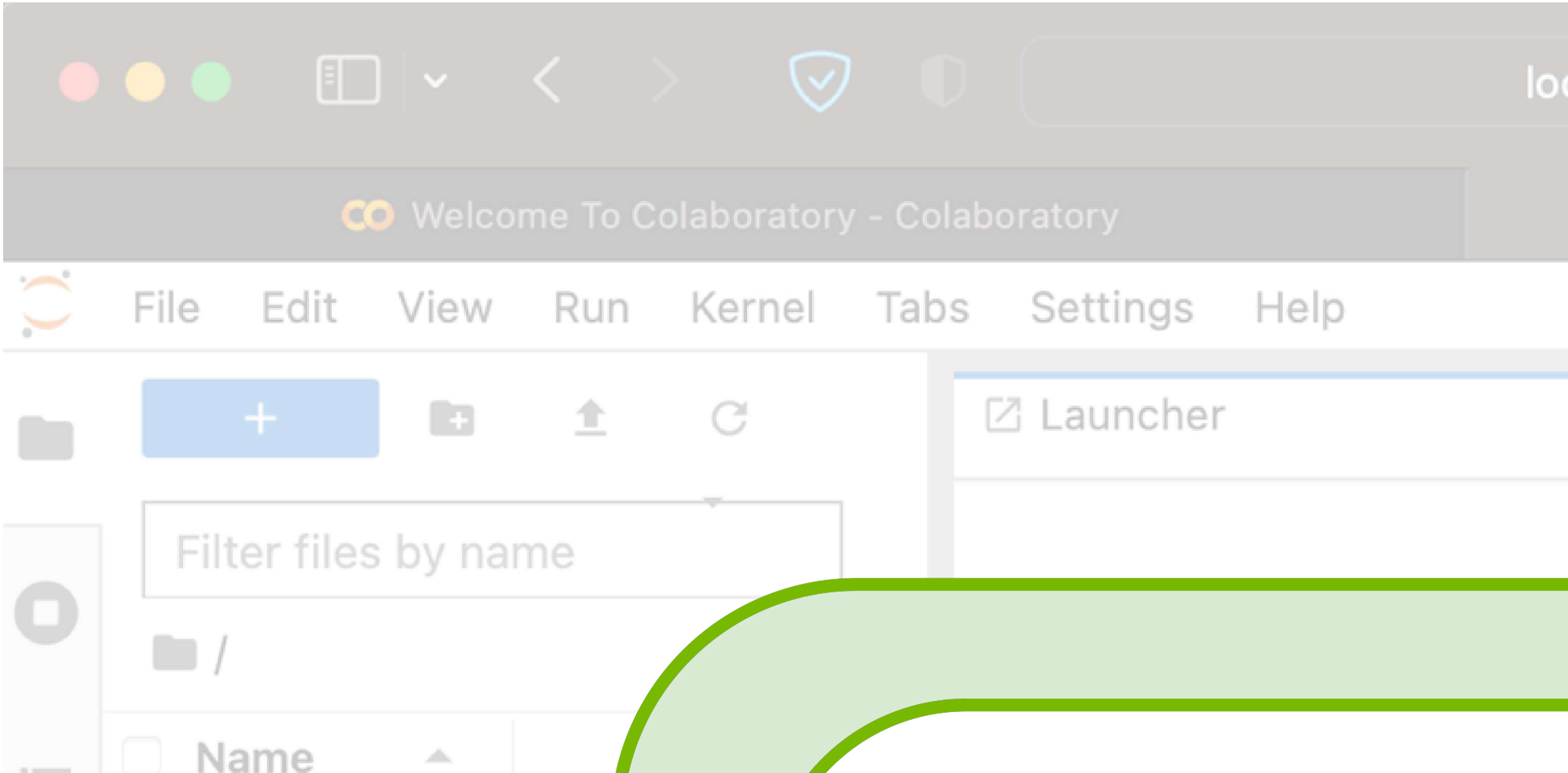

Welcome To Colaboratory - Colaboratory
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Launcher
Filter files by name
/
Name
Last Modified

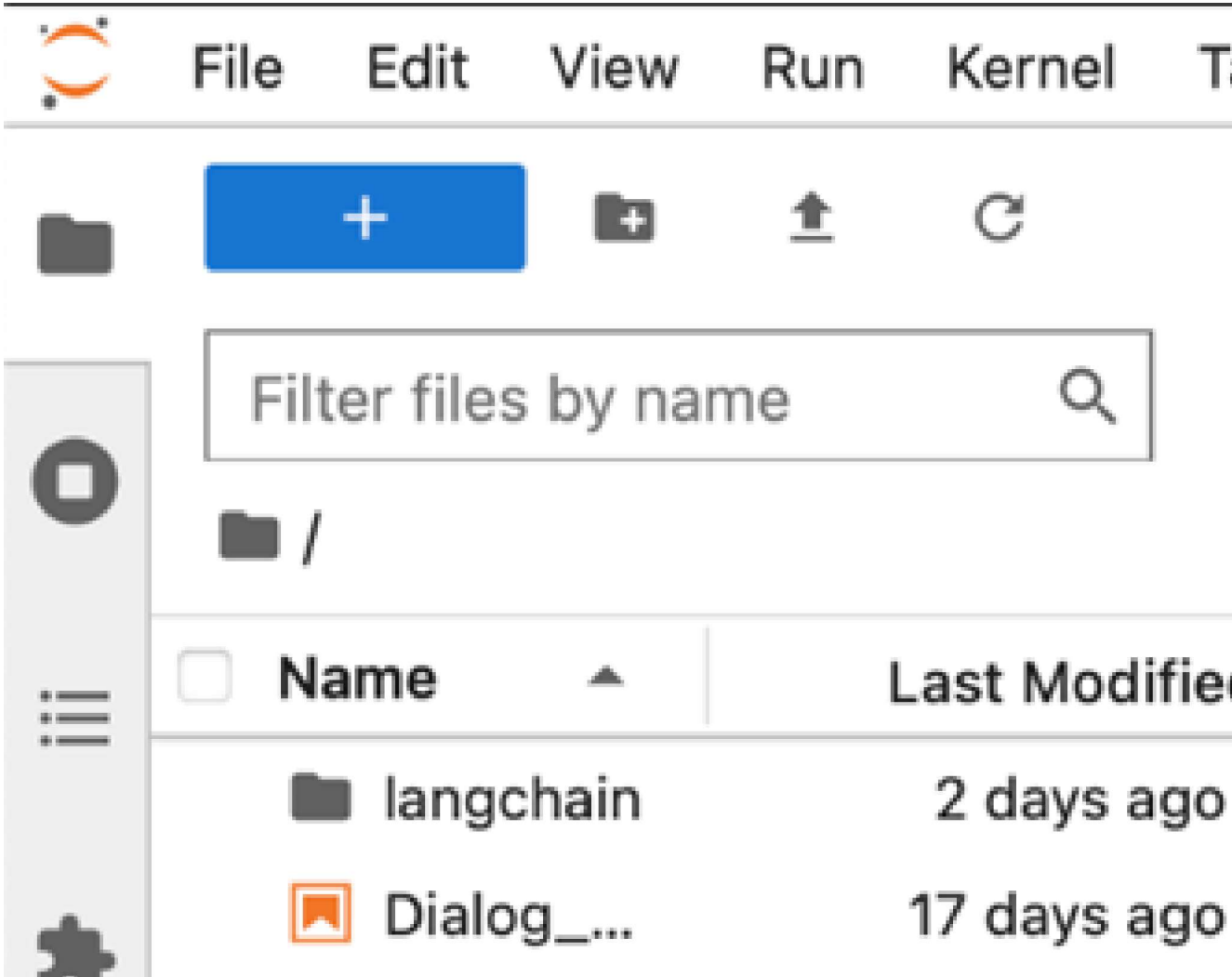

Welcome To Colaboratory - Colaboratory
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Launcher
Filter files by name
/
Name

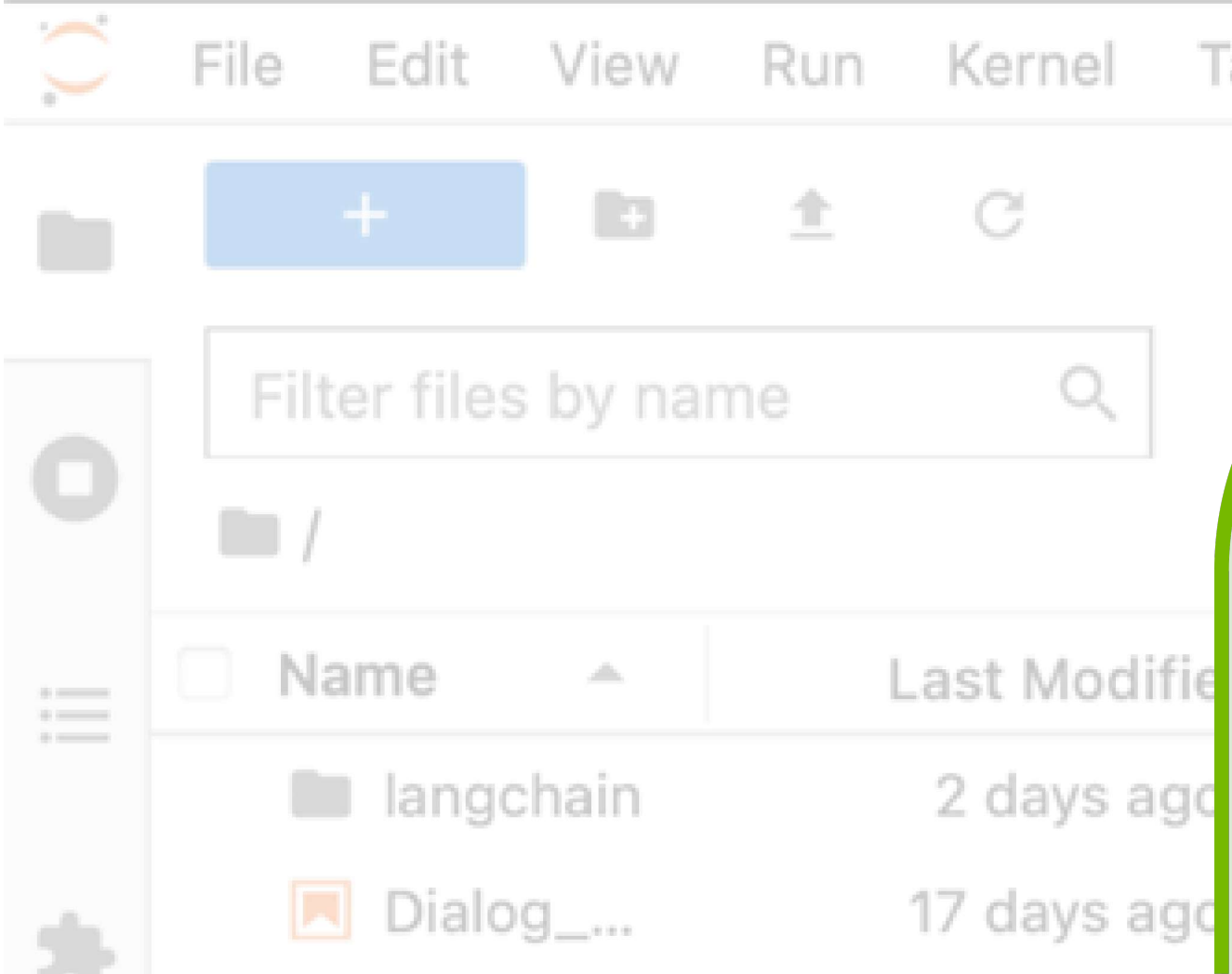

Welcome To Colaboratory - Colaboratory
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Launcher
Filter files by name
/
Name

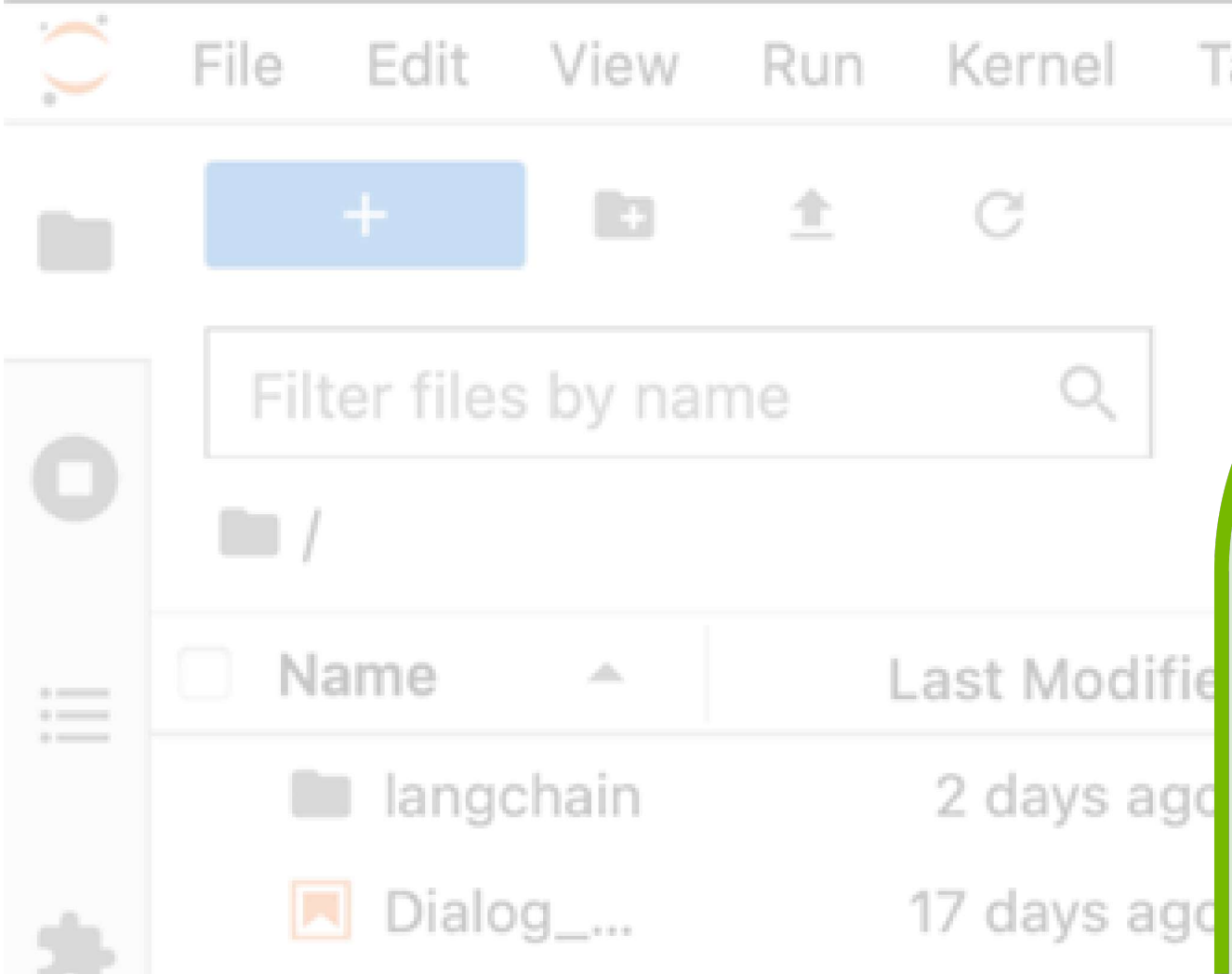

Welcome To Colaboratory - Colaboratory
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Launcher
Filter files by name
/
Name

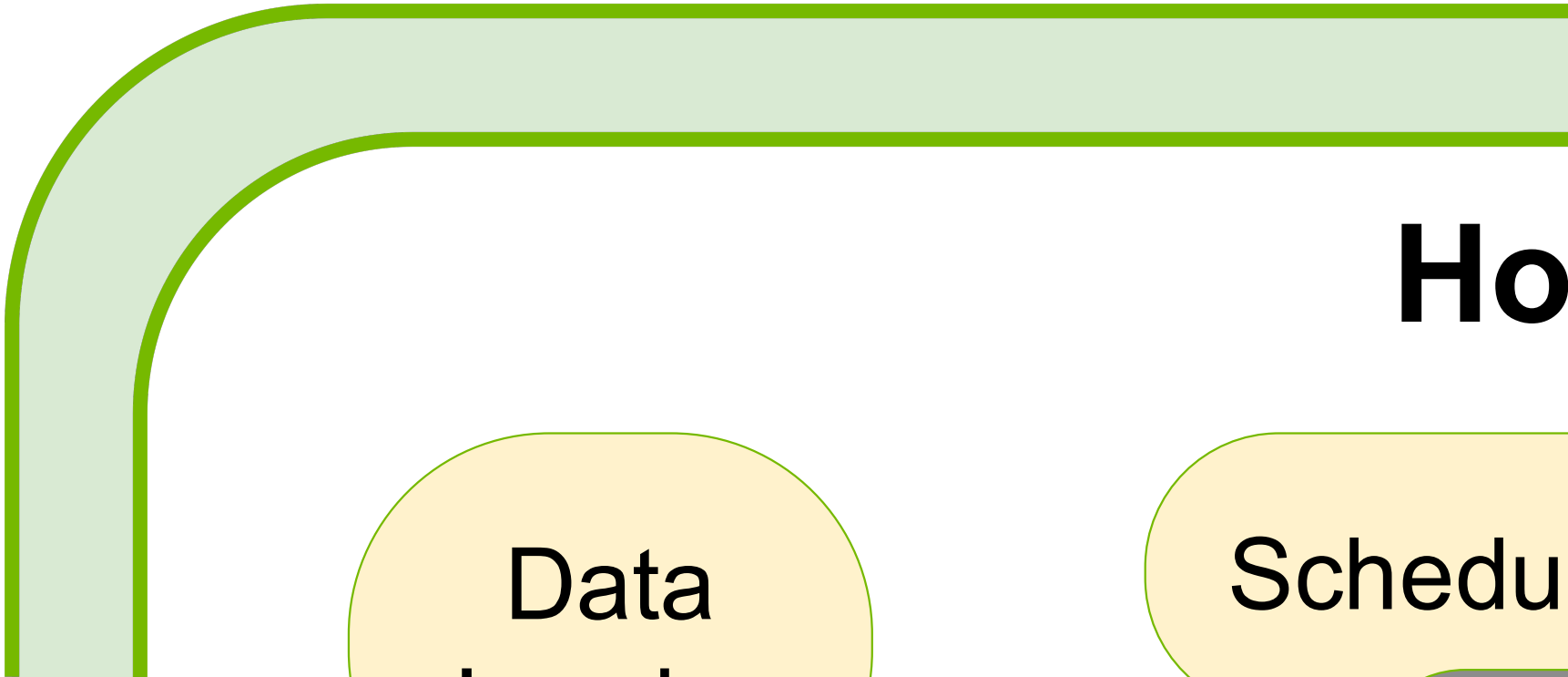

Welcome To Colaboratory - Colaboratory
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Launcher
Filter files by name
/
Name


Welcome To Colaboratory - Colaboratory
File
Edit
View
Run
Kernel
Tabs
Settings
Help
Launcher
Filter files by name
/
Name


File
Edit
View
Run
Kernel
Filter files by name
/
Name
Last Modifie
langchain
2 days ago
Dialog_...
17 days ago


File
Edit
View
Run
Kernel
Filter files by name
/
Name
Last Modifie
langchain
2 days ago
Dialog_...
17 days ago


File
Edit
View
Run
Kernel
Filter files by name
/
Name
Last Modifie
langchain
2 days ago
Dialog_...
17 days ago


C
Ho
Data
Schedu

1. Allocate Resources
2. Define Services

1. Allocate Resources
2. Define Services

1. Allocate Resources
2. Define Services

1. Allocate Resources
2. Define Services

S

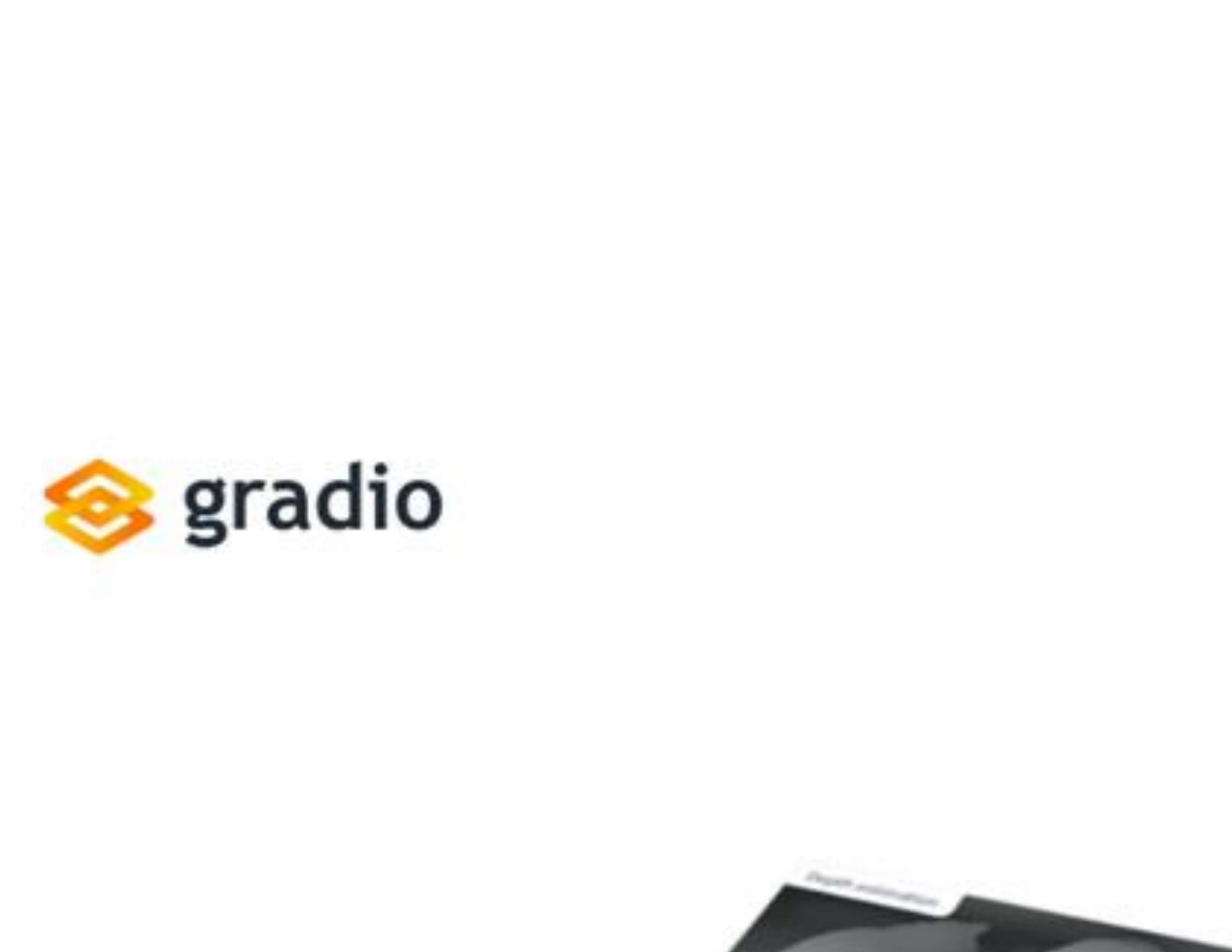
S

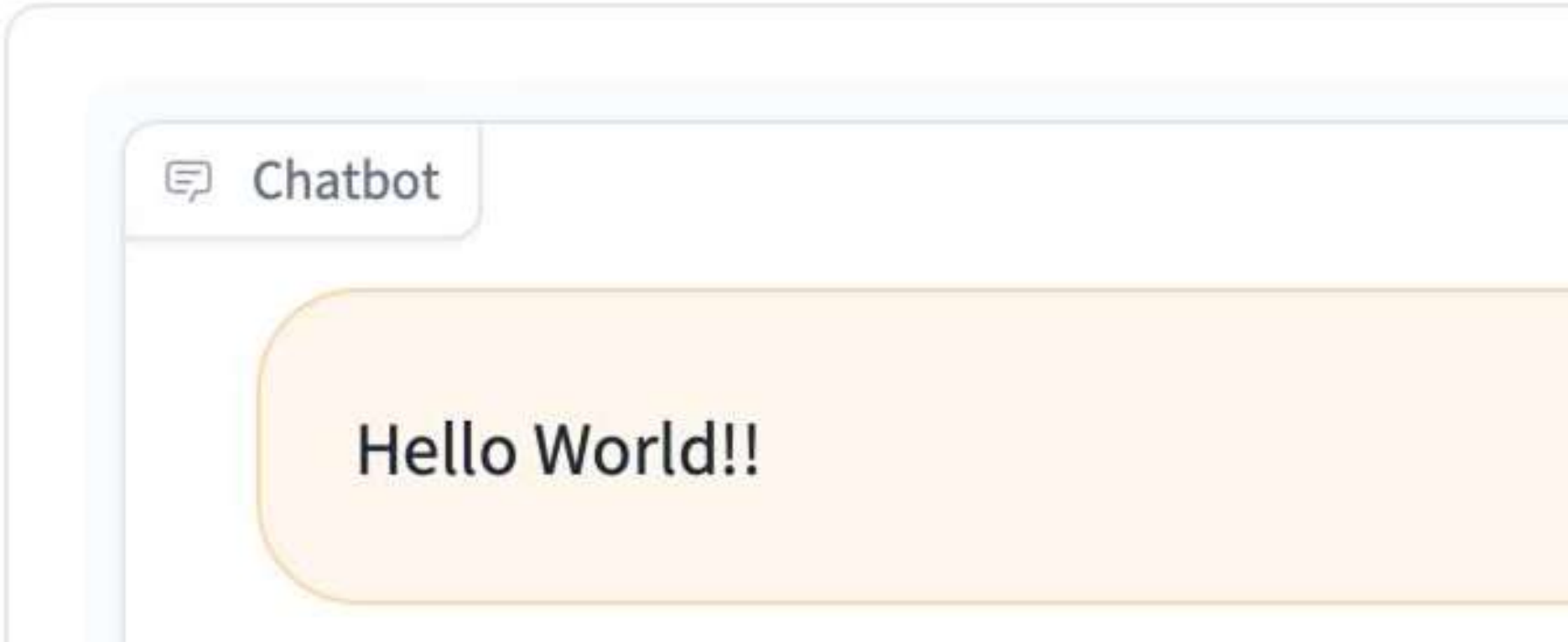
Chatbot

Hello World!!

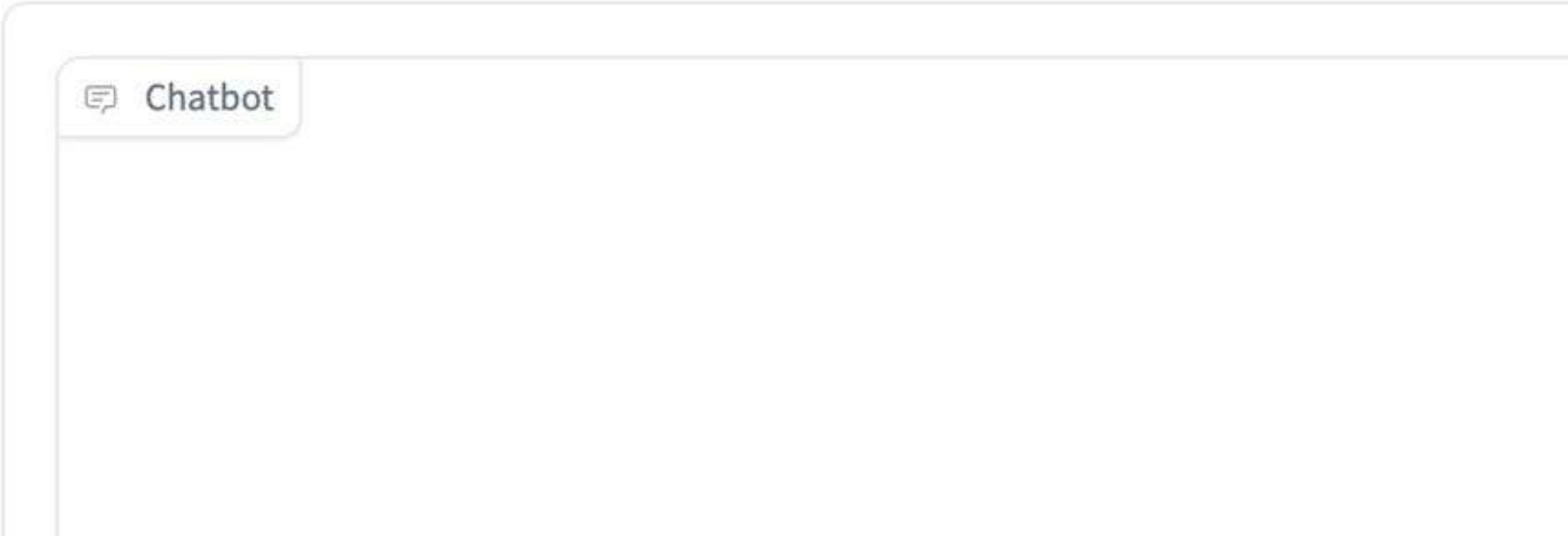
You typed: Hello World!!

gradio




Chatbot
Hello World!!

Hello World Airbnb Map Chatbot Streaming Diffu


Chatbot

Chatbot

Hello World!!

You typed: Hello World!!

Building RAG A

Chatbot

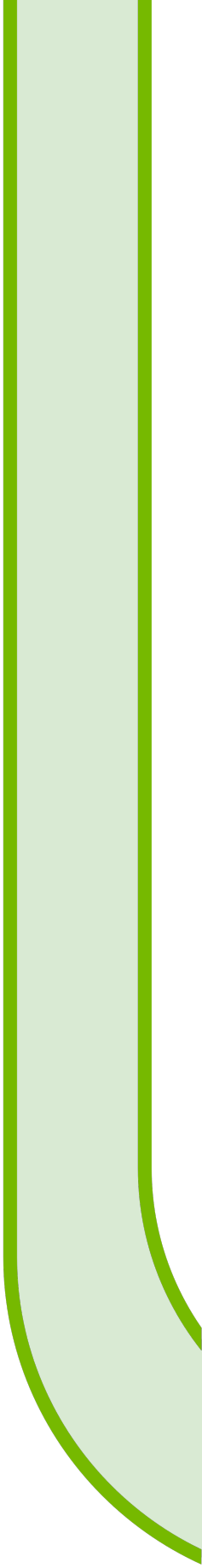
Hello World!!

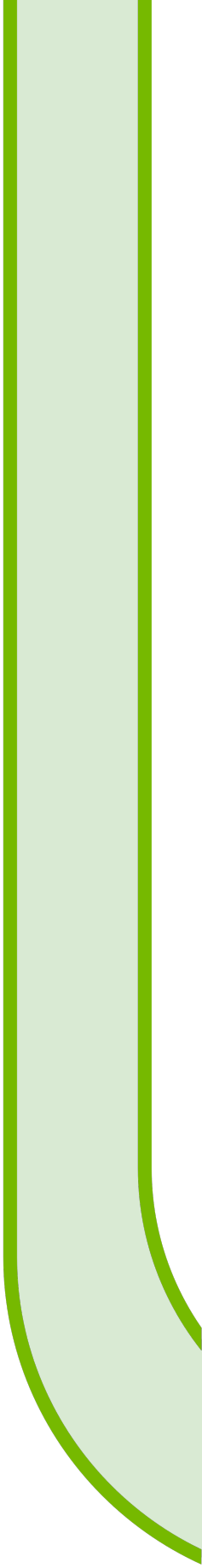
You typed: Hello World!!

Jupyter Notebook

Jupyter Notebook

Jupyter Notebook





OpenAI G

```
{
"messages": [{
```

```
{
"messages": [{
```

Larg

Larg

Larg

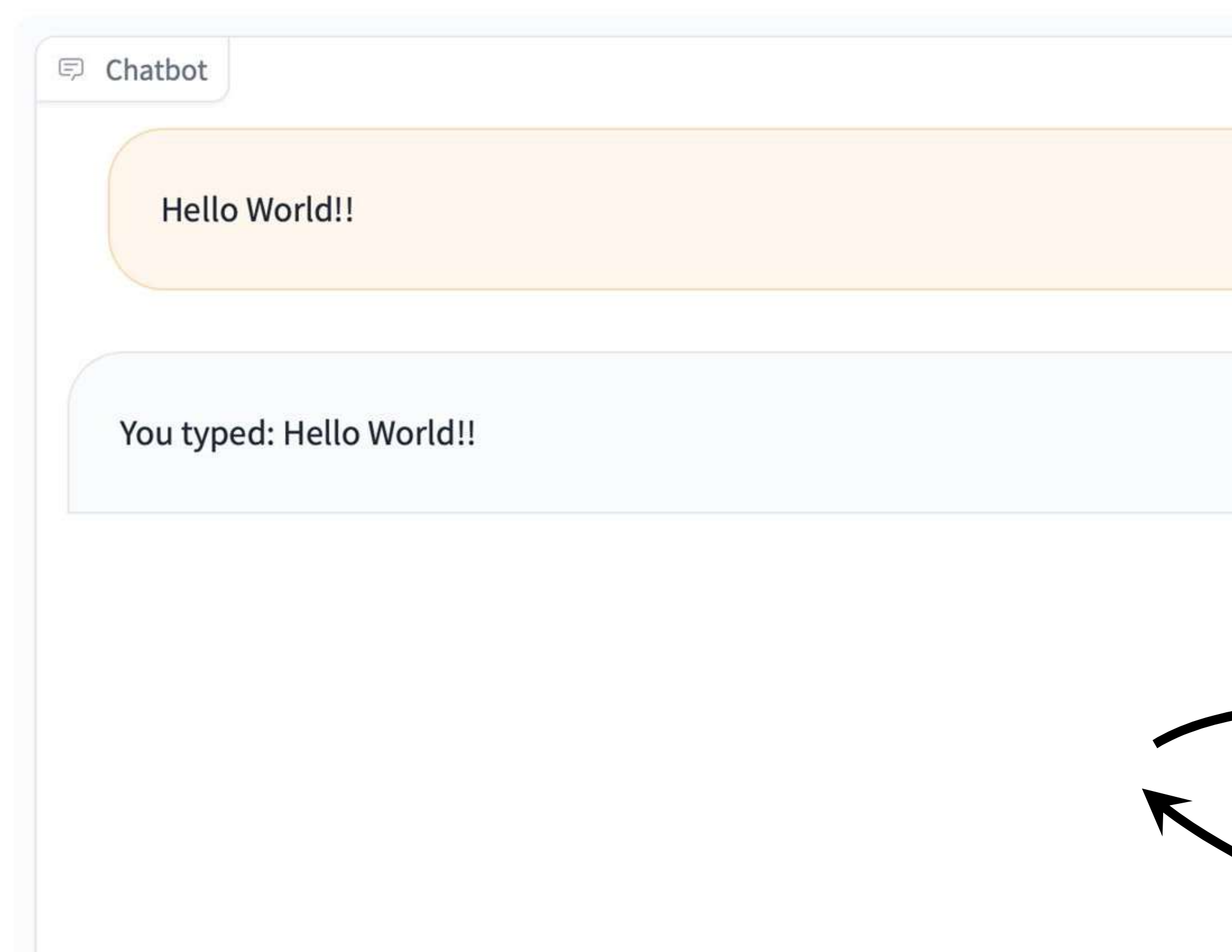
OpenAI G

OpenAI G

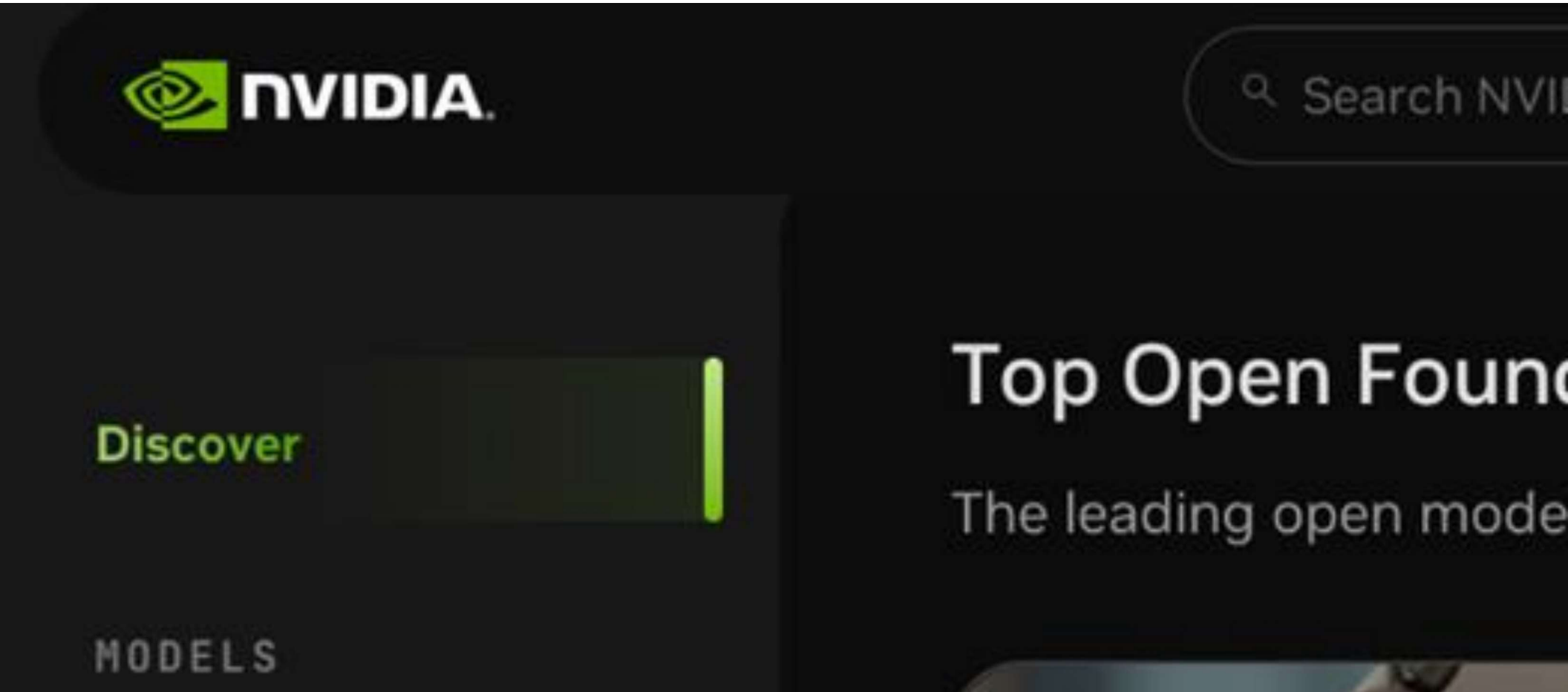
```
{
"messages": [{
```

Millions of Apps |

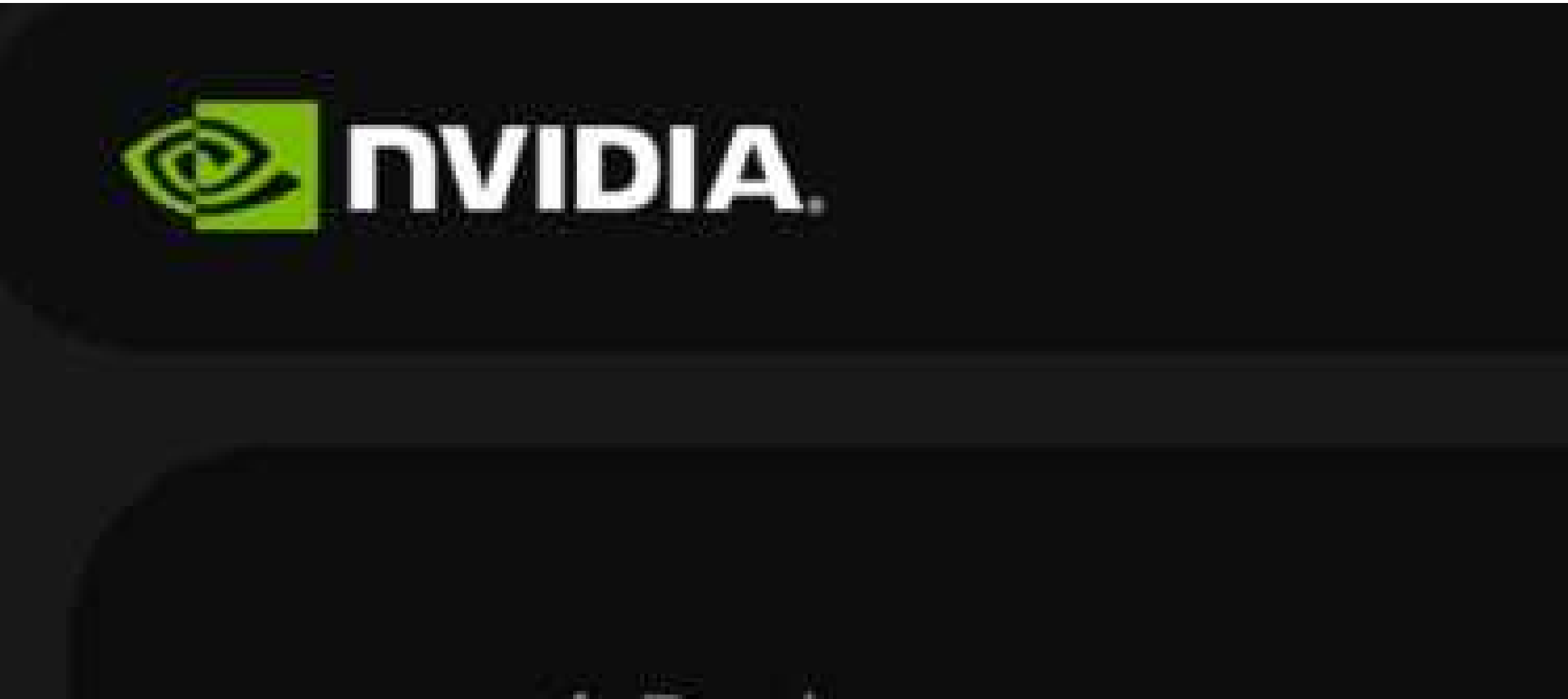
Application Layer


Chatbot
Hello World!!
You typed: Hello World!!

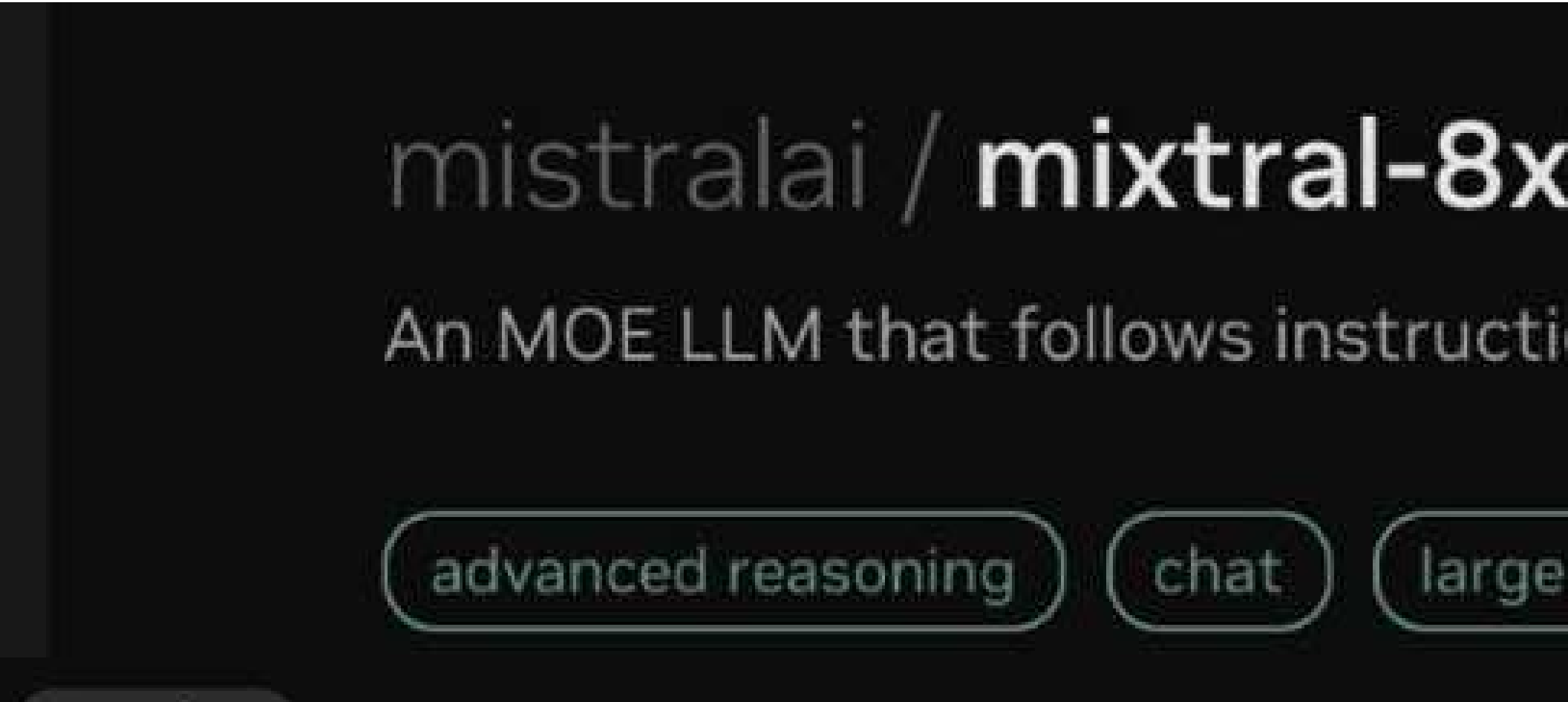
NV


NVIDIA.
Search NVI
Discover
MODELS
Top Open Found
The leading open mode

NV



NV


mistralai / mixtral-8x
An MOE LLM that follows instructi
advanced reasoning
chat
large

Fror

Fron

Output Parsers

Models

Prompts

Example Selectors

Output Parsers

Models

Prompts

Example Selectors

ChatNVIDIA

Building RAG A

Input

Prom

Input

Prom

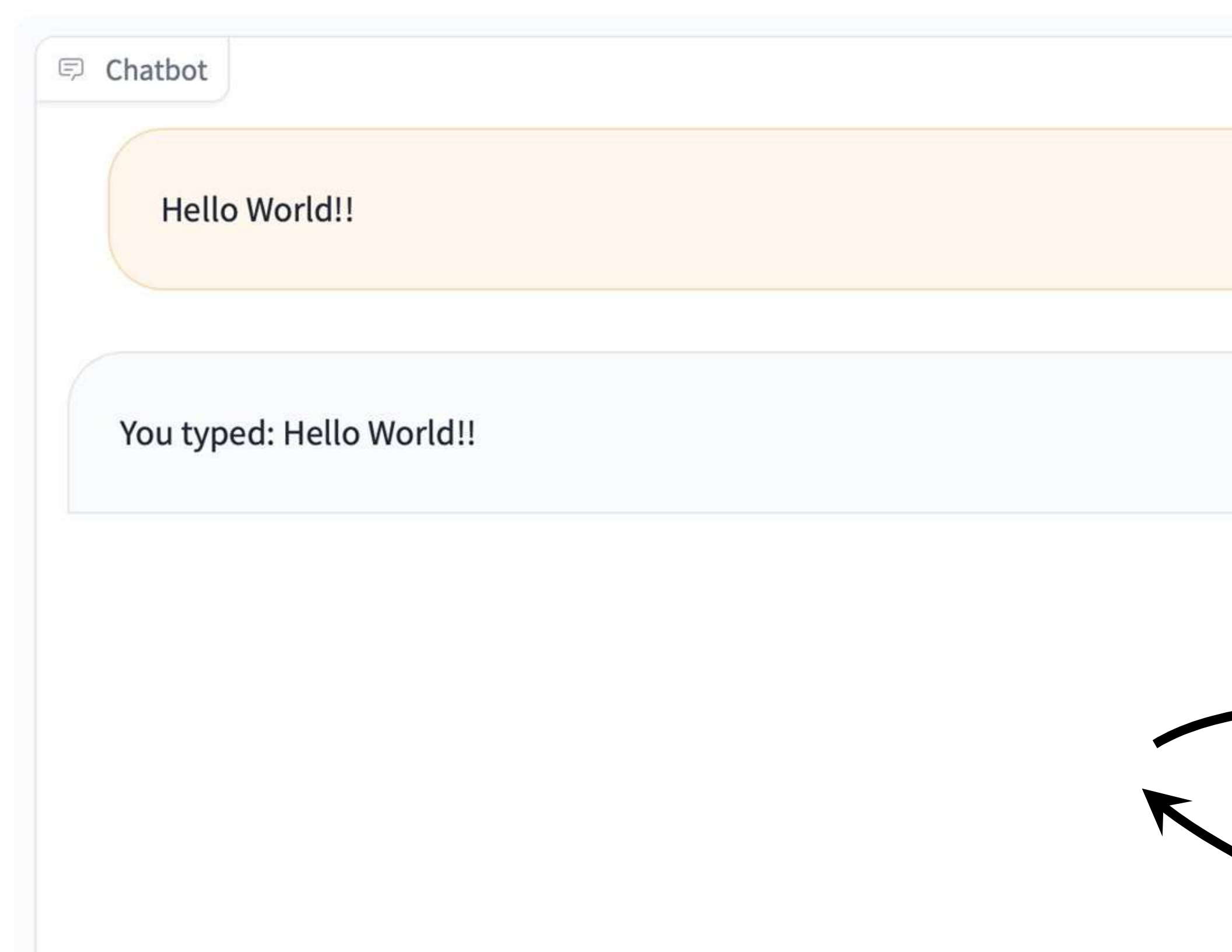
Input

Prom

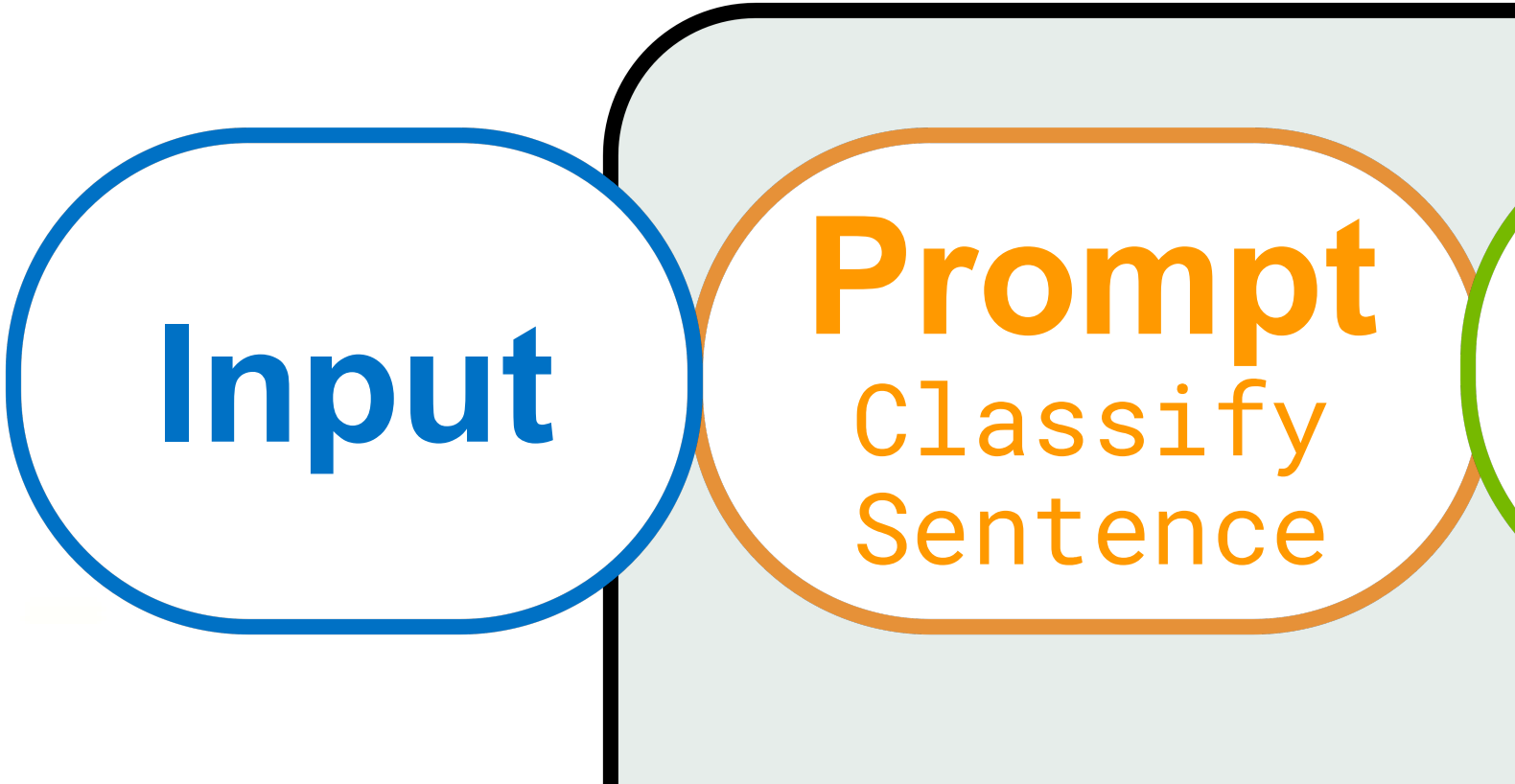
Input

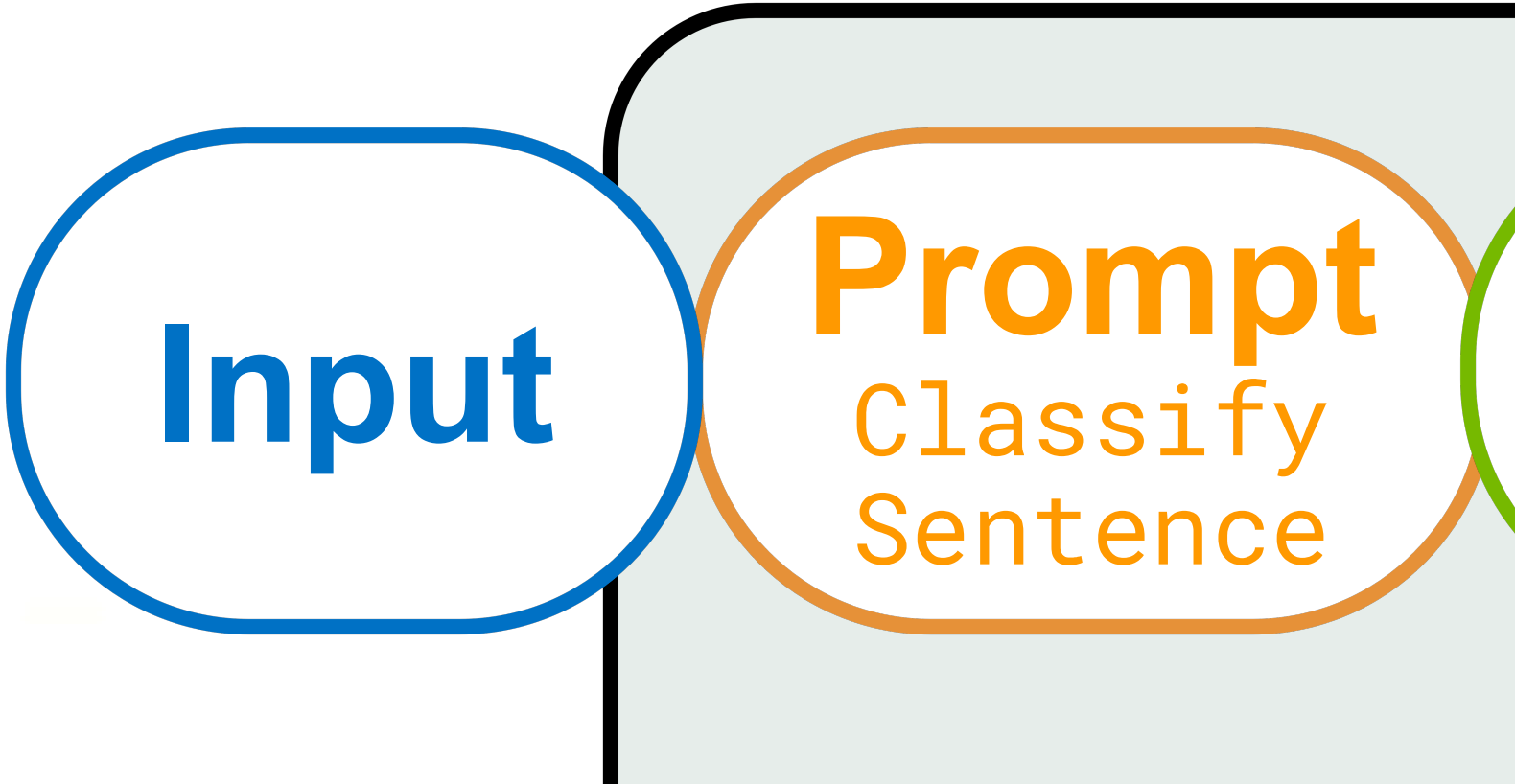
OBSE

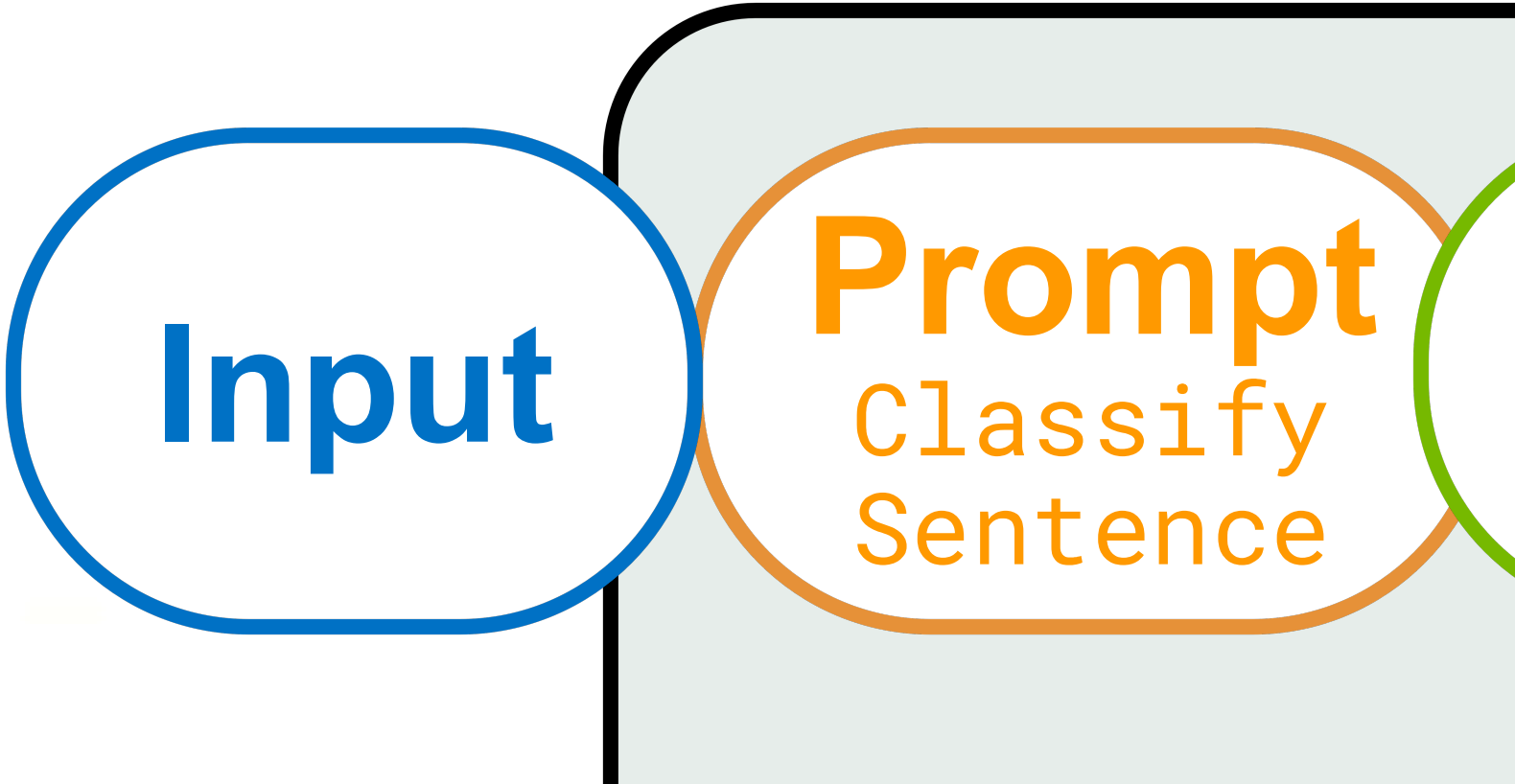
DEPL

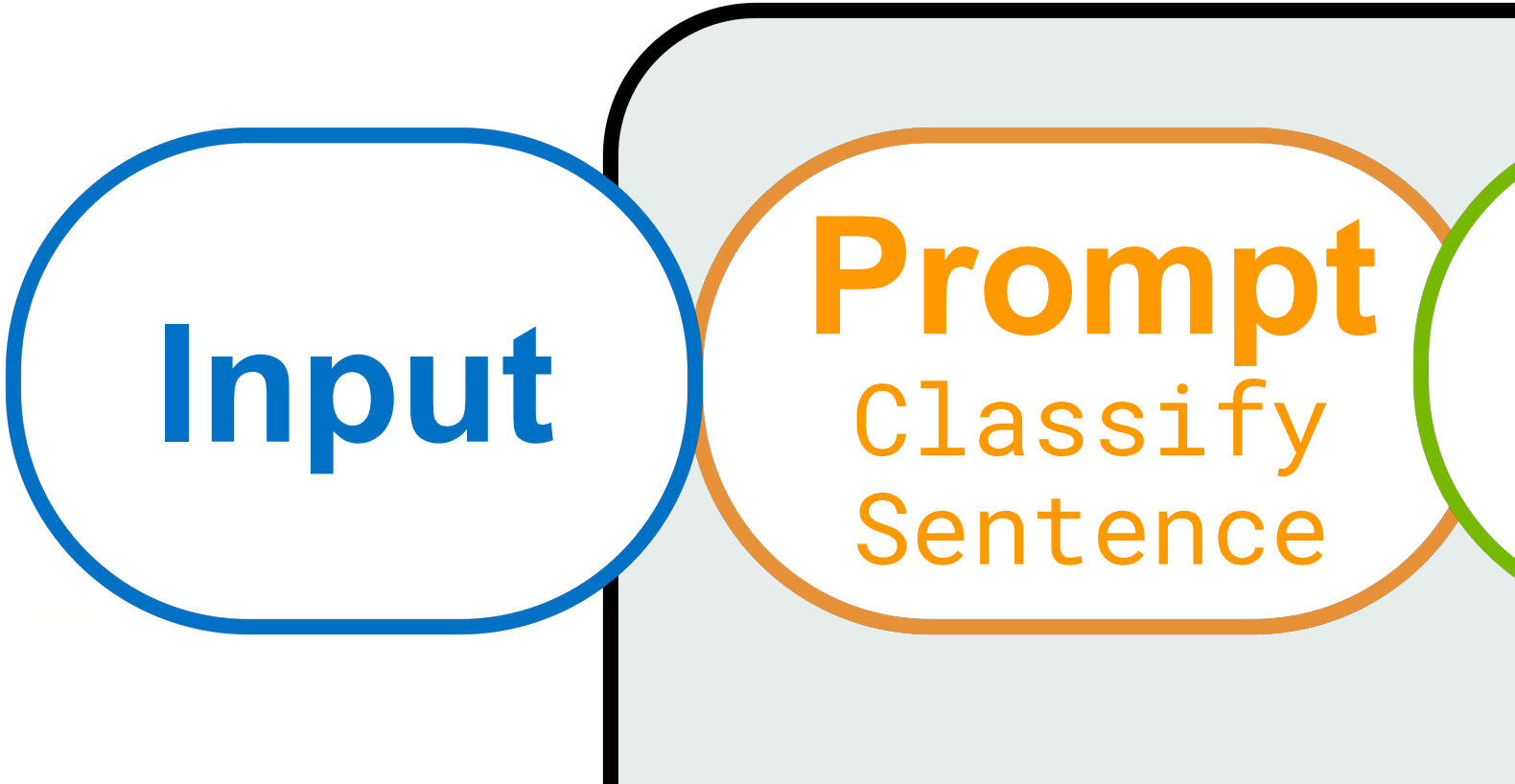

Chatbot
Hello World!!
You typed: Hello World!!

Building RAG A

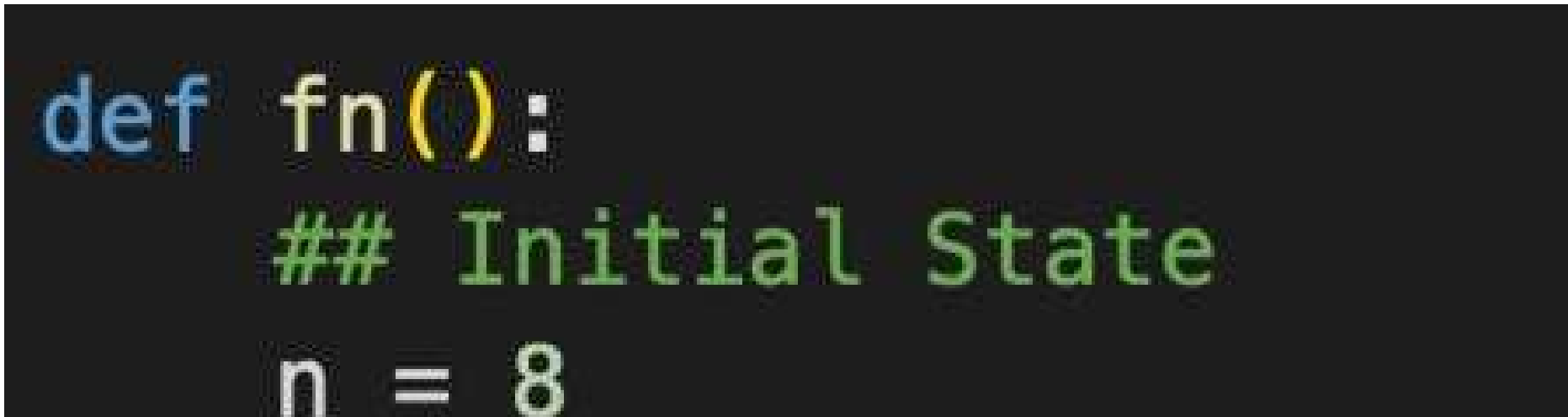

Input
Prompt
Classify
Sentence


Input
Prompt
Classify
Sentence


Input
Prompt
Classify
Sentence


Input
Prompt
Classify
Sentence


Input
Prompt
Classify
Sentence


def fn():
Initial State
n = 8

F

```
def get_initial_state():
    n = 8
    fib = [0,1]
    return locals()
```

E

```
running_state = {'n': 8,
running_state['n'] = 7
running_state['fib'] +=
running_
```

F

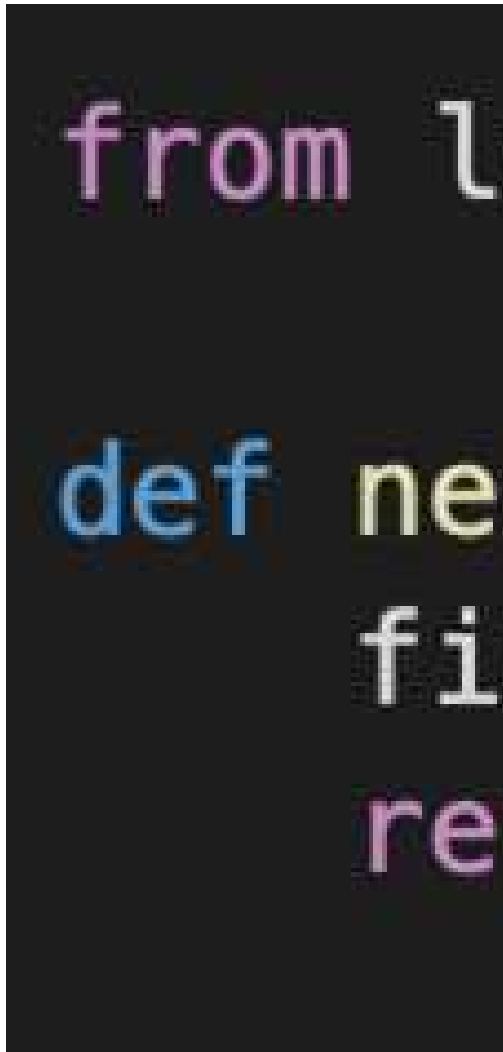
```
from langchain_core.runnables

def next_fib_fn(d):
    return d['fib'] + [d['fib
```

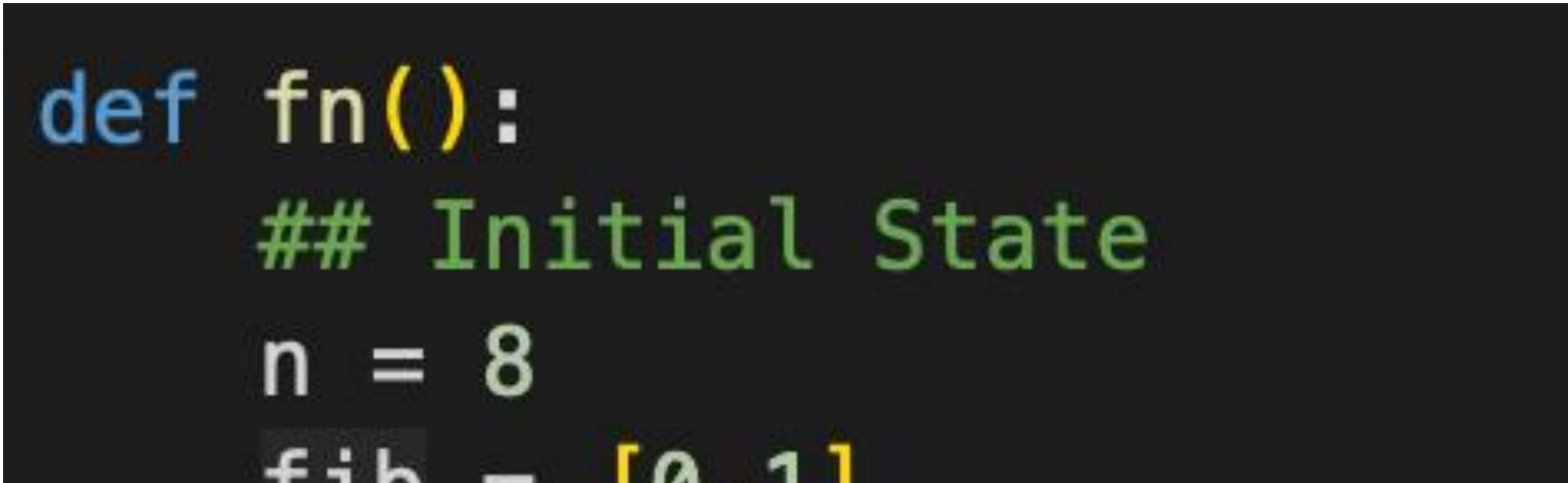
F

```
from langchain_core.runnables

def next_fib_fn(d):
    return d['fib'] + [d['fib
```

F

```
from langchain_core.runnables

def next_fib_fn(d):
    return d['fib'] + [d['fib
```


from l
def ne
fi
re


def fn():
Initial State
n = 8
fib = [0,1]

+Knowl

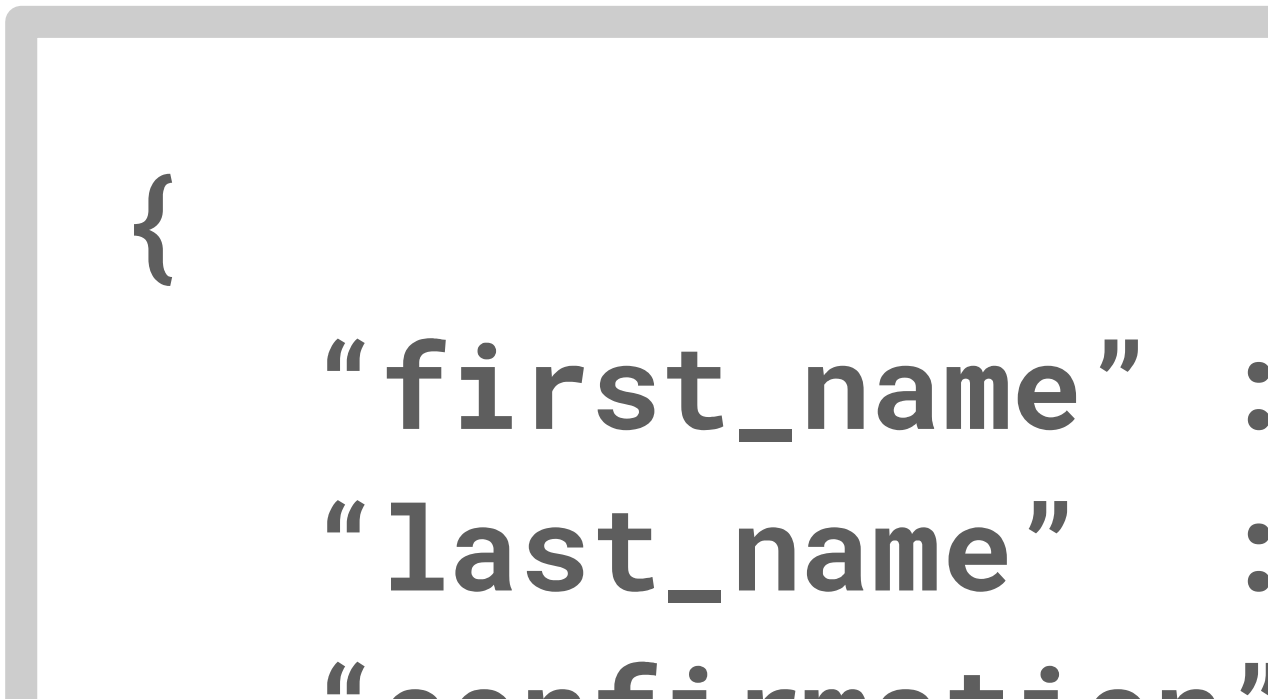
Pa

Pro

Pro

Pro

Pro


{
"first_name" :
"last_name" :
"confirmation"

+Knowl

Pa

+Knowl

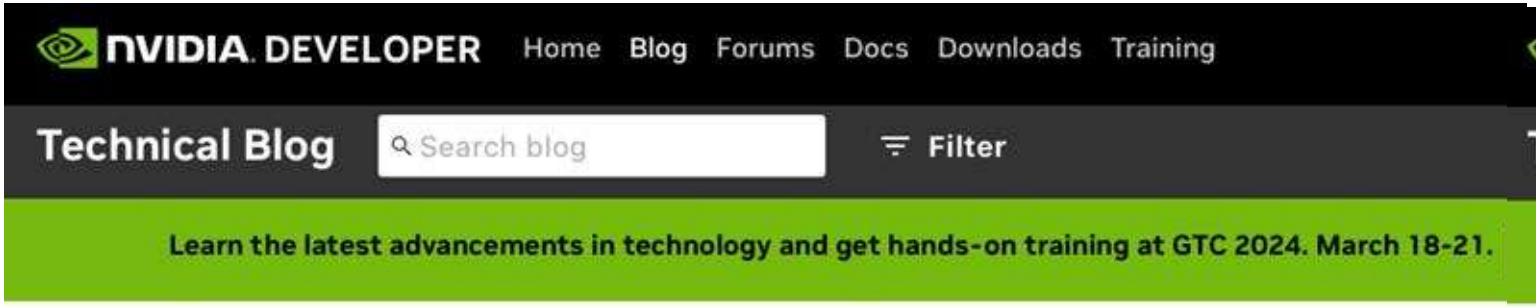
Pa

+Knowl

Pa

Building RAG A

Pro



Generative AI / LLMs

RAG 101: Retrieval-Augmented

Generative AI / LLMs

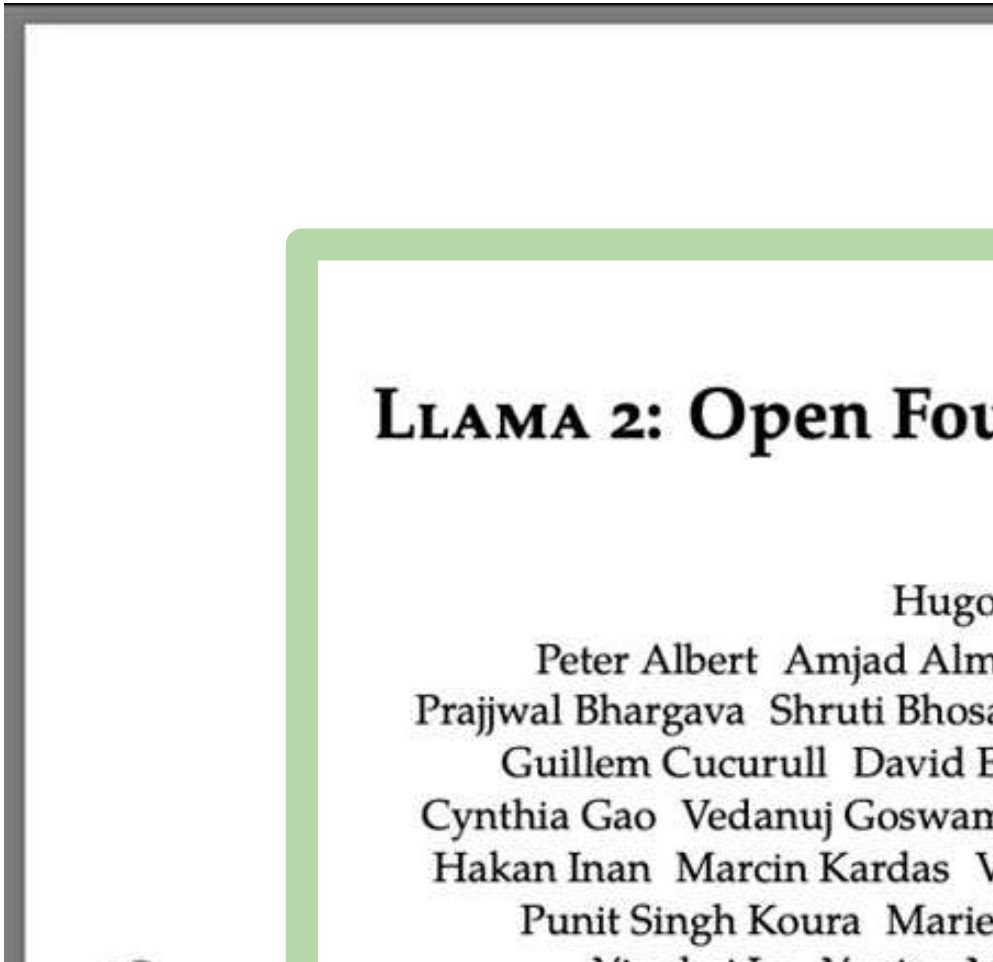
Deploying R


Company
Database


Company
Database


Company
Database


Company
Database


Llama 2: Open Fou
Hugo
Peter Albert Amjad Alm
Prajjwal Bhargava Shruti Bhosa
Guillem Cucurull David E
Cynthia Gao Vedanuj Goswam
Hakan Inan Marcin Kardas
Punit Singh Koura Marie

Llama 2: Open Foundation and Fine-Tune

Hugo Touvron* Louis Martin† Kevin Stone†
Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlyko
Prajjwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Ca
Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyi
Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Sa
Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Klou
Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee
Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pus

Llama 2: Open Foundation and Fine-Tuned

Hugo Touvron* Louis Martin† Kevin Stone†
Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov
Prajjwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton
Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin F
Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Sagha
Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Klouma
Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee D
Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushk
Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rung
Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing

How does Flying work according t

RunnableLamb

```
from langchain_openai import ChatOpe
from langchain_core.messages import H
from langgraph.graph import END, Mes
```

RunnableLamb

Building RAG A

How does Flying work according t

Pro

Pro

Dec

Autoregres

(Common

Dec

Autoregres

(Common

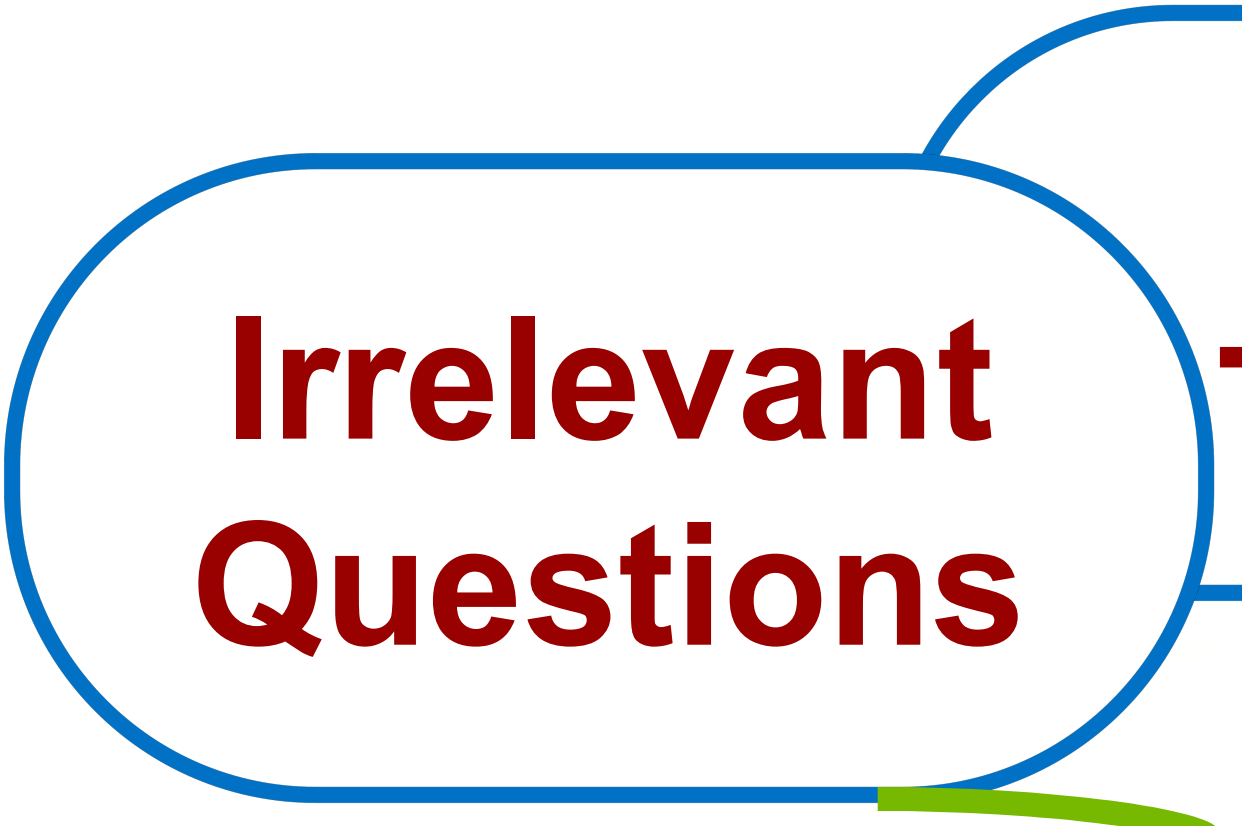

NVIDIA

Bi-En

Sym

f(query,

Building RAG A


Irrelevant
Questions

Illegal Topics

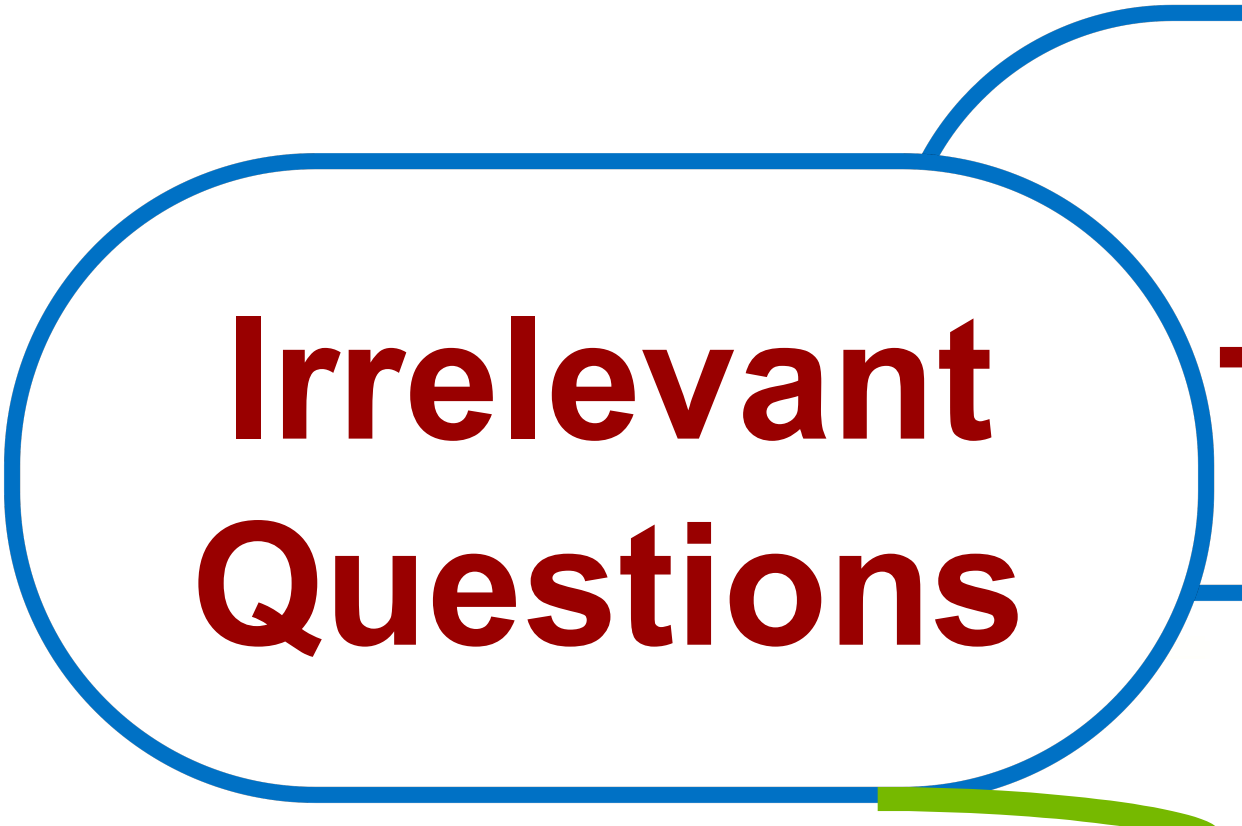
Irrelevant

Illegal Topics

Irrelevant

Illegal Topics

Irrelevant


Irrelevant
Questions

Building RAG A

Who is the


NVIDIA


Vector Stores

```
from langchain_nvidia_ai
from langchain.vectorsto


embedder = NVIDIAEmbeddi
```

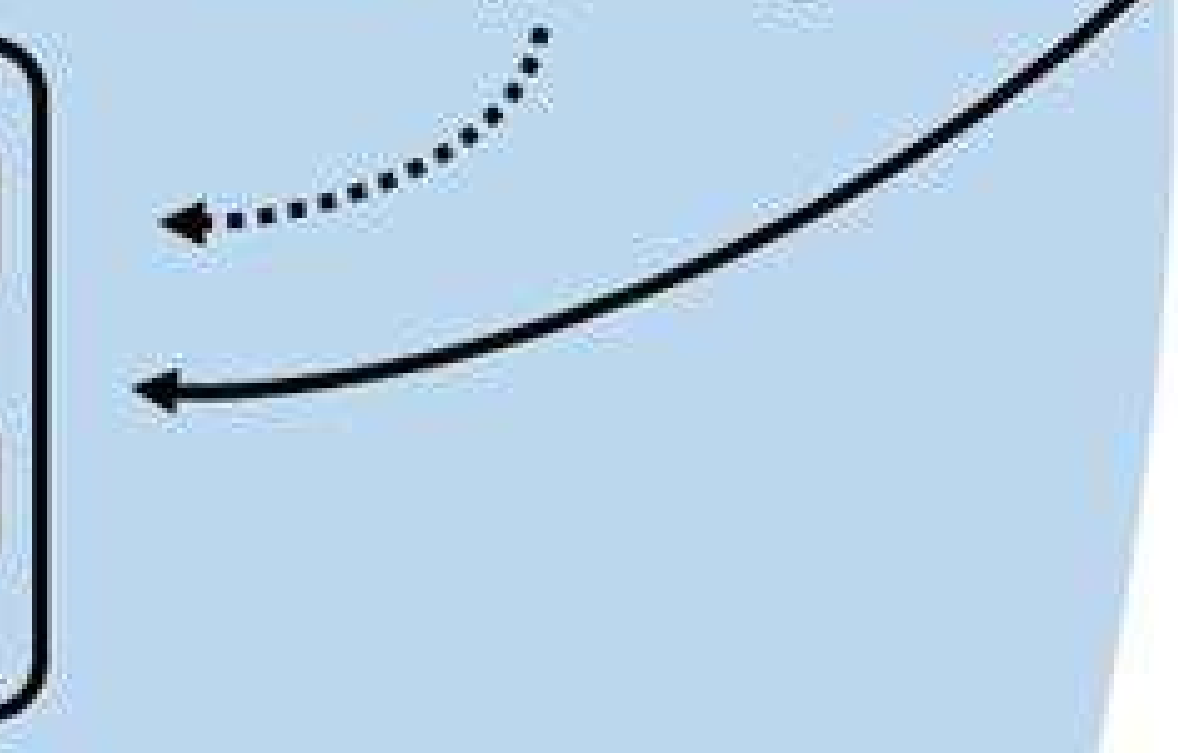
```
from langchain_nvidia_ai
from langchain.vectorsto


embedder = NVIDIAEmbeddi
```

```
from langchain_nvidia_ai
from langchain.vectorsto


embedder = NVIDIAEmbeddi
```

```
from langchain_nvidia_ai
from langchain.vectorsto


embedder = NVIDIAEmbeddi
```

Pron

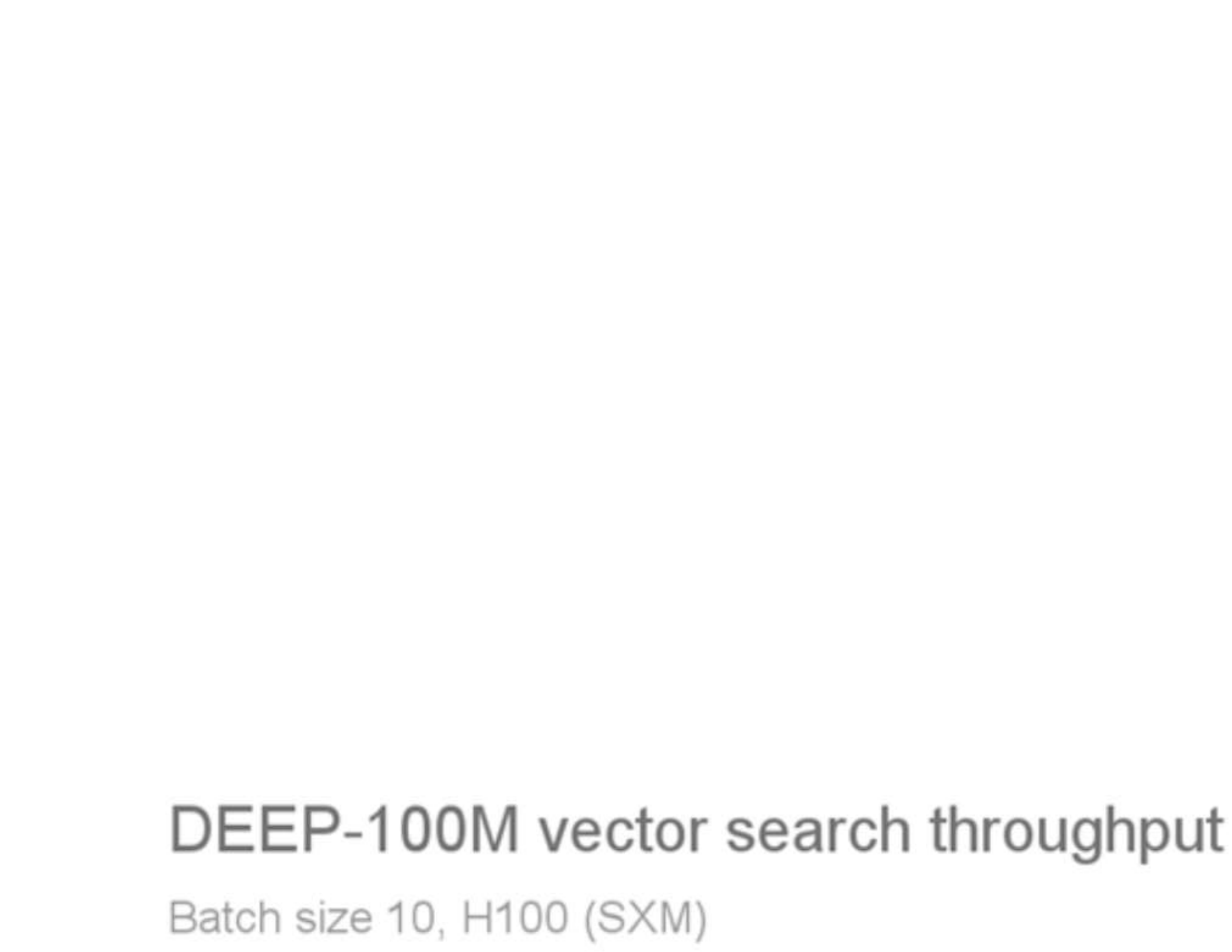
Pron



Loca

Local Host

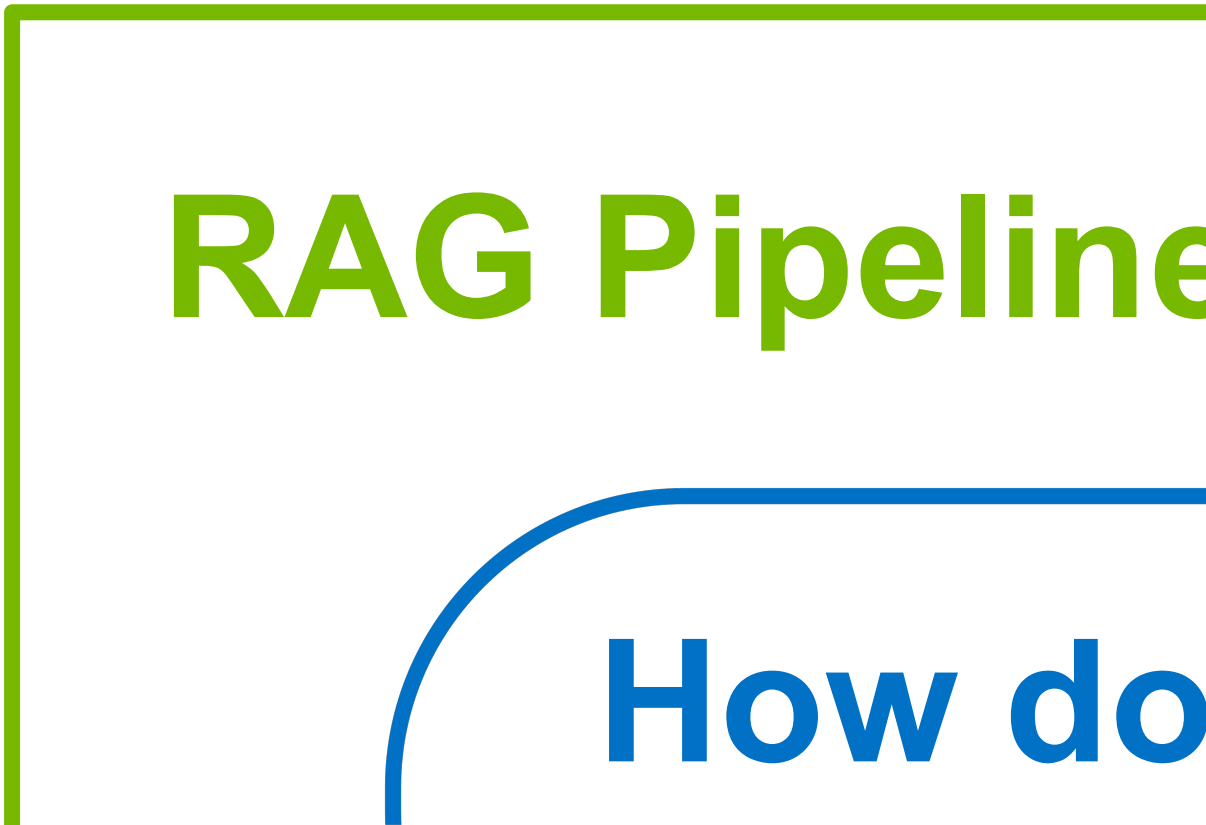
Milvus

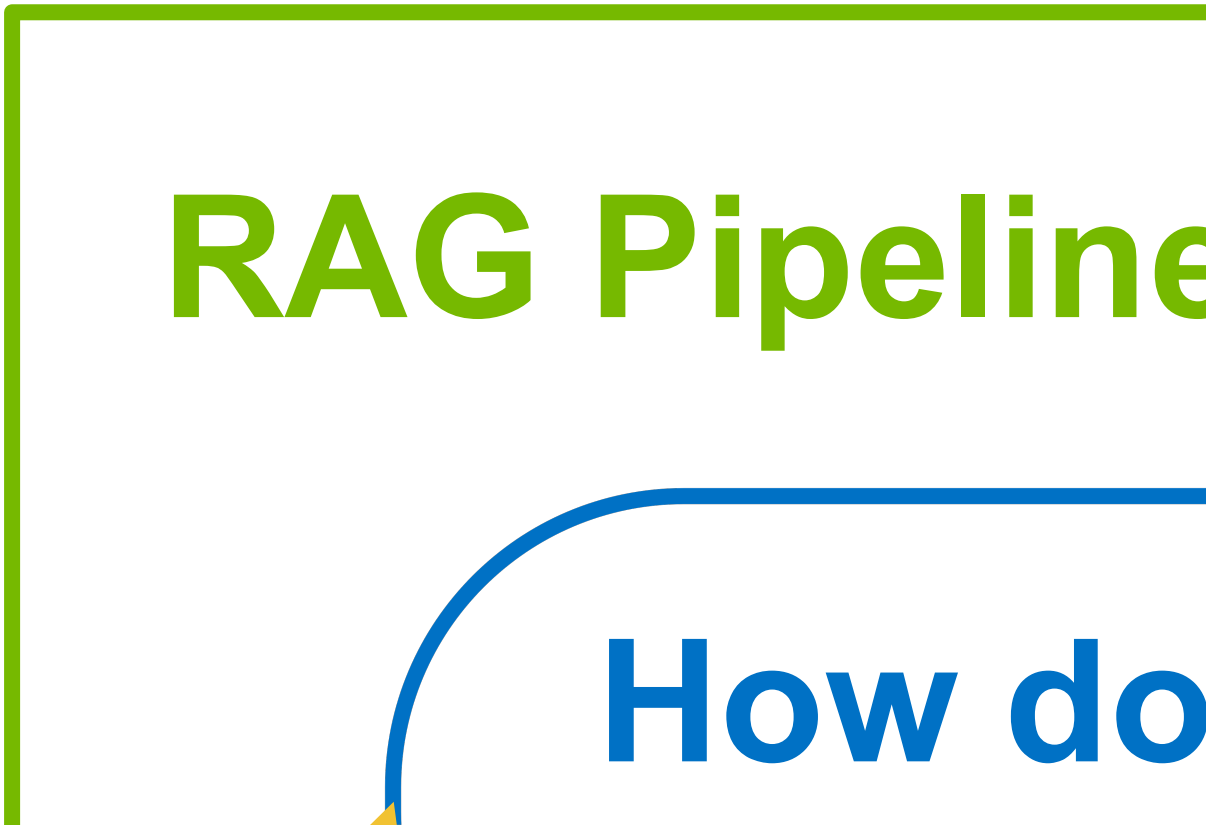

DEEP-100M vector search throughput
Batch size 10, H100 (SXM)


Hello! My name is Jane
Hello Jane! How are you?

How does RAG work?

Building RAG A


RAG Pipeline
How do


RAG Pipeline
How do

VDB

VDB

VDB

VDB

Ge

VDB

+ Embed

R

Q

```
54 ⌄  NLI
55
56      Na
57
58    Con
59    Joh
```

Prompt

Ask some questions

LLM

VDB

Remot

Congratulation