

Sustainability and ESG Data Analytics

NVIDIA Building RAG Agents with LLMs Part I: LLM Services and AI Foundation Models

1141ESGDA06

MBA, IM, NTPU (M5265) (Fall 2025)

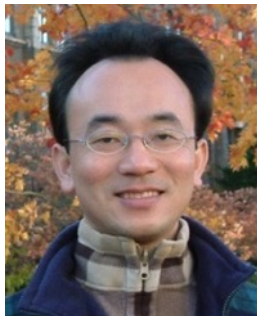
Wed 2, 3, 4 (9:10-12:00) (B3F17)



<https://meet.google.com/miy-fbif-max>

 **NVIDIA**
University Ambassador
Certified Instructor

 **aws** educate | Cloud
Ambassador
2020 Cohort



Min-Yuh Day, Ph.D,
Professor and Director

Institute of Information Management, National Taipei University

<https://web.ntpu.edu.tw/~myday>

2025-10-22



Syllabus

Week Date Subject/Topics

1 2025/09/10 Introduction Sustainability and ESG Data Analytics

**2 2025/09/17 Environmental, Social, and Governance (ESG) in
Net-Zero Digital Transformation**

3 2025/09/24 Data Science for Sustainability and ESG

4 2025/10/01 Case Study on Sustainability and ESG Data Analytics I

**5 2025/10/08 Web 3.0 and Big Data Analysis in Fintech, Green and
Sustainable Finance**

6 2025/10/15 ESG Data Gathering, Analysis, and Visualization

Syllabus

Week Date Subject/Topics

**7 2025/10/22 NVIDIA Building RAG Agents with LLMs Part I:
LLM Services and AI Foundation Models**

8 2025/10/29 Self-Learning

9 2025/11/05 Midterm Project Report

**10 2025/11/12 NVIDIA Building RAG Agents with LLMs Part II:
Document Loading, Chunking, and Embeddings**

**11 2025/11/19 NVIDIA Building RAG Agents with LLMs Part III:
Retrieval-Augmented Generation with
Vector Stores and RAG Evaluation**

Syllabus

Week Date Subject/Topics

12 2025/11/26 Case Study on Sustainability and ESG Data Analytics II

**13 2025/12/03 Artificial Intelligence of things (AIoT) in ESG and
Sustainability Applications**

14 2025/12/10 Generative AI for ESG Rating and Reporting Generation

15 2025/12/17 Final Project Report I

16 2025/12/24 Final Project Report II

Building RAG Agents with LLMs

Part 1: Your Course Environment

Part 2: LLM Services

Part 3: Intro to LangChain

Part 4: Running State Chain

Part 5: Working with Documents

Part 6: Embedding Model for Retrieval

Part 7: Document Retrieval with Vector Databases

Part 8: RAG Evaluation

NVIDIA

Building RAG Agents with LLMs

[0,1,2] [3, 4] [5, 6] [7,8]

NVIDIA Building RAG Agents with LLMs

[0, 1, 2]

Building RAG Agents with LLMs



DEEP
LEARNING
INSTITUTE

This Lab 0 : 05 : 59 / 2 : 00 : 00

Course 1 : 47 : 00 / 32 : 00 : 00



LAUNCH



STOP TASK

Course Slides

Course Slides

Environment and LLMs

Environment [0, 1, 2]

- chatbot
- composer
- docker_router
- frontend
- imgs
- llm_client
- slides

- 00_jupyterlab.ipynb
- 01_microservices.ipynb
- 02_llms.ipynb
- 99_table_of_contents.ipynb

Spring 2025

Generative AI Innovative Applications



University Ambassador



This certificate acknowledges that

Min-Yuh Day

has been certified to deliver NVIDIA instructor-led workshop for
academia

A handwritten signature in black ink, appearing to read "Greg Estes", written over a horizontal line.

Greg Estes

Vice President, NVIDIA

Issue Date: : March 7, 2025

Ambassador Certification ID: cCFh1ZWWTvqKTq7dcKkEWw

<https://learn.nvidia.com/certificates?id=cCFh1ZWWTvqKTq7dcKkEWw>



Certified Instructor



This certificate acknowledges that

Min-Yuh Day

has been certified to deliver the instructor-led workshop

Building RAG Agents with LLMs

A handwritten signature in black ink, appearing to read "Greg Estes", written over a horizontal line.

Greg Estes

Vice President, NVIDIA

Issue Date: : March 7, 2025

Certification ID: OVmqY4cSSya0BdMQBWHxzw

<https://learn.nvidia.com/certificates?id=OVmqY4cSSya0BdMQBWHxzw>



Certified Instructor



This certificate acknowledges that

Min-Yuh Day

has been certified to deliver the instructor-led workshop

Fundamentals of Deep Learning

A handwritten signature in black ink, appearing to read "Greg Estes", written over a horizontal line.

Greg Estes

Vice President, NVIDIA

Issue Date: : August 1, 2025

Certification ID: XhLz9YFzRPyew64462TNjQ



Certificate of Completion

This certificate is awarded to

Min-Yuh Day

for successfully completing

Building RAG Agents with LLMs

A handwritten signature in black ink, appearing to read "Greg A. Estes", written over a horizontal line.

Greg Estes

Vice President, NVIDIA

Issue Date: : December 8, 2024

Certification ID: ed-qOCIMQaatzU8SNUNxgw |

https://learn.nvidia.com/certificates?id=ed-qOCIMQaatzU8SNUNxgw/courses/course?course_id=course-v1:DLI+S-FX-15+V1

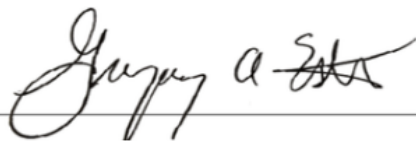
<https://learn.nvidia.com/certificates?id=ed-qOCIMQaatzU8SNUNxgw>

Certificate of Competency

This certificate is awarded to

Min-Yuh Day

for demonstrating competence in the completion of
Generative AI with Diffusion Models



Greg Estes

Vice President, NVIDIA

Issue Date: : February 28, 2025

Certification ID: q3oo-oBhTQKtyCCote2E-Q



Certificate of Competency

This certificate is awarded to

Min-Yuh Day

for demonstrating competence in the completion of

Getting Started with Deep Learning

A handwritten signature in black ink, appearing to read "Greg A. Estes", written over a horizontal line.

Greg Estes

Vice President, NVIDIA

Issue Date: : April 13, 2025

Certification ID: ny50I3TXT5W8Q_6kENklaQ

NVIDIA Certified Instructors (12 from Taiwan)

Shop Drivers Support 

Don't miss NVIDIA CEO Jensen Huang's GTC keynote March 18. Join early for the pregame event. 01d:04h:15m:12s [Add to Calendar](#)

Deep Learning Institute Find Training Self Paced Courses Instructor-Led Workshops Educator Programs ...

Filters

Workshop Certification Location [1] Organization Specialization Reset filters Search for instructors by name. Minimum

Sort by: Name A-Z

Name	Organization	Featured Workshop
Chi-Hung Chuang	Chung Yuan Christian University	Fundamentals of Deep Learning, Taiwanese
Chia Yu Hsu	National Taiwan University of Science and Technology (NTUST)	Applications of AI for Predictive Maintenance, Taiwanese
Chien-Yu Chen	National Taiwan University (NTU)	Fundamentals of Deep Learning, Taiwanese
Chun-Yi Lee	National Taiwan University (NTU)	Fundamentals of Deep Learning, Taiwanese
David Tseng	Cavedu	Getting Started with AI on Jetson Nano, Taiwanese
Hsinmin Lu	National Taiwan University (NTU)	Fundamentals of Deep Learning, English
Min-Yuh Day	National Taipei University (NTPU)	Building RAG Agents with LLMs, English
Ming-Che Chen	Southern Taiwan University of Science and Technology (STUST)	Getting Started with AI on Jetson Nano, Taiwanese
MingChe Hu	Chung Yuan Christian University	Fundamentals of Deep Learning, Taiwanese
Ping-Chun Hsieh	National Yang Ming Chiao Tung University (NYCU)	Fundamentals of Deep Learning, English
Po-Chih Kuo	National Tsing Hua University (NTHU)	Fundamentals of Deep Learning, English
Shu-Kai Hsieh	National Taiwan University (NTU)	Fundamentals of Deep Learning, Taiwanese

<https://www.nvidia.com/en-us/training/instructor-directory/search/>

NVIDIA Certified Instructors (14 from Taiwan) (2 from NTPU)(April 2025)

Filters

Workshop Certification Location [1] Organization Specialization Reset filters Search for instructors by name: Minimum

Sort by: Name A-Z

Name	Organization	Featured Workshop
Chi-Hung Chuang	Chung Yuan Christian University	Fundamentals of Deep Learning, Taiwanese
Chia Yu Hsu	National Taiwan University of Science and Technology (NTUST)	Applications of AI for Predictive Maintenance, Taiwanese
Chien-Yu Chen	National Taiwan University (NTU)	Fundamentals of Deep Learning, Taiwanese
Chun-Yi Lee	National Taiwan University (NTU)	Fundamentals of Deep Learning, Taiwanese
David Tseng	Cavedu	Building Transformer-Based Natural Language Processing Applications, Taiwanese
Hsinmin Lu	National Taiwan University (NTU)	Fundamentals of Deep Learning, English
Hung-Wen Chen	National Tsing Hua University (NTHU)	Fundamentals of Deep Learning, Taiwanese
Ko-Chia Yu	National Taipei University (NTPU)	Building RAG Agents with LLMs, English
Min-Yuh Day	National Taipei University (NTPU)	Building RAG Agents with LLMs, English
Ming-Che Chen	Southern Taiwan University of Science and Technology (STUST)	Getting Started with AI on Jetson Nano, Taiwanese
MingChe Hu	Chung Yuan Christian University	Fundamentals of Deep Learning, Taiwanese
Ping-Chun Hsieh	National Yang Ming Chiao Tung University (NYCU)	Fundamentals of Deep Learning, English
Po-Chih Kuo	National Tsing Hua University (NTHU)	Fundamentals of Deep Learning, English
Shu-Kai Hsieh	National Taiwan University (NTU)	Fundamentals of Deep Learning, Taiwanese

NVIDIA Developer Program

<https://developer.nvidia.com/join-nvidia-developer-program>

NVIDIA Deep Learning Institute (DLI)

<https://learn.nvidia.com/>

Get NVIDIA DLI Certificate Before the NVIDIA Workshop

- **Step 1. Join NVIDIA Developer Program (Free)**
<https://developer.nvidia.com/join-nvidia-developer-program>
- **Step 2. Visit NVIDIA Deep Learning Institute (DLI)**
<https://learn.nvidia.com/>
- **Step 3. Enroll "Building RAG Agents with LLMs" Self-Paced Course (Free)**
https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Get NVIDIA DLI Certificate

- Welcome to NVIDIA DLI "**Building RAG Agents with LLMs**" **workshop**.
- Step 1. Visit <https://learn.nvidia.com/dli-event>
- Step 2. Enter the event code: **NTPU_RAG_AMBASSADOR_OC25**
- Step 3. Complete the NVIDIA DLI course and review the course datasheet at <https://developer.nvidia.com/dli/getready>

Get NVIDIA DLI Certificate Workshop

DLI Event

Event Code

NTPU_RAG_AMBASSADOR_OC25

Enter your event code.

ENROLL

Welcome to NVIDIA DLI

" Building RAG Agents with LLMs "
workshop.

Step 1. Visit <https://learn.nvidia.com/dli-event>

Step 2. Enter the event code:

NTPU_RAG_AMBASSADOR_OC25

<https://learn.nvidia.com/dli-event>

Building RAG Agents with LLMs

Deep Learning Institute Find Training Self Paced Courses Instructor-Led Workshops Educator Programs Enterprise Solutions Certification Resources

Building RAG Agents with LLMs

Course Progress Bookmarks Updates

Building RAG Agents with LLMs Environment and LLMs Environment [0, 1, 2]

Building RAG Agents with LLMs

Course Slides

Course Slides

Environment and LLMs

Environment [0, 1, 2]

LangChain

Documents and Embeddings

Retrieval-Augmented
Generation

Next Steps

Previous

Next

Welcome to **Building RAG Agents with LLMs**. In this first section, we will get introduced to the overall course environment, LLM services, and recommended workflows!

This tab contains the course environment for this section!

Please click the "Start" button to start up your own private server for hands-on coding practice. It will take a few minutes to start up, so go ahead and click it now and then proceed to the next video! After a few minutes when the server has loaded, click "Launch" to access the code labs.



Workshop


START

Generative AI with Diffusion Models

Deep Learning Institute

Find Training

Self Paced Courses

Instructor-Led Workshops

Educator Programs

Enterprise Solutions

Certification

Generative AI with Diffusion Models

Course

Progress

Bookmarks

Updates

Generative AI with Diffusion Models

Start Here

Launch the Course

Generative AI with Diffusion Models

Start Here

Launch the Course

Next Steps

Feedback

 Bookmark this page

"Generative AI with Diffusion Models" workshop

Step 1. Visit <https://learn.nvidia.com/dli-event>

Step 2. Enter the event code:

NTPU_GENAI_AMBASSADOR_NO25



Please click the start button to the top-right. The course will take about 5 minutes to load. The course lecture slide decks will appear here when you launch the course. The numbers on the lectures correspond to the numbers of the interactive notebooks in the course.

https://learn.nvidia.com/courses/course?course_id=course-v1:DLI+C-FX-08+V1

NVIDIA Deep Learning Institute (DLI)

Building RAG Agents with LLMs Workshop

Self-Paced Course

Building RAG Agents With LLMs

Certificate available
Free
8 hours

Instructor-Led Workshop

Building RAG Agents With LLMs

Certificate available
\$500
8 hours

Step 1.

Visit

<https://learn.nvidia.com/dli-event>

Step 2.

Enter the event code:

NTPU_RAG_AMBASSADOR_OC25



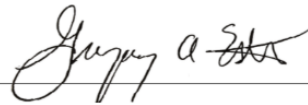
Certificate of Competency

This certificate is awarded to

Min-Yuh Day

for demonstrating competence in the completion of

Building RAG Agents with LLMs



Greg Estes

Vice President, NVIDIA

Issue Date: : December 8, 2024

Certification ID: ed-qOCIMQaatzU8SNUNxgw

[Download Certificate](#)[Add to Profile](#)

Join the NVIDIA Developer Program

take one of the
complimentary
technical self-
paced courses
(worth up to \$90)

Generative AI and LLMs Graphics and Simulation Accelerated Computing Data Science Deep Learning

8 hours
Getting Started With Deep Learning
Explore the fundamentals of deep learning by training neural networks and using results to improve performance and capabilities.

2 hours
Modeling Time-Series Data With Recurrent Neural Networks in Keras
Explore how to classify and forecast time-series data using recurrent neural networks (RNNs), such as modeling a patient's health over time.

4 hours
Deploying a Model for Inference at Production Scale
Learn how to deploy your own machine learning models on a GPU server.

8 hours
Building Real-Time Video AI Applications
Gain the knowledge and skills needed to enable the real-time transformation of raw video data from widely deployed camera sensors into deep learning-based insights.

2 hours
Introduction to Graph Neural Networks
Learn the basic concepts, models, and applications of graph neural networks.

4 hours
Introduction to Physics-Informed Machine Learning With Modulus
Learn the various building blocks of NVIDIA Modulus, which turbocharges use cases by building physics-based deep learning models that are 100,000X faster than traditional methods and offers high-fidelity simulation results.

2 hours
Get Started With Highly Accurate Custom ASR for Speech AI
Learn to build, train, fine-tune, and deploy a GPU-accelerated automatic speech recognition (ASR) service with NVIDIA® Riva that includes customized features.

2 hours
Integrating Sensors With NVIDIA DRIVE
Find out how to integrate automotive sensors into your applications using NVIDIA DRIVE®.

<https://developer.nvidia.com/join-nvidia-developer-program>

NVIDIA Deep Learning Institute (DLI)

Self-Paced Course

Generative AI Explained

Free
2 hours

Self-Paced Course

Getting Started With Deep Learning

Certificate available
\$90
8 hours

Instructor-Led Workshop

Fundamentals of Deep Learning

Certificate available
\$500
8 hours

Self-Paced Course

Introduction to Transformer-Based Natural Language Processing

Certificate available
\$30
6 hours

Self-Paced Course

Building RAG Agents With LLMs

Certificate available
Free
8 hours

Instructor-Led Workshop

Building RAG Agents With LLMs

Certificate available
\$500
8 hours

Self-Paced Course

Generative AI with Diffusion Models

Certificate available
\$90
8 hours

Instructor-Led Workshop

Generative AI with Diffusion Models

Certificate available
\$500
8 hours

What do you want to learn today?

Filters

Level +

Format +

Topics -

- ☐ Deep Learning
- ☐ Accelerated Computing
- ☐ Generative AI/LLM
- ☐ Graphics and Simulation
- ☐ OpenUSD
- ☐ Data Science
- ☐ NIMS
- ☐ NIM
- ☐ RAPIDS

Free / Paid +

Language +

Generative AI 

Sort by: -- v

Showing 19 results

Generative AI x

Generative AI

All Courses

Self-paced

Generative AI Explained

Free
02:00

Self-paced

Generative AI with Diffusion
Models\$90
08:00

Instructor-Led

Generative AI with Diffusion
Models

08:00

Self-paced

Augment your LLM Using

Self-paced

Introduction to Transformer-

Instructor-Led

Rapid Application

<https://learn.nvidia.com/en-us/training/find-training?q=Generative+AI>

Building RAG Agents with LLMs

Deep Learning Institute Find Training Self Paced Courses Instructor-Led Workshops Educator Programs Enterprise Solutions Certification Resources

Self-paced Course

Building RAG Agents with LLMs

Agents powered by large language models (LLMs) have shown great retrieval capability for using tools, looking at documents, and plan their approaches. This course will show you how to deploy an agent system in practice with the flexibility to scale up your system to meet the demands of users and customers.

[About Course](#)[Objectives](#)[Topics Covered](#)[Course Outline](#)[Stay Informed](#)[Contact Us](#)[Continue Learning](#)

About this Course

This course is free for a limited time.

The evolution and adoption of large language models (LLMs) have been nothing short of revolutionary, with retrieval-based systems at the forefront of this technological leap. These models are not just tools for automation; they are partners in enhancing productivity, capable of holding informed conversations by interacting with a vast array of tools and documents. This course is designed for those eager to explore the potential of these systems, focusing on practical deployment and the efficient implementation required to manage the considerable demands of both users and deep learning models. As we delve into the intricacies of LLMs, participants will gain insights into advanced orchestration techniques that include internal reasoning, dialog management, and effective tooling strategies.

Course Details

Duration: 08:00

Price: Free

Level: Technical - Intermediate

Subject: Generative AI/LLM

Language: English

Course Prerequisites:

Introductory deep learning knowledge, with comfort

https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Generative AI Explained

Self-paced Course

Generative AI Explained

In this no-coding course, learn Generative AI concepts and applications, as well as the challenges and opportunities in this exciting field.

[About Course](#)[Objectives](#)[Topics Covered](#)[Course Outline](#)[Stay Informed](#)[Contact Us](#)[Continue Learning](#)

About this Course

Generative AI describes technologies that are used to generate new content based on a variety of inputs. In recent time, Generative AI involves the use of neural networks to identify patterns and structures within existing data to generate new content. In this course, you will learn Generative AI concepts, applications, as well as the challenges and opportunities in this exciting field.

Learning Objectives

Upon completion, you will have a basic understanding of Generative AI and be able to more effectively use the various tools built on this

Course Details

Duration: 02:00

Price: Free

Level: Technical - Beginner

Subject: Generative AI/LLM

Language: English

https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Introduction to Transformer-Based Natural Language Processing

Self-paced Course

Introduction to Transformer-Based Natural Language Processing

Learn how Transformers are used as the building blocks of modern large language models (LLMs). You'll then use these models for various NLP tasks, including text classification, named-entity recognition (NER), author attribution, and question answering.

[About Course](#) [Objectives](#) [Topics Covered](#) [Course Outline](#) [Stay Informed](#) [Contact Us](#)

Continue Learning

About this Course

Large Language Models (LLMs), or Transformers, have revolutionized the field of natural language processing (NLP). Driven by recent advancements, applications of NLP and generative AI have exploded in the past decade. With the proliferation of applications like chatbots and intelligent virtual assistants, organizations are infusing their businesses with more interactive human-machine experiences. Understanding how Transformer-based large language models (LLMs) can be used to manipulate, analyze, and generate text-based data is essential. Modern pre-trained LLMs can encapsulate the nuance, context, and sophistication of language, just as humans do. When fine-tuned and deployed correctly, developers can use these LLMs to build powerful NLP applications that provide natural and seamless human-computer interactions within chatbots, AI voice agents, and more. In this course, you'll learn how Transformers are used as the building blocks of modern large language models (LLMs). You'll then use these models for various NLP

https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-08+V1

Course Details

Duration: 06:00

Price: \$30

Level: Technical - Beginner

Subject: Generative AI/LLM

Language: English

Generative AI with Diffusion Models

Self-paced Course

Generative AI with Diffusion Models

Take a deeper dive into denoising diffusion models, which are a popular choice for text-to-image pipelines, with applications in creative content generation, data augmentation, simulation and planning, anomaly detection, drug discovery, personalized recommendations, and more.



About this Course

Thanks to improvements in computing power and scientific theory, generative AI is more accessible than ever before. Generative AI plays a significant role across industries due to its numerous applications, such as creative content generation, data augmentation, simulation and planning, anomaly detection, drug discovery, personalized recommendations, and more. In this course, learners will take a deeper dive into denoising diffusion models, which are a popular choice for text-to-image pipelines.

Learning Objectives

Course Details

Duration: 08:00

Price: \$90

Subject: Generative AI/LLM

Language: English

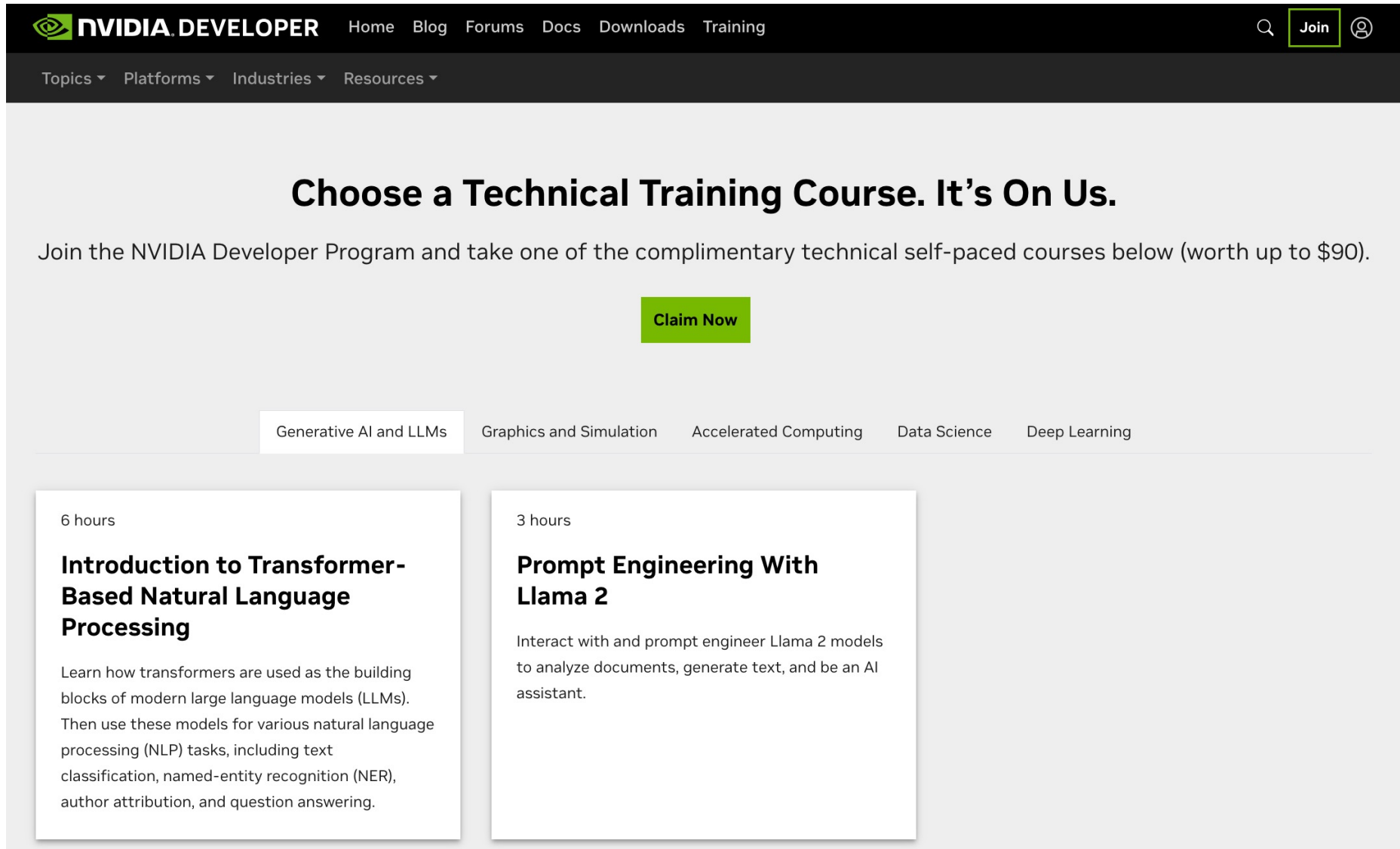
Course Prerequisites:

A basic understanding of [Deep Learning Concepts](#).

https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-14+V1

Join the NVIDIA Developer Program

take one of the complimentary technical self-paced courses (worth up to \$90)



The screenshot shows the NVIDIA Developer Program website. At the top is a dark navigation bar with the NVIDIA Developer logo, links for Home, Blog, Forums, Docs, Downloads, and Training, a search icon, a 'Join' button, and a user profile icon. Below this is a secondary dark bar with dropdown menus for Topics, Platforms, Industries, and Resources. The main content area has a light gray background. It features a large heading 'Choose a Technical Training Course. It's On Us.' followed by a paragraph: 'Join the NVIDIA Developer Program and take one of the complimentary technical self-paced courses below (worth up to \$90).' A green 'Claim Now' button is centered below the paragraph. Underneath is a horizontal menu with categories: Generative AI and LLMs (highlighted), Graphics and Simulation, Accelerated Computing, Data Science, and Deep Learning. Two course cards are displayed. The first card, 'Introduction to Transformer-Based Natural Language Processing', is 6 hours long and describes learning about transformers and their applications in NLP tasks like text classification and question answering. The second card, 'Prompt Engineering With Llama 2', is 3 hours long and describes interacting with Llama 2 models to analyze documents and generate text.

Choose a Technical Training Course. It's On Us.

Join the NVIDIA Developer Program and take one of the complimentary technical self-paced courses below (worth up to \$90).

Claim Now

Generative AI and LLMs Graphics and Simulation Accelerated Computing Data Science Deep Learning

6 hours

Introduction to Transformer-Based Natural Language Processing

Learn how transformers are used as the building blocks of modern large language models (LLMs). Then use these models for various natural language processing (NLP) tasks, including text classification, named-entity recognition (NER), author attribution, and question answering.

3 hours

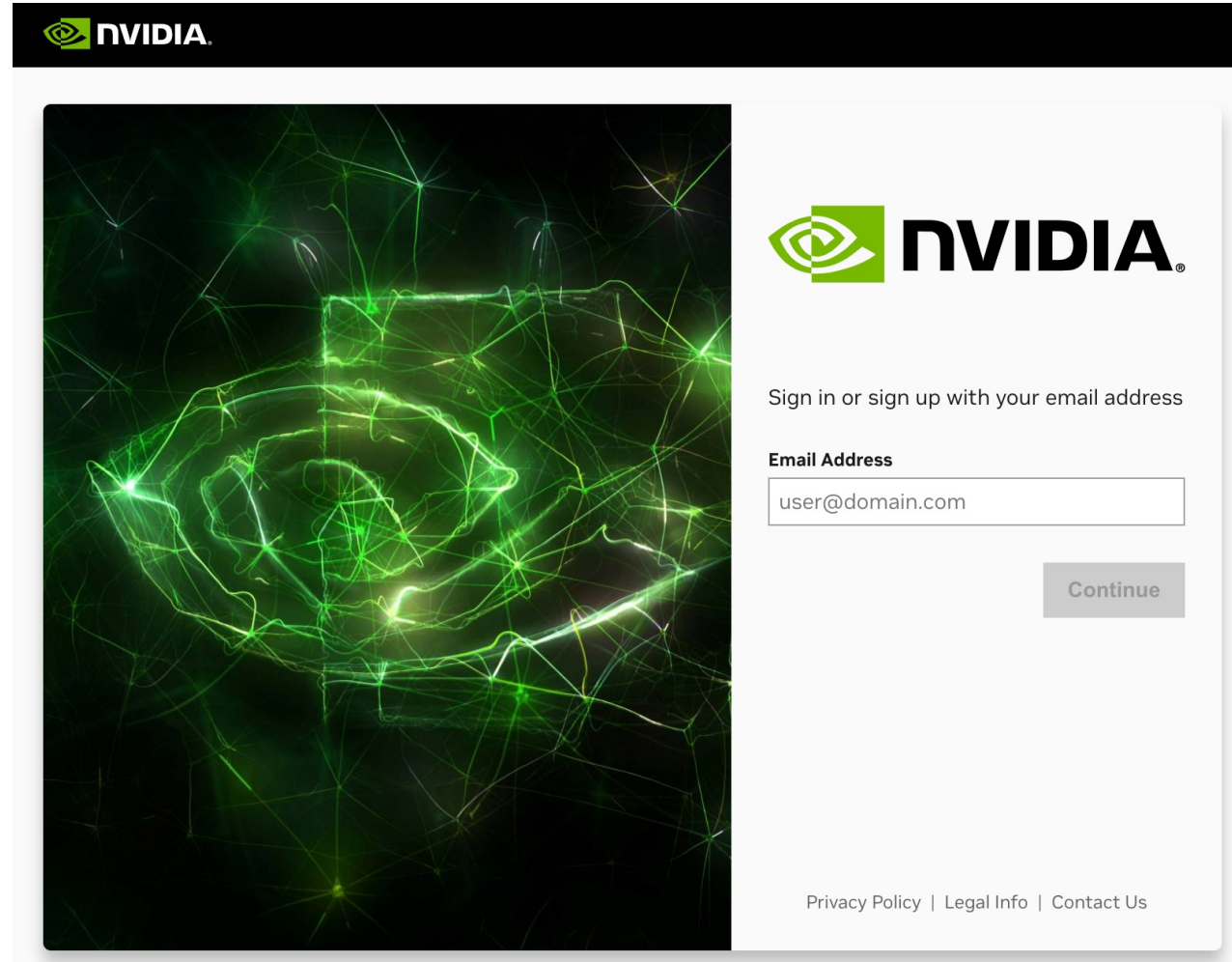
Prompt Engineering With Llama 2

Interact with and prompt engineer Llama 2 models to analyze documents, generate text, and be an AI assistant.

<https://developer.nvidia.com/join-nvidia-developer-program>

Join the NVIDIA Developer Program

take one of the complimentary technical self-paced courses (worth up to \$90)



The screenshot shows the NVIDIA Developer Program sign-up page. On the left is a large image of a glowing green wireframe cube. On the right, the NVIDIA logo is displayed above the text "Sign in or sign up with your email address". Below this is a label "Email Address" and a text input field containing "user@domain.com". A "Continue" button is positioned to the right of the input field. At the bottom right, there are links for "Privacy Policy", "Legal Info", and "Contact Us".

<https://developer.nvidia.com/join-nvidia-developer-program>

Join the NVIDIA Developer Program

take one of the complimentary technical self-paced courses (worth up to \$90)

Date of birth
1970-08-28

Password
.....

Confirm password
.....

☒ Stay logged in

☐ By proceeding, I agree to the Terms of Service

Privacy Policy | Legal Info | Contact Us

Copyright © 2024 NVIDIA Corporation

Select the images matching the toy in the example image.
If there are None, click Skip



EN

Verify

English (US) ▾

<https://developer.nvidia.com/join-nvidia-developer-program>

Join the NVIDIA Developer Program

take one of the complimentary technical self-paced courses (worth up to \$90)

 **NVIDIA DEVELOPER** DEEP LEARNING INSTITUTE PROGRAM BENEFITS

SELECT YOUR FREE COURSE.

Thank you for your participation in the NVIDIA Developer Program. Please select your free DLI course below.

English

Integrating Sensors with NVIDIA DRIVE®
Getting Started with Deep Learning
Deploying a Model for Inference at Production Scale
Get Started with Highly Accurate Custom ASR for Speech AI
Introduction to Graph Neural Networks
Introduction to Transformer-Based Natural Language Processing
Prompt Engineering with LLaMA-2 (Access Expires Dec. 5th 2025)

Generative AI with Diffusion Models
Building Real-Time Video AI Applications
Introduction to Robotic Simulations in Isaac Sim

Generative AI with Diffusion Models

Take a deeper dive into denoising diffusion models, which are a popular choice for text-to-image pipelines, with applications in creative content generation, data augmentation, simulation and planning, anomaly detection, drug discovery, personalized recommendations, and more.

 Certificate Available

 Duration: 08:00

Continue >

Join the NVIDIA Developer Program

take one of the complimentary technical self-paced courses (worth up to \$90)

 NVIDIA DEVELOPER DEEP LEARNING INSTITUTE PROGRAM BENEFITS

SELECT YOUR FREE COURSE.

Thank you for your participation in the NVIDIA Developer Program. Please select your free DLI course below.

English ▾

Modeling Time Series Data with Recurrent Neural Networks in Keras (Access ends 10/16/2024)

Optimizing CUDA Machine Learning Codes With Nsight Profiling Tools

Getting Started with Accelerated Computing in CUDA C/C++

Fundamentals of Accelerated Computing with CUDA Python

Fundamentals of Accelerated Computing with OpenACC

Integrating Sensors with NVIDIA DRIVE®

Getting Started with Deep Learning

Deploying a Model for Inference at Production Scale

Get Started with Highly Accurate Custom ASR for Speech AI

Getting Started with Deep Learning

Learn how deep learning works through hands-on exercises in computer vision and natural language processing.

 Certificate Available

 Duration: 08:00

Continue >

Join the NVIDIA Developer Program

take one of the complimentary technical self-paced courses (worth up to \$90)

 NVIDIA DEVELOPER DEEP LEARNING INSTITUTE PROGRAM BENEFITS

SELECT YOUR FREE COURSE.

Thank you for your participation in the NVIDIA Developer Program. Please select your free DLI course below.

English

Get Started with Highly Accurate Custom ASR for Speech AI

Introduction to Graph Neural Networks

Introduction to Transformer-Based Natural Language Processing

Prompt Engineering with LLaMA-2 (Access Expires Dec. 5th 2025)

Generative AI with Diffusion Models

Building Real-Time Video AI Applications

Introduction to Robotic Simulations in Isaac Sim

Introduction to Physics-informed Machine Learning with Modulus

Essentials of USD in Omniverse: Access Expires 09/18/2025

Synthetic Data Generation for Training Computer Vision Models

Generative AI with Diffusion Models

Take a deeper dive into denoising diffusion models, which are a popular choice for text-to-image pipelines, with applications in creative content generation, data augmentation, simulation and planning, anomaly detection, drug discovery, personalized recommendations, and more.

 Certificate Available

 Duration: 08:00

Continue >



NVIDIA

NVIDIA Deep Learning Institute (DLI)

Deep Learning Institute

Find Training

Self Paced Courses

Instructor-Led Workshops

Educator Programs

Enterprise Solutions

Certification

Resources

Search



Monthly Activity

Skill Points 0

Time Spent

Courses in Progress 1

Courses Completed 0

Watched Videos

Assessments

Skills

Certificates



No Certificates

You don't have any certificates yet.

Courses in Progress

Self-paced

Generative AI with Diffusion Models

0% Completed

08:00

<https://learn.nvidia.com/my-learning>

Generative AI with Diffusion Models

[Course](#) [Progress](#) [Bookmarks](#) [Updates](#)

Generative AI with Diffusion Models > Start Here > 0: Server Access

Generative AI with Diffusion Models

[Start Here](#)[Next Steps](#)[Feedback](#)[Previous](#)[Next](#)

0: Server Access

[Bookmark this page](#)

Welcome to Generative AI with Diffusion Models. Please click "Next" below to get started.

Underneath each video is a link to start your own private server for hands-on coding practice. Click the "Start" button to boot up the server. In a few minutes after the server is done loading, click "Launch" to access the code labs.

1: From U-Nets to Diffusion

[Bookmark this page](#)

Theory

<https://learn.nvidia.com/my-learning>

Deep Learning Institute (DLI)

 NVIDIA
 Products Solutions Industries For You

Shop Drivers Support

Min-Yuh Day

Deep Learning Institute
 Find Training Self Paced Courses Instructor-Led Workshops Educator Programs Enterprise Solutions Certification Resources

Search

Monthly Activity

Skill Points	0
Time Spent	
Courses in Progress	16
Courses Completed	12
Watched Videos	
Assessments	

Skills

Certificates


 Introduction to Transformer-Based Natural Language Processing


 Building RAG Agents with LLMs


 Building RAG Agents with LLMs


 Accelerating End-to-End Data Science Workflows


 Generative AI with Diffusion Models


 Building Agentic AI Applications with LLMs

Completed Courses

View more < >

Self-paced

Sizing LLM Inference Systems

100% Completed
03:00

Self-paced

Augment your LLM Using Retrieval Augmented Generation

100% Completed
01:00

Self-paced

Building RAG Agents with LLMs

100% Completed
08:00

Self-paced

Generative AI Explained

100% Completed
02:00

Self-paced

Introduction to Transform Based Natural Language Processing

100% Completed
06:00


 Introduction to Transformer-Based Natural Language Processing


 Building RAG Agents with LLMs


 Building RAG Agents with LLMs


 Accelerating End-to-End Data Science Workflows


 Generative AI with Diffusion Models


 Building Agentic AI Applications with LLMs


 Building Agentic AI Applications with LLMs


 Rapid Application Development with Large Language Models (LLMs)


 Getting Started with Deep Learning


 Generative AI with Diffusion Models


 Building LLM Applications With Prompt Engineering


 基於擴散模型的生成式人工智慧


 Fundamentals of Deep Learning


 Domain-Adaptive Pre-Training: Tailoring LLMs for Specialized Applications



Certificate of Completion

This certificate is awarded to

Min-Yuh Day

for successfully completing

Building RAG Agents with LLMs

A handwritten signature in black ink, appearing to read "Greg A. Estes", written over a horizontal line.

Greg Estes

Vice President, NVIDIA

Issue Date: : December 8, 2024

Certification ID: ed-qOCIMQaatzU8SNUNxgw |

https://learn.nvidia.com/certificates?id=ed-qOCIMQaatzU8SNUNxgw/courses/course?course_id=course-v1:DLI+S-FX-15+V1

<https://learn.nvidia.com/certificates?id=ed-qOCIMQaatzU8SNUNxgw>

All Self-Paced Courses

[Accelerated Computing](#)
[Data Science](#)
[Deep Learning](#)
[Generative AI/LLM](#)
[Graphics and Simulation](#)
[Infrastructure](#)
[Share Generative AI/LLM Courses](#)

Self-paced

Generative AI Explained

Free
02:00

Self-paced

Introduction to NVIDIA NIM™
Microservices

Free
02:00

Self-paced

Introduction to Deploying
RAG Pipelines for Production
at Scale

\$90
03:00

Self-paced

Generative AI with Diffusion
Models

\$90
08:00

Self-paced

Techniques for Improving the
Effectiveness of RAG Systems

\$30
03:00

Self-paced

Introduction to Transformer-
Based Natural Language
Processing

\$30
06:00

Self-paced

Building LLM Applications
With Prompt Engineering

\$90
08:00

Self-paced

Synthetic Tabular Data
Generation Using
Transformers

\$30
04:00

Self-paced

Sizing LLM Inference Systems

Free
03:00

Self-paced

Building RAG Agents with
LLMs

Free
08:00

Self-paced

Augment your LLM Using
Retrieval Augmented
Generation

Free
01:00

NVIDIA Deep Learning Institute (DLI)

Building RAG Agents with LLMs

Self-paced Course

Building RAG Agents with LLMs

Agents powered by large language models (LLMs) have shown great retrieval capability for using tools, looking at documents, and plan their approaches. This course will show you how to deploy an agent system in practice with the flexibility to scale up your system to meet the demands of users and customers.

[About Course](#)[Objectives](#)[Topics Covered](#)[Course Outline](#)[Stay Informed](#)[Contact Us](#)[Continue Learning](#)

About this Course

This course is free for a limited time.

The evolution and adoption of large language models (LLMs) have been nothing short of revolutionary, with retrieval-based systems at the forefront of this technological leap. These models are not just tools for automation; they are partners in enhancing productivity, capable of holding informed conversations by interacting with a vast array of tools and documents. This course is designed for those eager to explore the potential of these systems, focusing on practical deployment and the efficient implementation required to manage the considerable demands of both users and deep learning models. As we delve into the intricacies of LLMs, participants will gain insights into advanced orchestration techniques that include internal reasoning, dialog management, and effective tooling strategies.

Course Details

Duration: 08:00

Price: Free

Level: Technical - Intermediate

Subject: Generative AI/LLM

Language: English

Course Prerequisites:

Introductory deep learning knowledge, with comfort

https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Building RAG Agents with LLMs

Deep Learning Institute Find Training Self Paced Courses Instructor-Led Workshops Educator Programs Enterprise Solutions Certification Resources

Building RAG Agents with LLMs

Course Progress Bookmarks Updates

Building RAG Agents with LLMs Introduction Introduction

Building RAG Agents with LLMs

Introduction

Environment and LLMs

LangChain

Documents and Embeddings

Retrieval-Augmented
Generation

Next Steps

Feedback

Previous

Next



Building RAG Agents with LLMs

Introduction



Building RAG Agents with LLMs

Deep Learning Institute

Find Training

Self Paced Courses

Instructor-Led Workshops

Educator Programs

Enter

Training Home

My Learning

NVIDIA Account

Logout

Building RAG Agents with LLMs

Course

Progress

Bookmarks

Updates

Building RAG Agents with LLMs

Introduction

Introduction

Building RAG Agents with LLMs

Deep Learning Institute Find Training Self Paced Courses Instructor-Led Workshops Educator Programs Enterprise Solutions Certification Resources

Building RAG Agents with LLMs

Course Progress Bookmarks Updates

Building RAG Agents with LLMs Environment and LLMs Environment [0, 1, 2]

Building RAG Agents with LLMs

Introduction

Environment and LLMs

Environment [0, 1, 2]

Part 1: Course Environment

Part 2: LLM Services

LangChain

Environment [3, 4]

Part 3: LangChain

Part 4: Running States

Documents and Embeddings

Environment [5, 6]

Previous

Next

 Bookmark this page

Welcome to **Building RAG Agents with LLMs**. In this first section, we will get introduced to the overall course environment, LLM services, and recommended workflows!

This tab contains the course environment for this section, which will contain the notebooks for the next two videos! Please click through the videos in the remaining tabs to watch the material and work through the exercises!

Please click the "Start" button to start up your own private server for hands-on coding practice. It will take a few minutes to start up, so go ahead and click it now and then proceed to the next video! After a few minutes when the server has loaded, click "Launch" to access the code labs.




START

Building RAG Agents with LLMs

Deep Learning Institute

Find Training

Self Paced Courses

Instructor-Led Workshops

Educator Programs

Enterprise Solutions

Certification

Building RAG Agents with LLMs

Course

Progress

Bookmarks

Updates

Building RAG Agents with LLMs

Environment and LLMs

Environment [0, 1, 2]

Building RAG Agents with LLMs

Introduction

Introduction

Course Slides

Environment and LLMs

Environment [0, 1, 2]

Part 1: Course Environment

Part 2: LLM Services

LangChain

Environment [3, 4]

Part 3: LangChain

Previous

Next

welcome to **Building RAG Agents with LLMs**. In this first section, we will get introduced to the overall course environment, LLM services, and recommended workflows!

This tab contains the course environment for this section, which will contain the notebooks for the next two videos! Please click through the videos in the remaining tabs to watch the material and work through the exercises!

Please click the "Start" button to start up your own private server for hands-on coding practice. It will take a few minutes to start up, so go ahead and click it now and then proceed to the next video! After a few minutes when the server has loaded, click "Launch" to access the code labs.



DEEP
LEARNING
INSTITUTE

This Lab 0 : 01 : 06 / 2 : 00 : 00

Course 13 : 45 : 51 / 32 : 00 : 00



LAUNCH



STOP TASK

Building RAG Agents with LLMs

← → ↻ ⚠ Not Secure 54.211.119.108/lab/lab ☆ 📁 🔍 ⋮

File Edit View Run Kernel Tabs Settings Help

+

📁

⬆

↻

Filter files by name 🔍

/

Name	Last Modified
chatbot	yesterday
composer	yesterday
docker_router	yesterday
frontend	yesterday
imgs	yesterday
llm_client	yesterday
slides	yesterday
00_jupyterlab.ipynb	yesterday
01_microservices.ipynb	yesterday
02_llms.ipynb	yesterday
99_table_of_contents.ipynb	yesterday

🔖 Launcher

+

⚙

🛡

🔖 Notebook

Python 3 (ipykernel)

> Console

Python 3 (ipykernel)

\$ Other

Terminal

Text File

Markdown File

Python File

Show Contextual Help

Simple ☐ 0 \$ 0 ⚙

Launcher 1 🔔

Building RAG Agents with LLMs

The screenshot displays the JupyterLab web interface in a browser. The address bar shows a URL: 54.211.119.108/lab/lab/tree/00_jupyterlab.ipynb. The interface includes a menu bar (File, Edit, View, Run, Kernel, Tabs, Settings, Help), a file browser on the left, and a main work area with a launcher tab and a notebook tab.

File Browser (Left Sidebar):

Name	Last Modified
chatbot	yesterday
composer	yesterday
docker_router	yesterday
frontend	yesterday
imgs	yesterday
llm_client	yesterday
slides	yesterday
00_jupyterlab.ipynb	yesterday
01_microservices.ipynb	yesterday
02_llms.ipynb	yesterday
99_table_of_contents.ipynb	yesterday

Launcher Tab: Displays the NVIDIA logo and the text "DEEP LEARNING INSTITUTE".

Notebook Tab (00_jupyterlab.ipynb):

Notebook 0: JupyterLab

We use [JupyterLab](#) to manage our environment for this hands-on lab. The [JupyterLab Interface](#) is a dashboard that provides access to interactive iPython notebooks, as well as the folder structure of our environment and a terminal window into the Ubuntu operating system. The view includes a **menu bar** at the top, a **file browser** in the **left sidebar**, and a **main work area** initially open to the "Launcher" page.

A green arrow points to the menu bar at the bottom of the interface.

Building RAG Agents with LLMs

Environment and LLMs

Environment [0, 1, 2]

Part 1: Course Environment

Part 2: LLM Services

LangChain

Environment [3, 4]

Part 3: LangChain

Part 4: Running States

Documents and Embeddings

Retrieval-Augmented
Generation

Next Steps

Feedback



This Lab 0 : 10 : 32 / 2 : 00 : 00



Course 13 : 55 : 17 / 32 : 00 : 00

LAUNCH



STOP TASK



Stopping Task

Click confirm to stop this task.

Confirm

Cancel

Building RAG Agents with LLMs

Building RAG Agents with LLMs

Introduction

Environment and LLMs

LangChain

Documents and Embeddings

Environment [5, 6]

Part 5: Documents

Part 6: Embeddings

Retrieval-Augmented Generation

Environment [7, 8, Assessment]

Part 7: Vector Stores

Part 8: Evaluation

Next Steps

Previous

Next

Bookmark this page

In this section, we will combine all of our prior efforts to integrate and evaluate retrieval-augmented generation pipelines! Along the way, you will also get the opportunity to work through the assessment, which will involve Gradio, LangServe, FAISS, RAG, and Evaluation! **Good Luck!**

Please click the "Start" button to start up your own private server for hands-on coding practice. It will take a few minutes to start up, so go ahead and click it now and then proceed to the next video! After a few minutes when the server has loaded, click "Launch" to access the code labs.



This Lab 0 : 15 : 39 / 4 : 00 : 00

Course 14 : 12 : 18 / 32 : 00 : 00



LAUNCH



STOP TASK



ASSESS TASK

https://learn.nvidia.com/courses/course?course_id=course-v1:DLI+S-FX-15+V1



NVIDIA.

NIM (NVIDIA Inference Microservice)

<https://build.nvidia.com/>

The screenshot shows the NVIDIA NIM website interface. At the top, there is a navigation bar with links for Explore, Models, Blueprints, GPUs, and Docs. A search bar is also present. The main heading is "Start Building Your AI Here." Below this, there are two main sections: "Use Inference Endpoints" and "Launch a GPU Instance".

Use Inference Endpoints
Free inference with leading models

Launch a GPU Instance
Prototype in a sandbox

Inference Endpoints:

Model	Capabilities	Count
qwen qwen2.5-coder-7b-instruct	code completion	+2
nvidia nvidia-nemotron-nano-9b-v2	reasoning	+1
deepseek-ai deepseek-v3.1	reasoning	+1
openai gpt-oss-120b	math	+3

GPU Instances:

Instance	GPU
B200	NVIDIA B200
H200	NVIDIA H200
H100	NVIDIA H100
L40S	NVIDIA L40S

At the bottom, there are two buttons: "More Models >" and "More Instances >".



Building RAG Agents with LLMs

Introduction



Building RAG Agents with LLMs

Introduction

Course Environment

 EDIT     

 DEEP
LEARNING
INSTITUTE


START

Please click the start button to the top-right. The course will take about 5 minutes to load. The course lecture slide decks will appear here when you launch the course. The numbers on the lectures correspond to the numbers of the interactive notebooks in the course.

Large Language Models

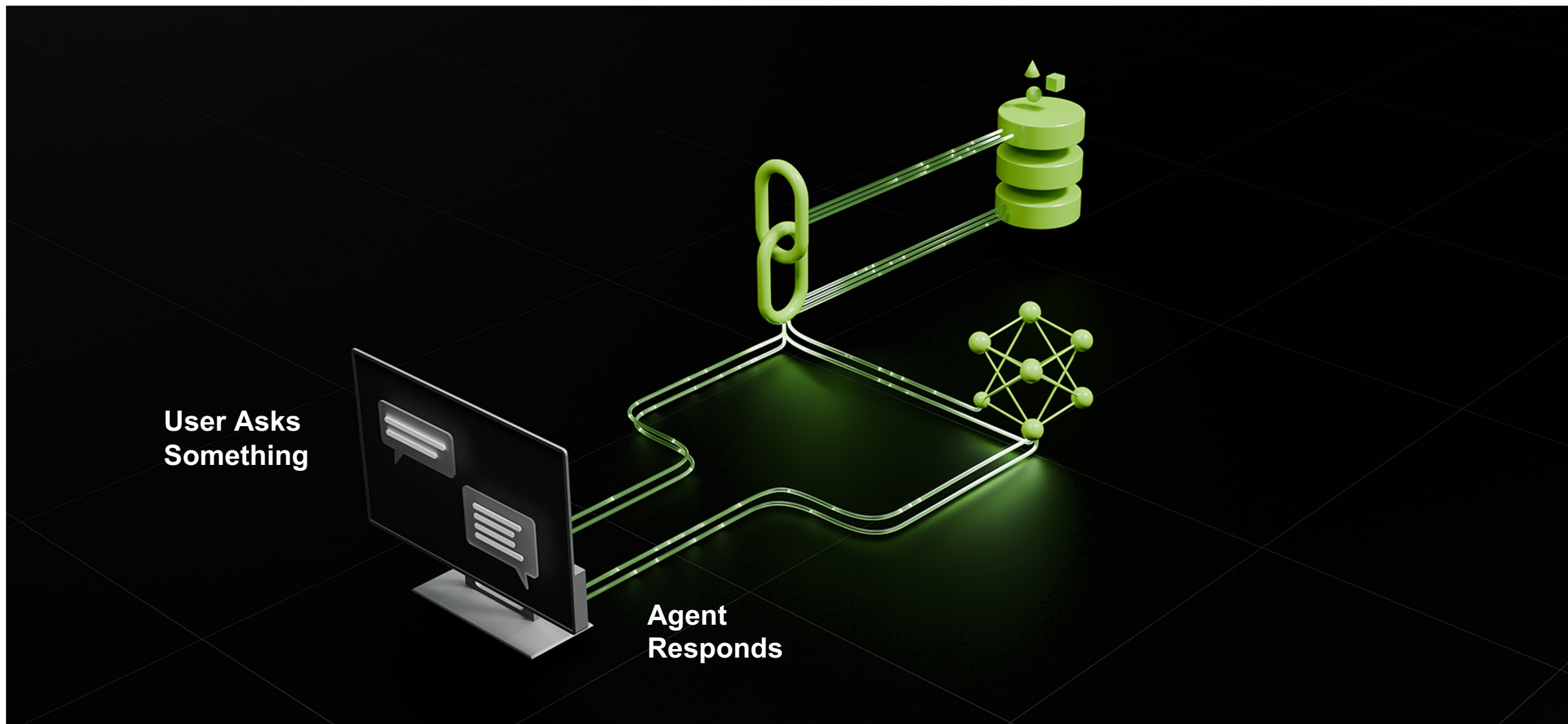
Backbones for Language Understanding



Backbones for language tasks, including classification and generation.

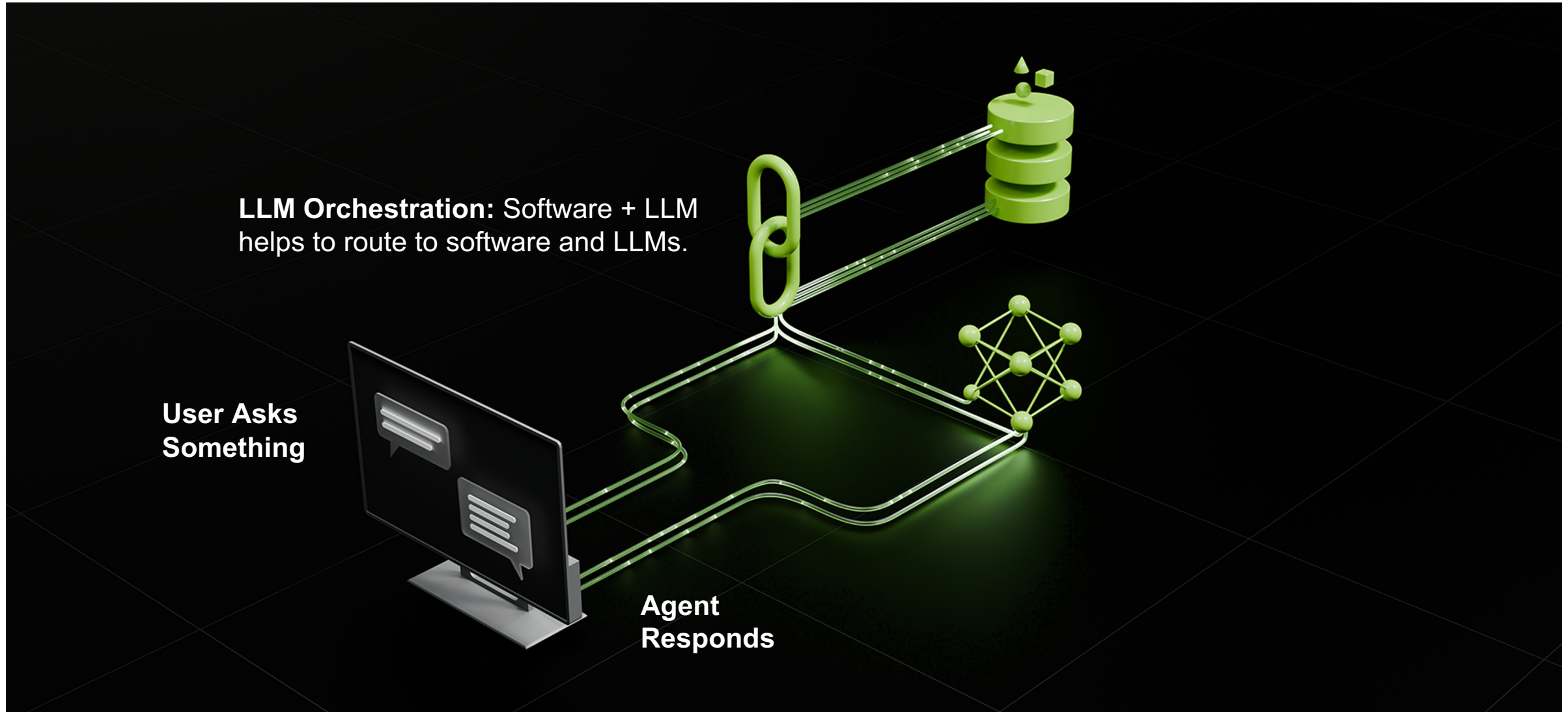
Dialog/Retrieval Agents

LLMs with Context and Control



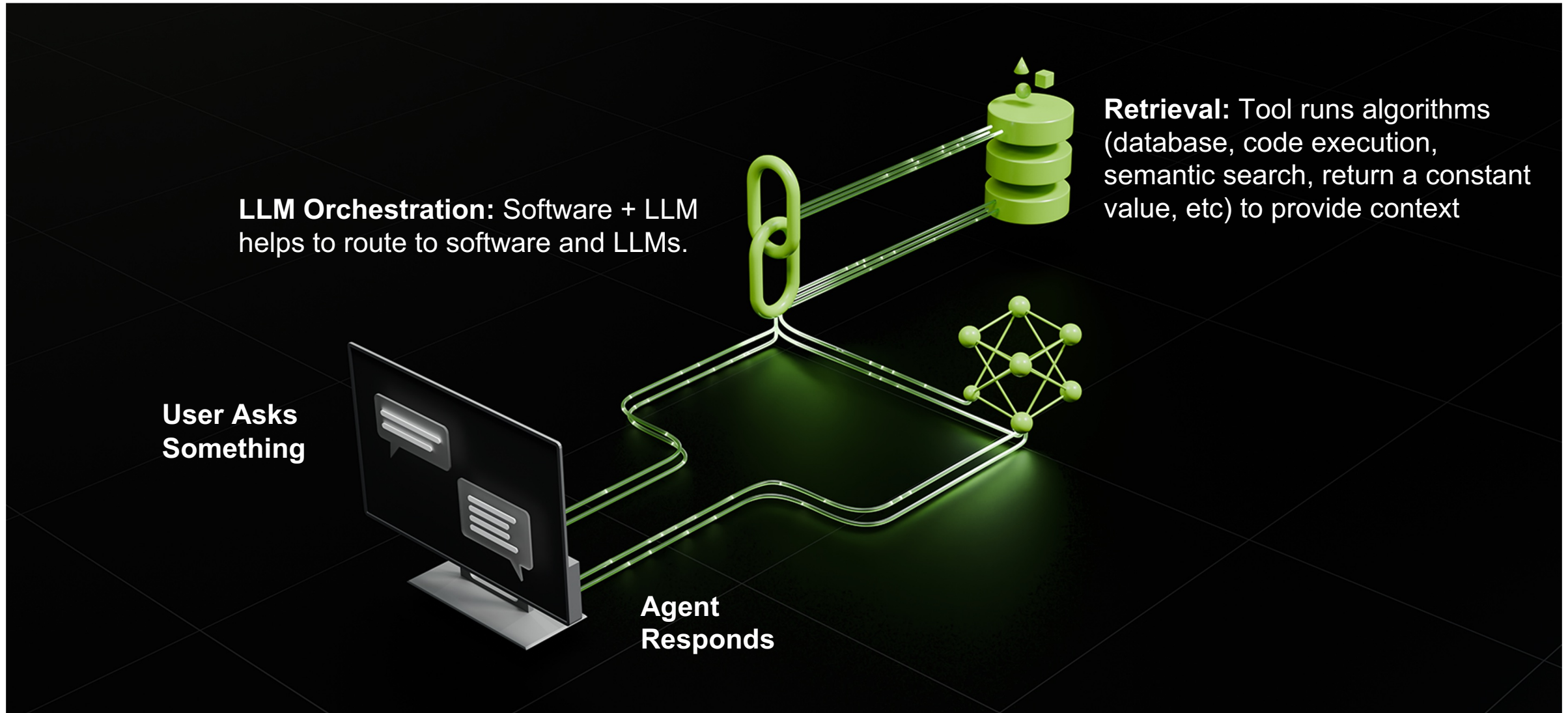
Dialog/Retrieval Agents

LLMs with Context and Control



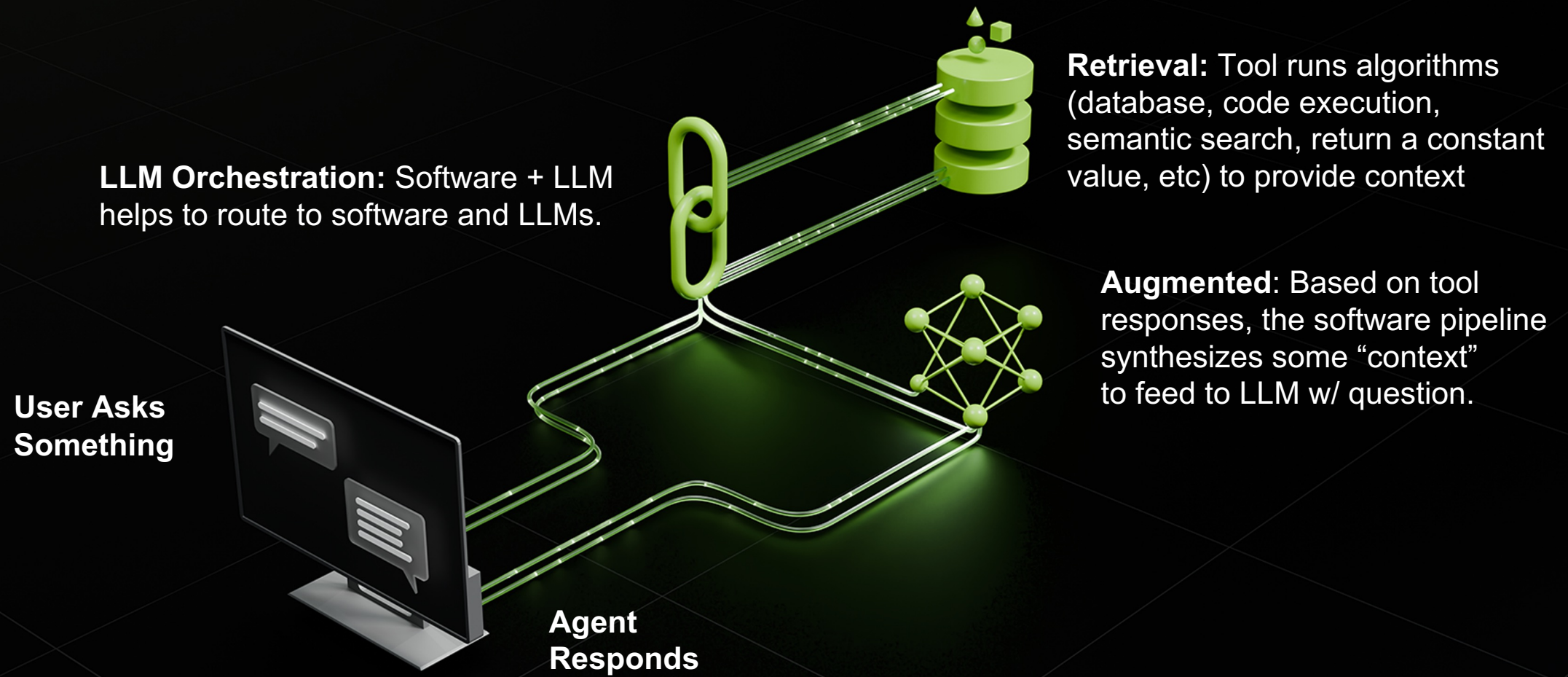
Dialog/Retrieval Agents

LLMs with Context and Control



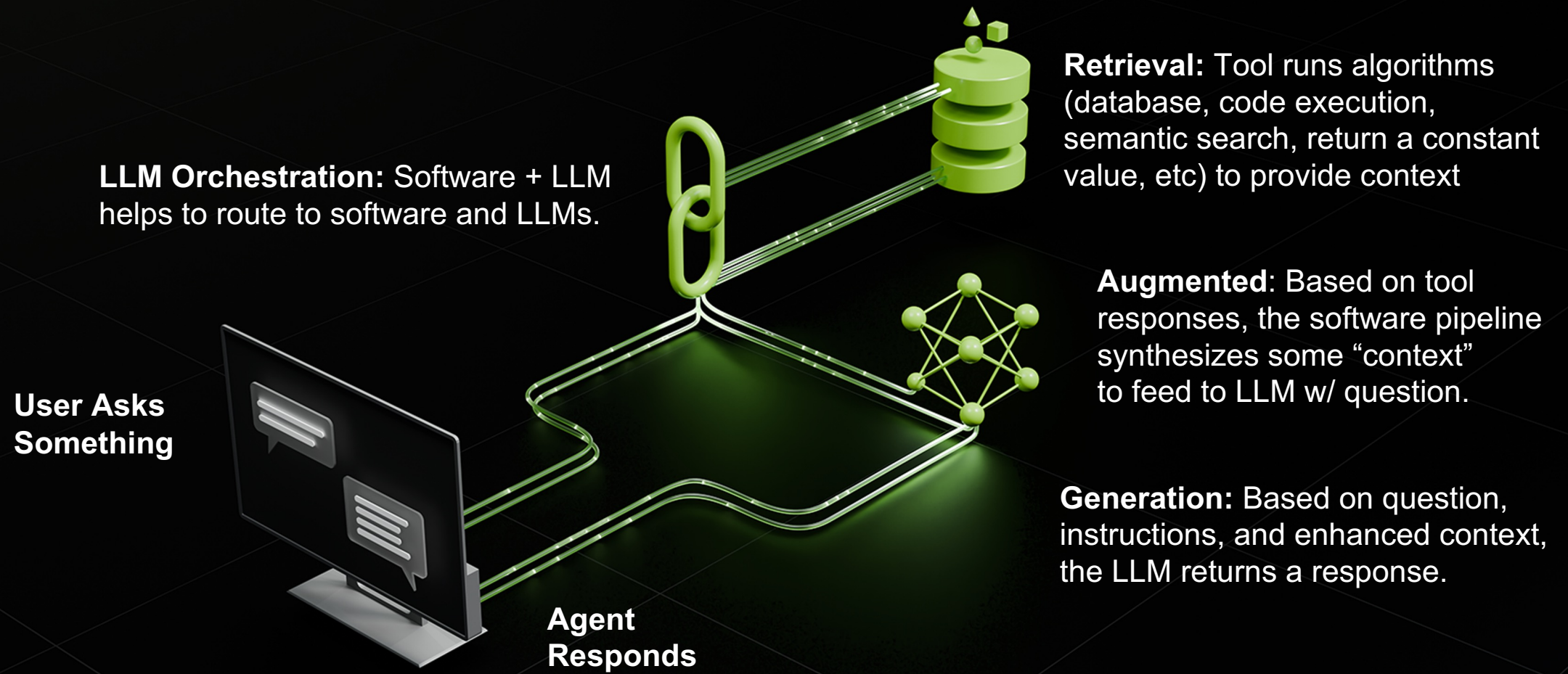
Dialog/Retrieval Agents

LLMs with Context and Control



Dialog/Retrieval Agents

LLMs with Context and Control



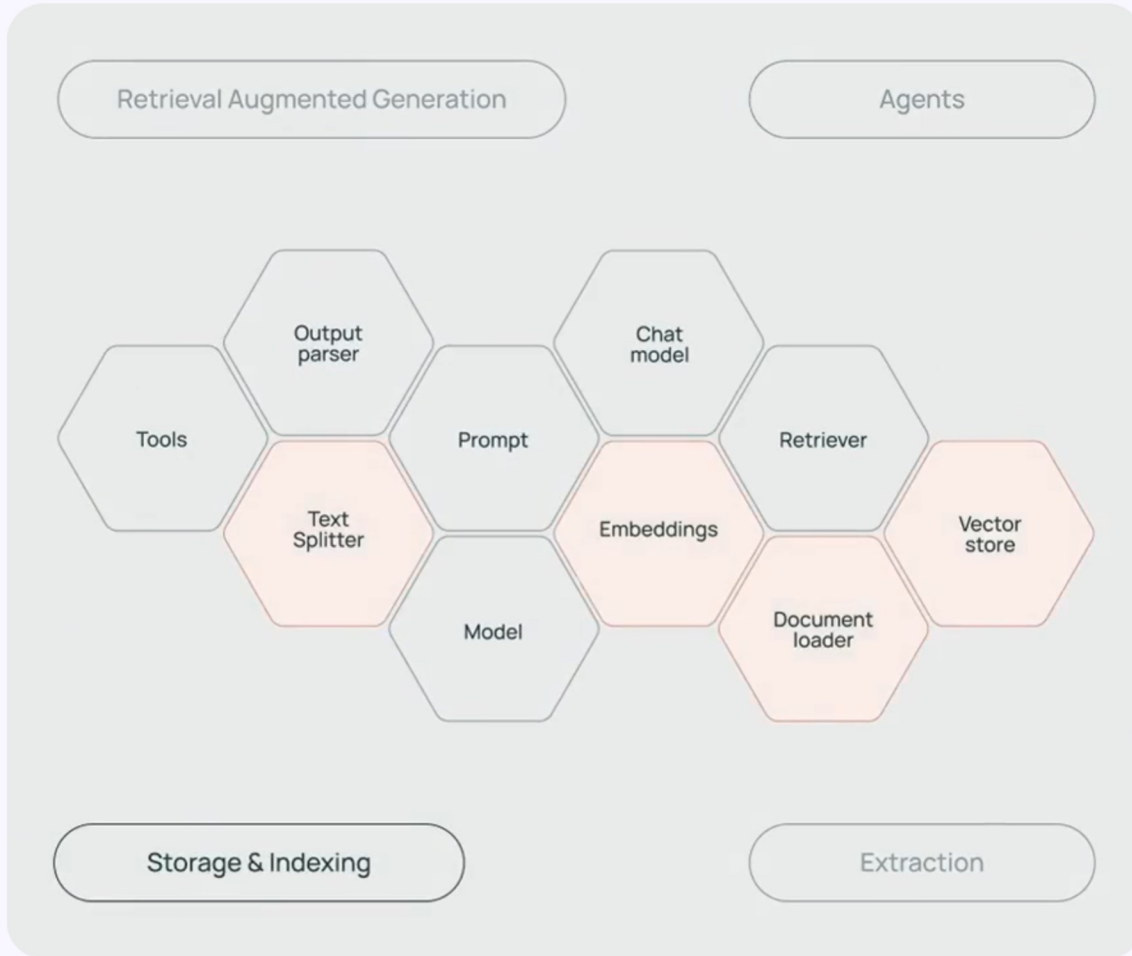
Chat Applications

Full Applications Build with LLMs



Chat Applications

Full Applications Build with LLMs



Build your app with LangChain

Build context-aware, reasoning applications with LangChain's flexible framework that leverages your company's data and APIs. Future-proof your application by making vendor optionality part of your LLM infrastructure design.

[Learn more about LangChain ↗](#)

Chat Applications

Full Applications Build with LLMs

Instantly Run and Deploy Generative AI

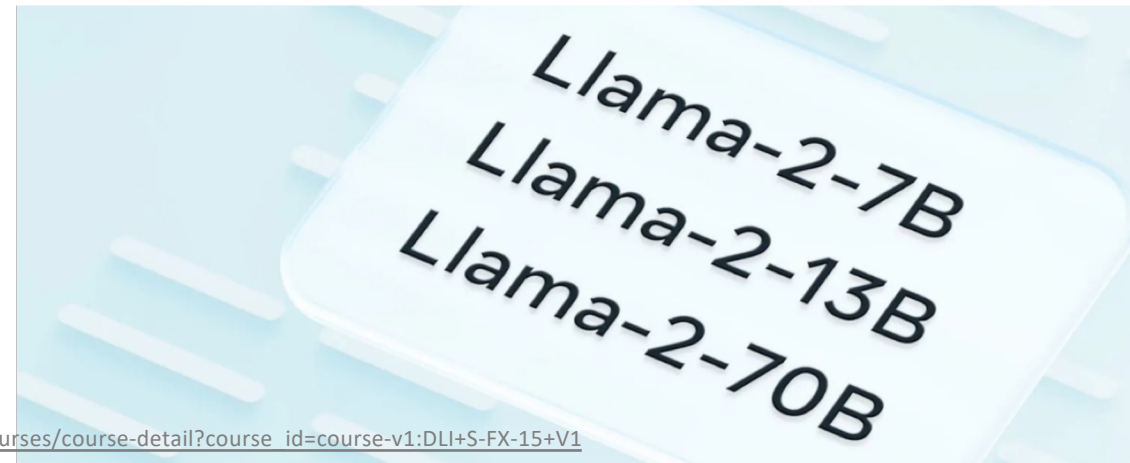
Explore the latest community-built AI models with APIs optimized and accelerated by NVIDIA, then deploy anywhere with NVIDIA NIM.



Prerequisites

RAG Agents in Production

- Prior LLM/LangChain Exposure
- Intermediate Python Experience
- Exposure to Web Engineering



Course Objectives

RAG Agents in Production

1. Environment
2. LLM Services
3. Intro to LangChain (LCEL)
4. Running State Chains
5. Document Loading
6. Embeddings
7. Document Retrieval
8. RAG Evaluation



Building RAG Agents with LLMs

Part 1: Your Course Environment

Part 2: LLM Services

Part 3: Intro to LangChain

Part 4: Running State Chain

Part 5: Working with Documents

Part 6: Embedding Model for Retrieval

Part 7: Document Retrieval with Vector Databases

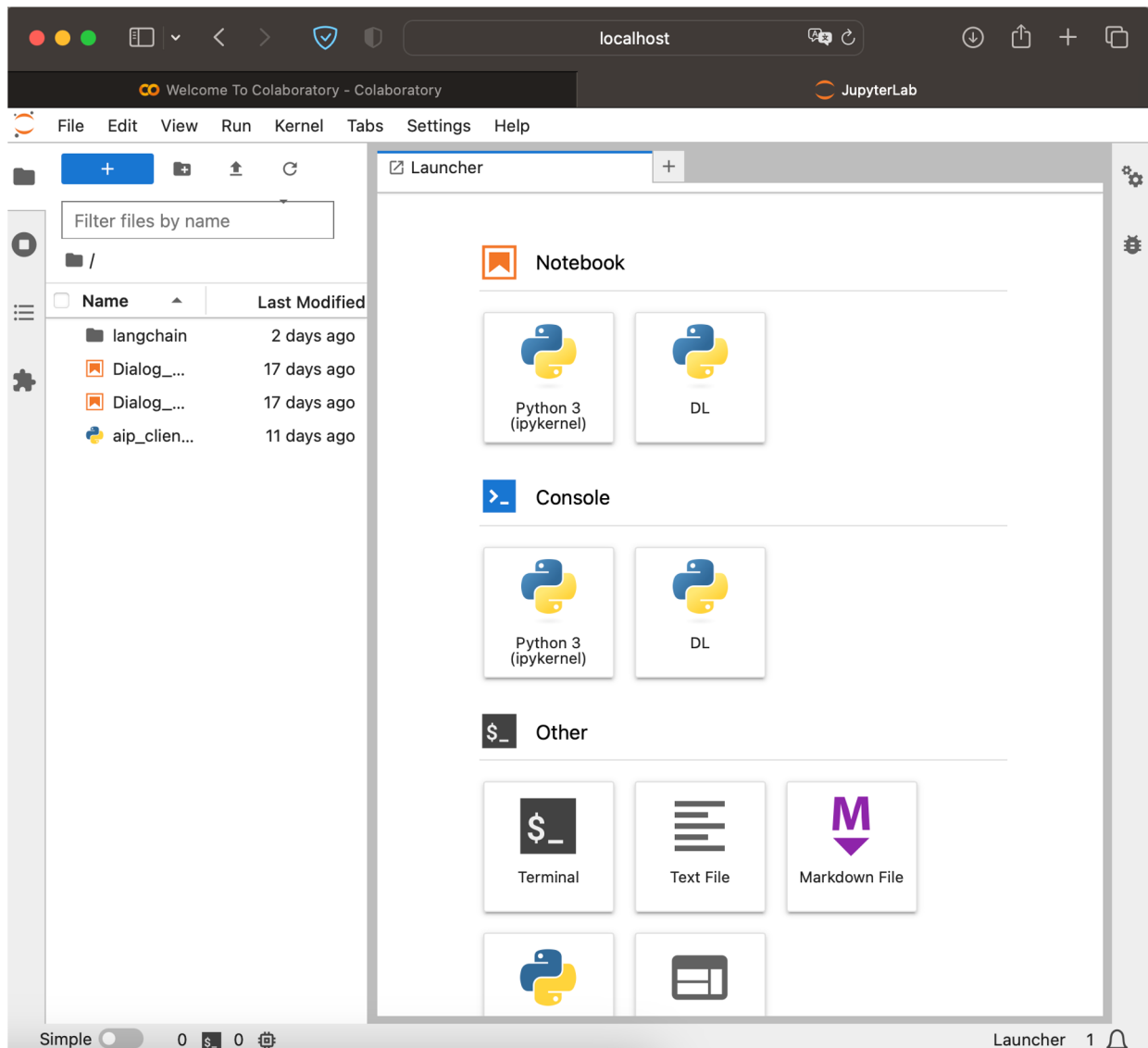
Part 8: RAG Evaluation



Building RAG Agents with LLMs

Part 1: Your Course Environment

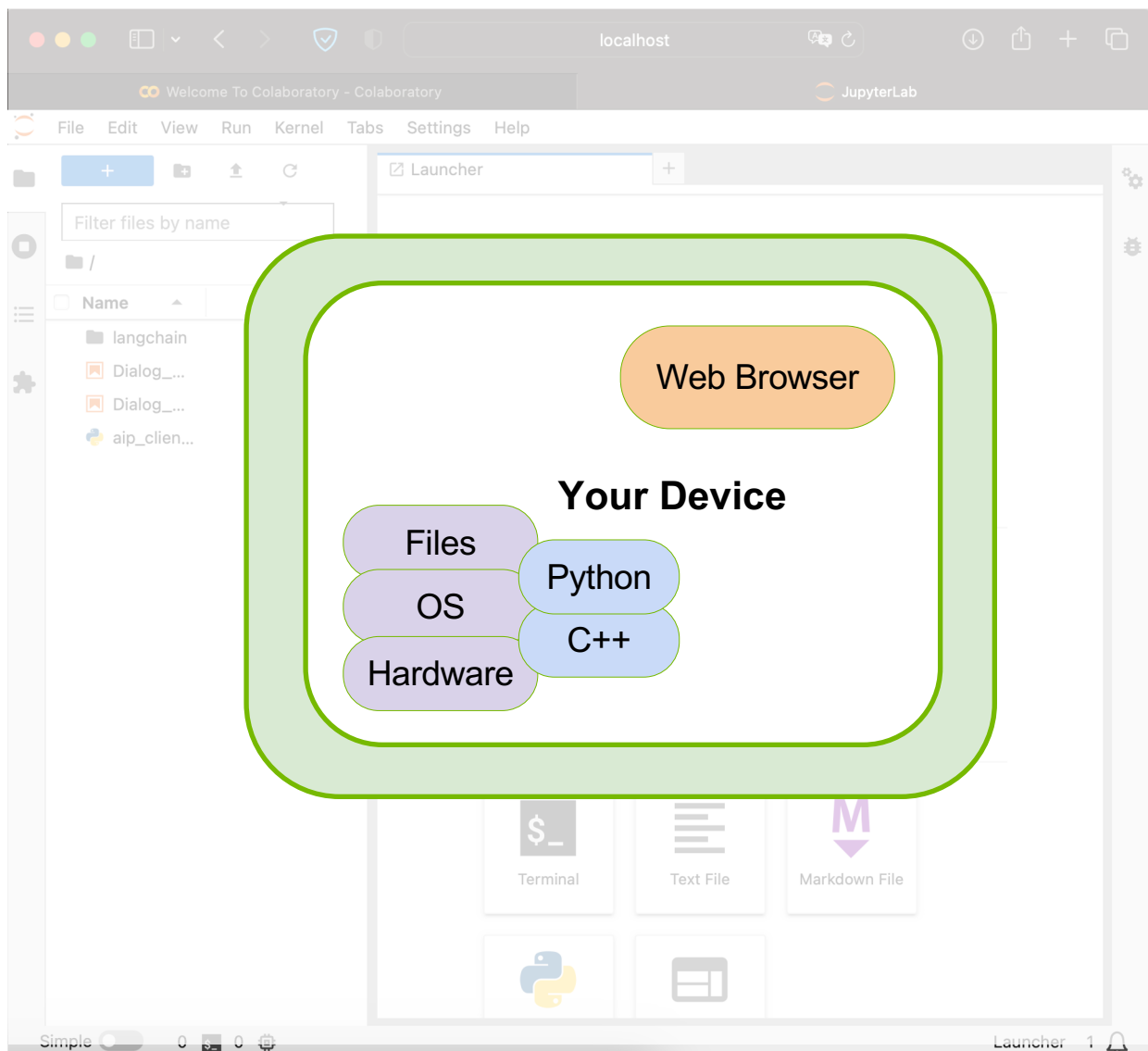
Typical Jupyter Labs Interface



`jupyter lab . &`; open -a Safari <http://localhost:8888/lab>

Source: NVIDIA DLI (2024), Building RAG Agents with LLMs, https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

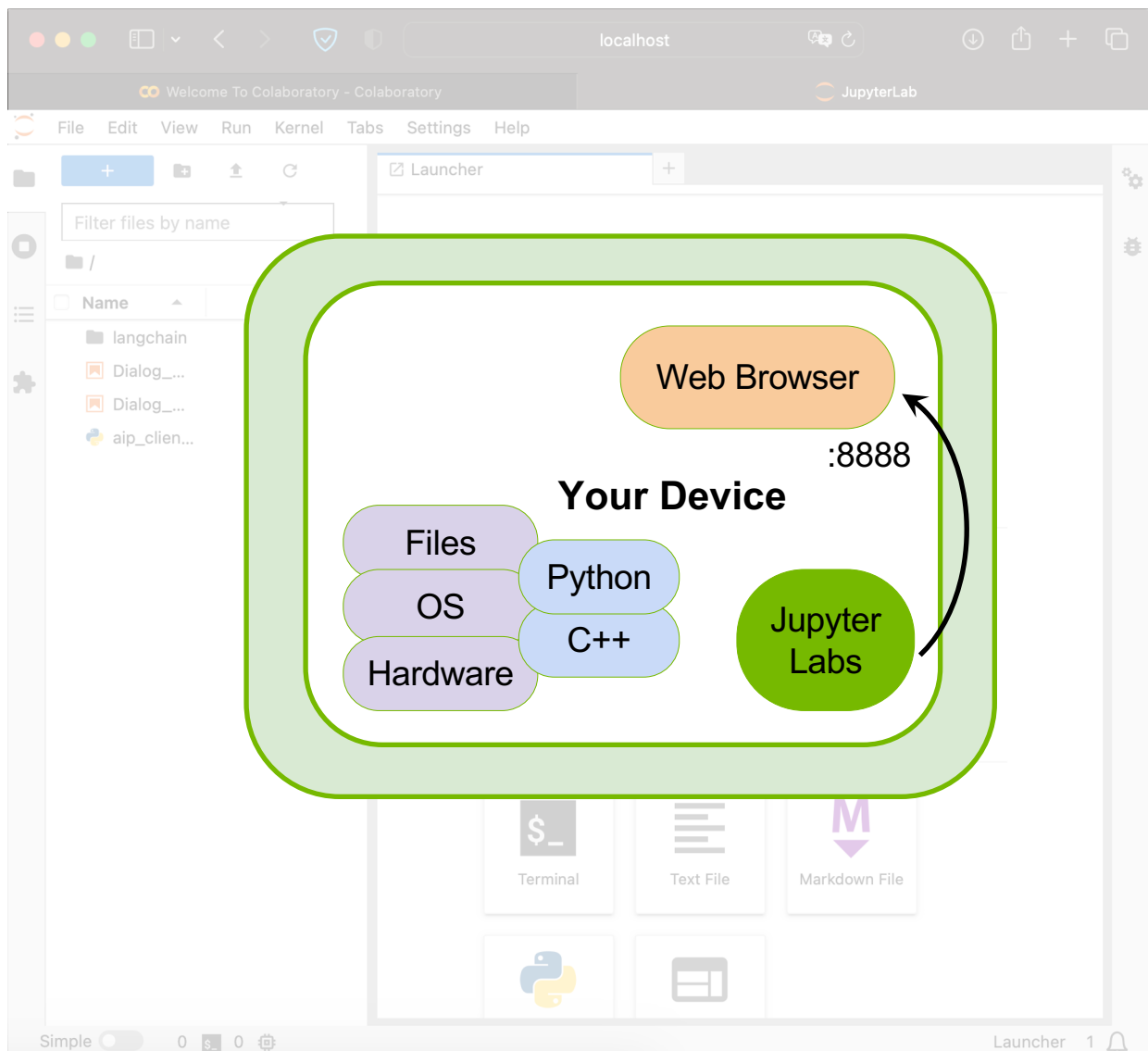
Typical Jupyter Labs Interface



jupyter lab . & open -a Safari <http://localhost:8888/lab>

Source: NVIDIA DLI (2024), Building RAG Agents with LLMs, https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

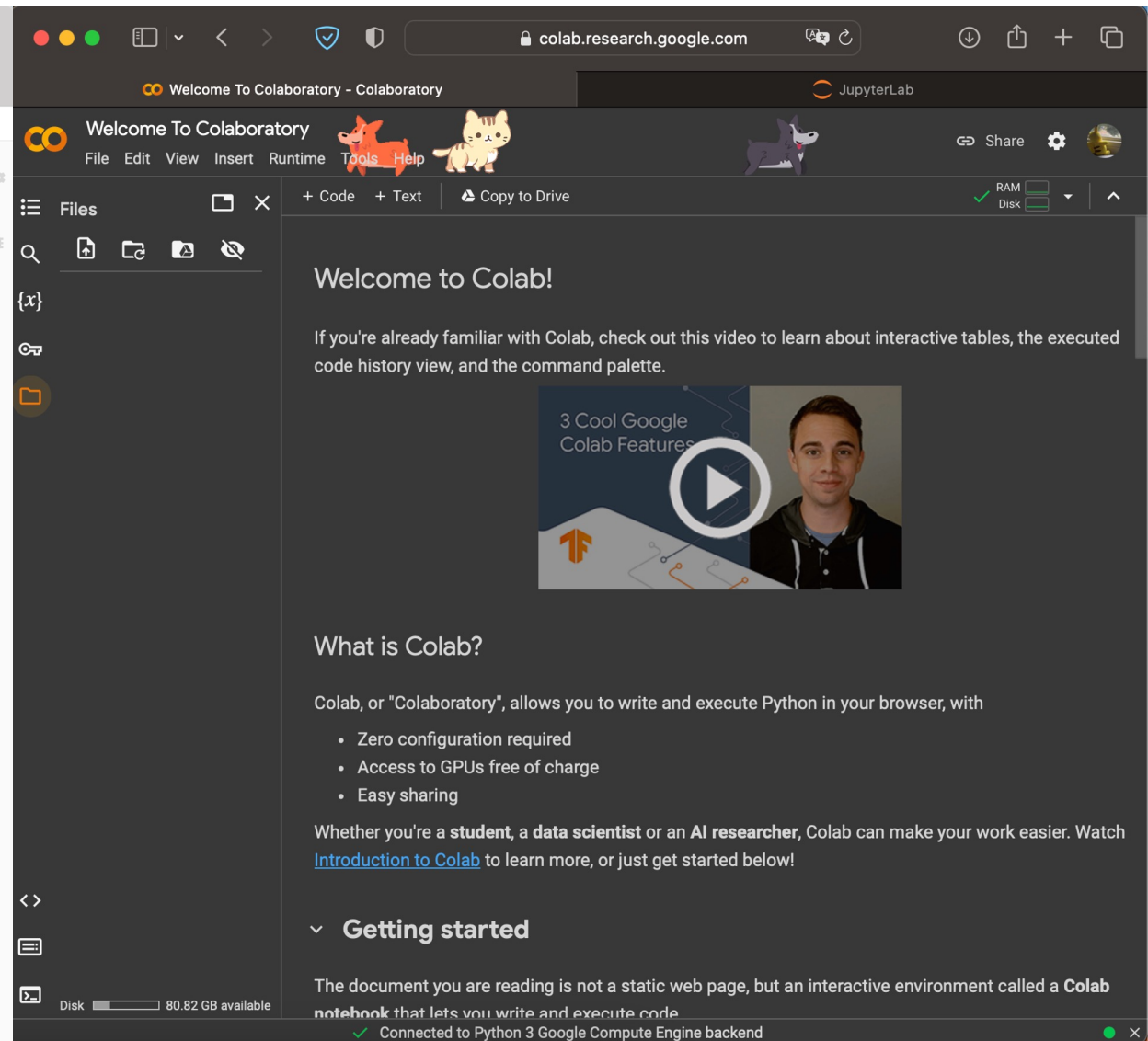
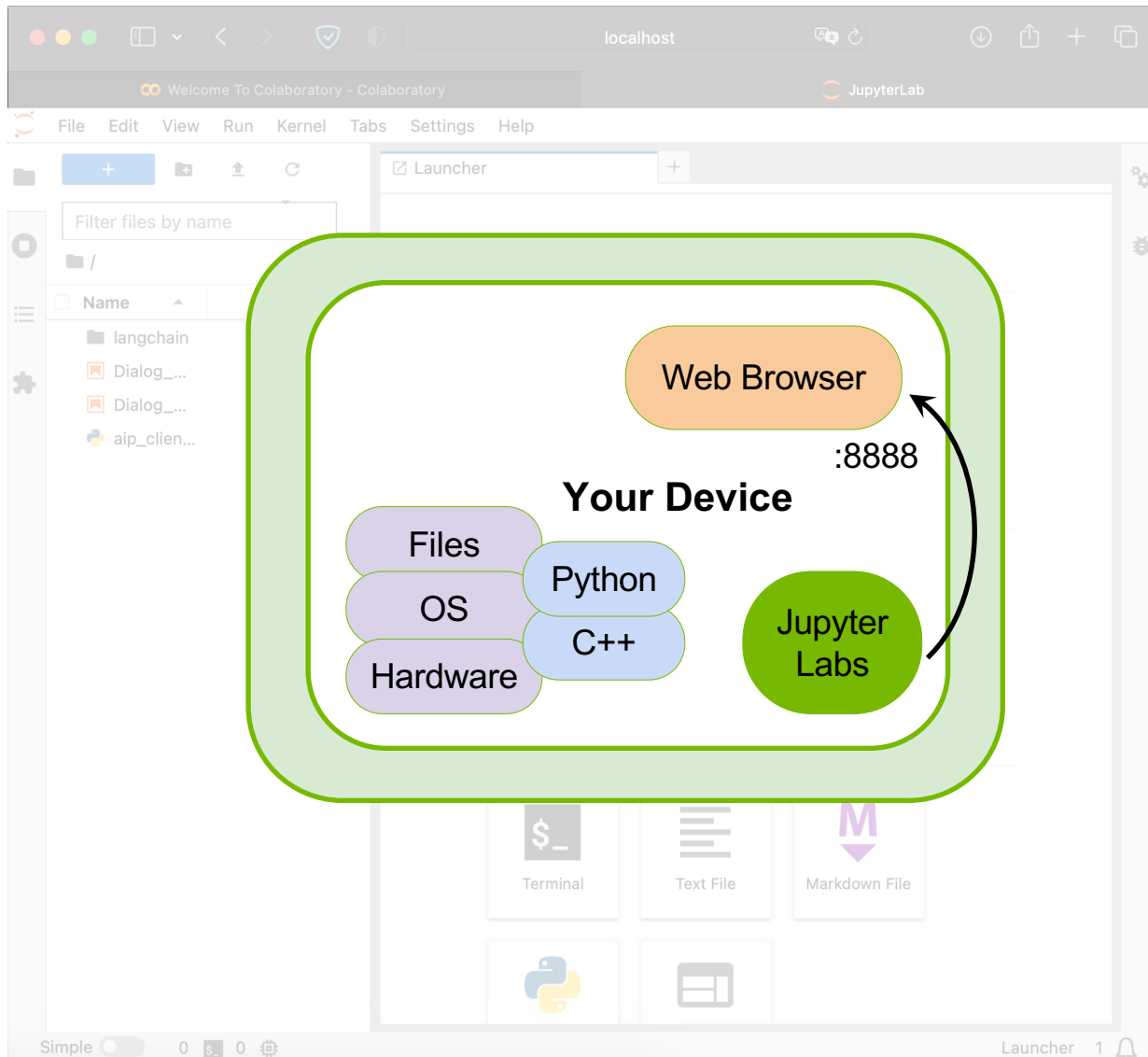
Typical Jupyter Labs Interface



jupyter lab . & open -a Safari <http://localhost:8888/lab>

Source: NVIDIA DLI (2024), Building RAG Agents with LLMs, https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Typical Jupyter Labs Interface

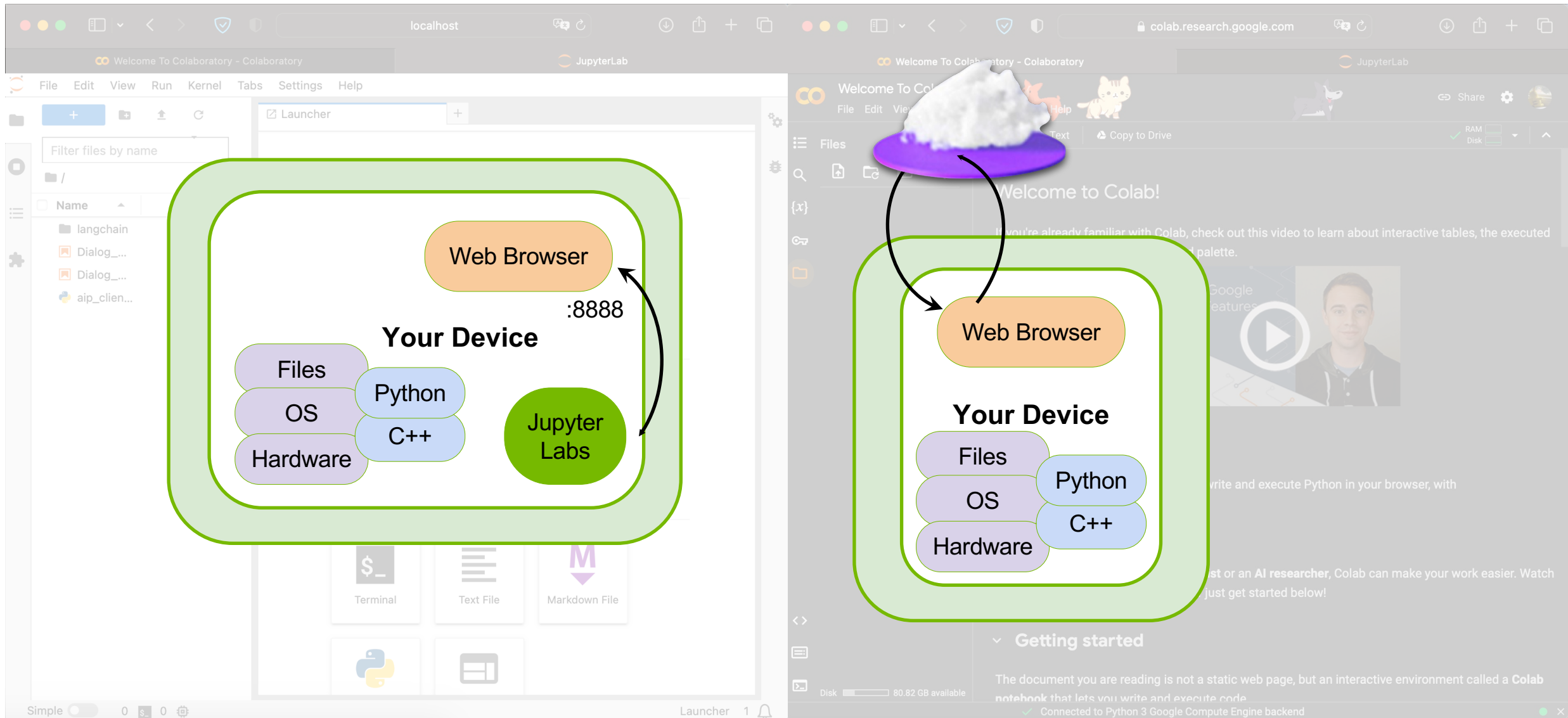


jupyter lab . & open -a Safari <http://localhost:8888/lab>

<https://colab.research.google.com>

Source: NVIDIA DLI (2024), Building RAG Agents with LLMs, https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Typical Jupyter Labs Interface

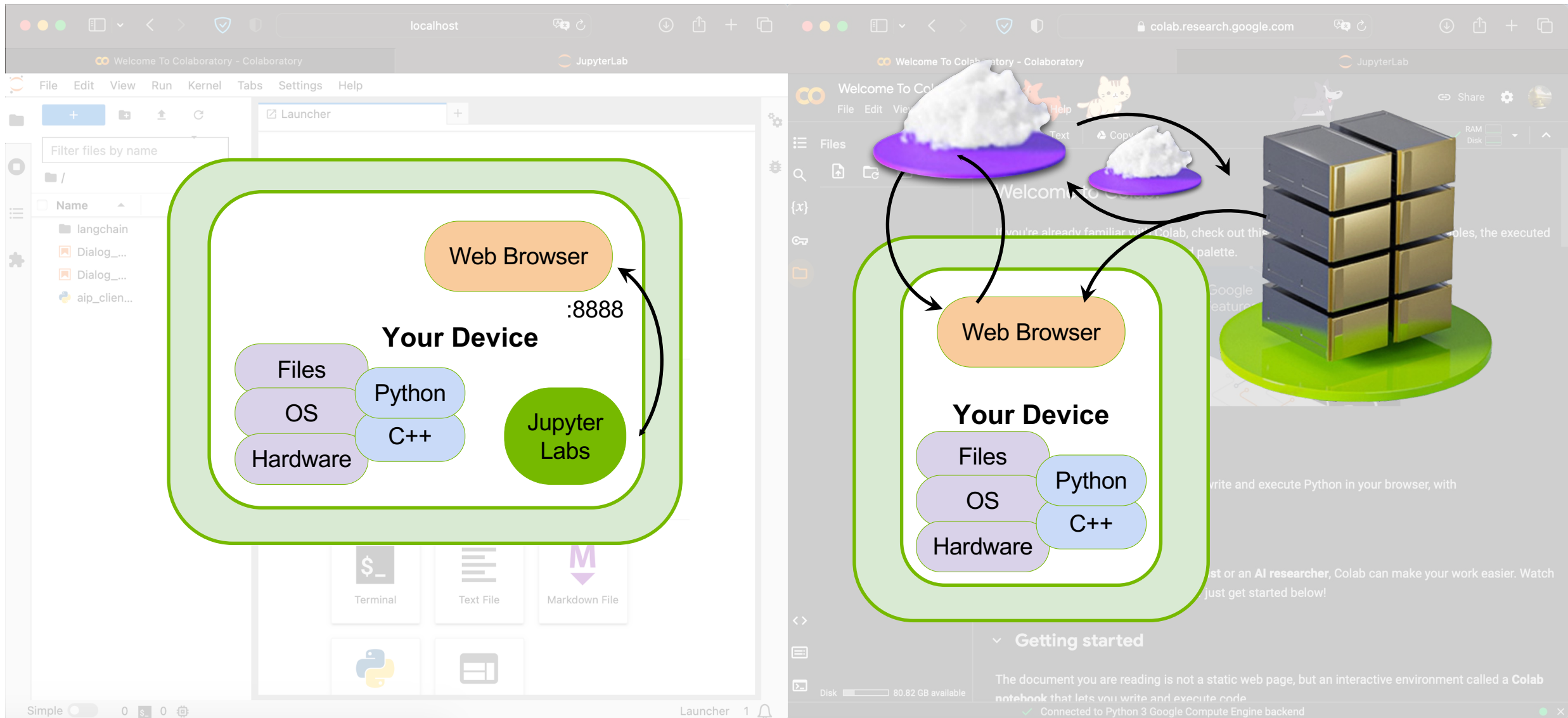


jupyter lab . & open -a Safari <http://localhost:8888/lab>

<https://colab.research.google.com>

Source: NVIDIA DLI (2024), Building RAG Agents with LLMs, https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

Typical Jupyter Labs Interface

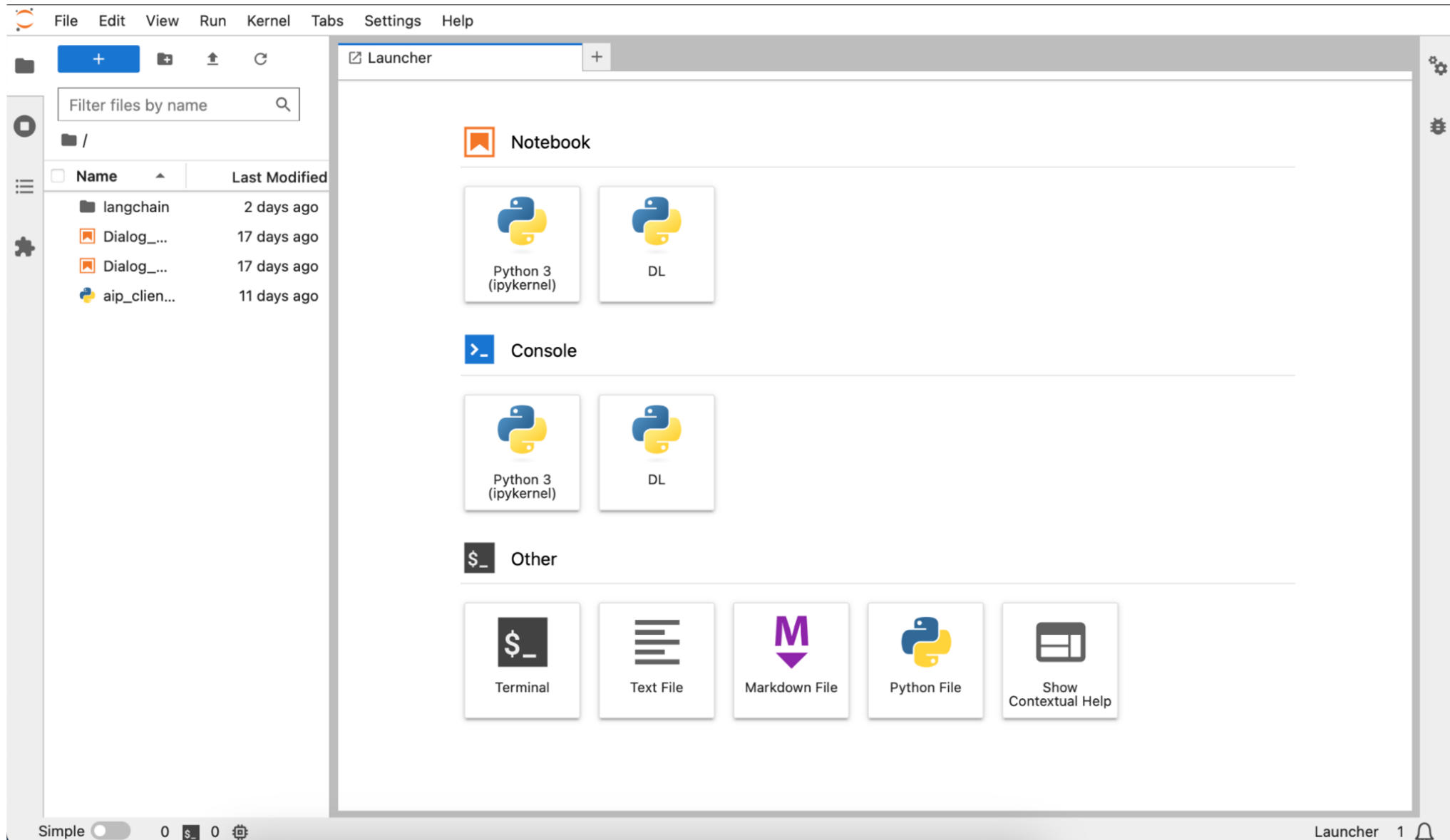


jupyter lab . & open -a Safari <http://localhost:8888/lab>

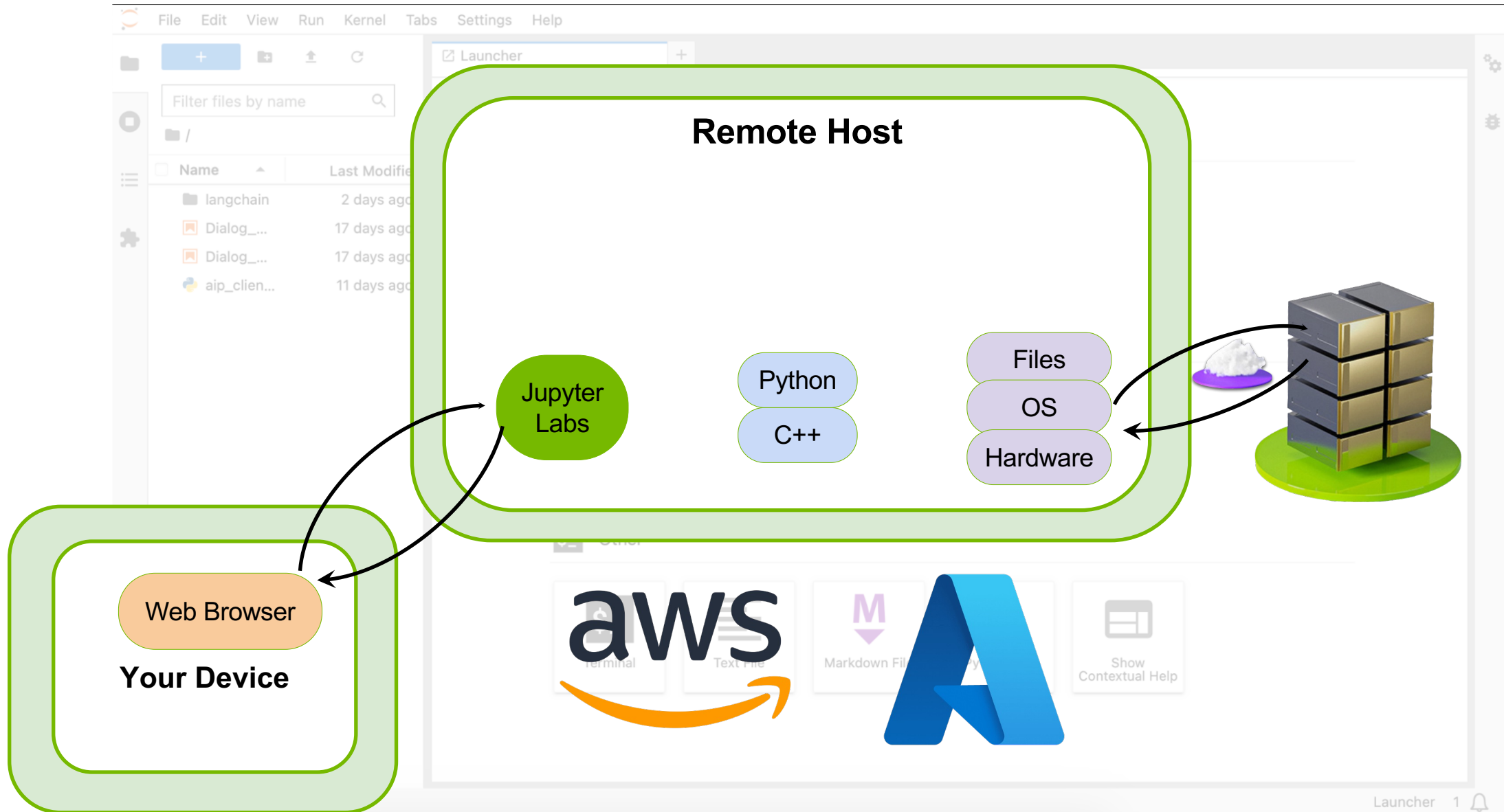
<https://colab.research.google.com>

Source: NVIDIA DLI (2024), Building RAG Agents with LLMs, https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1

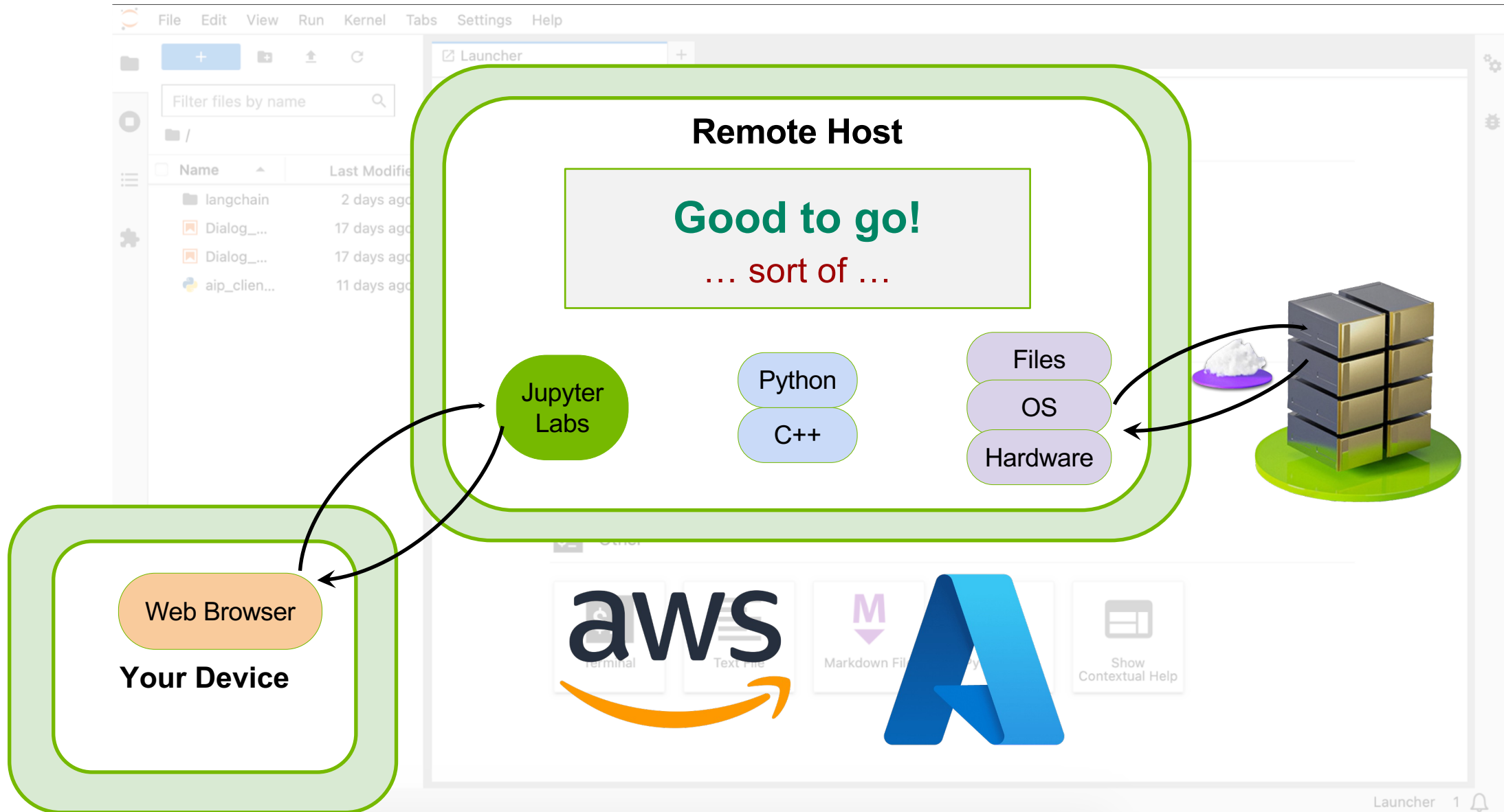
DLI Jupyter Labs Interface



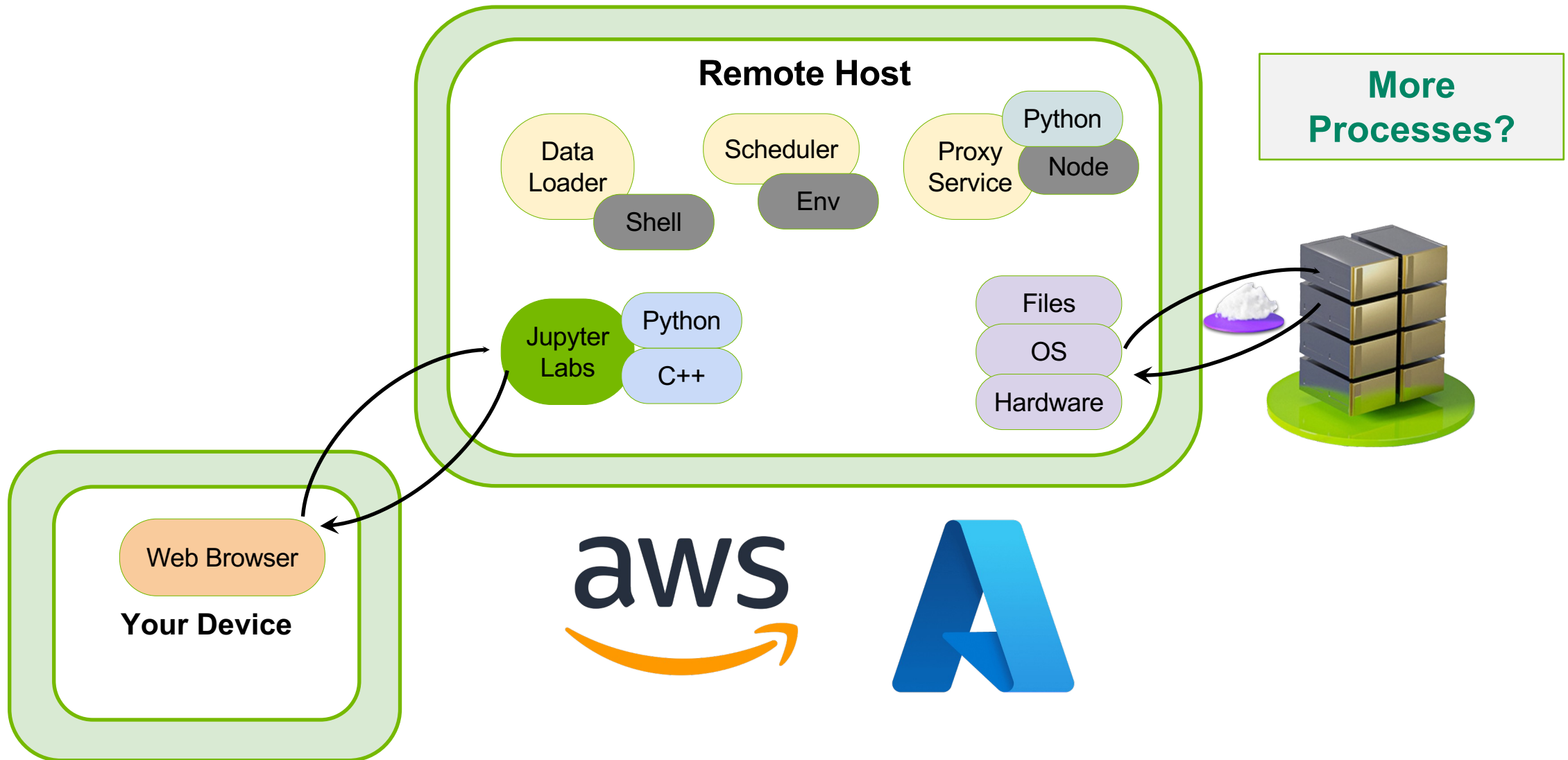
DLI Jupyter Labs Interface



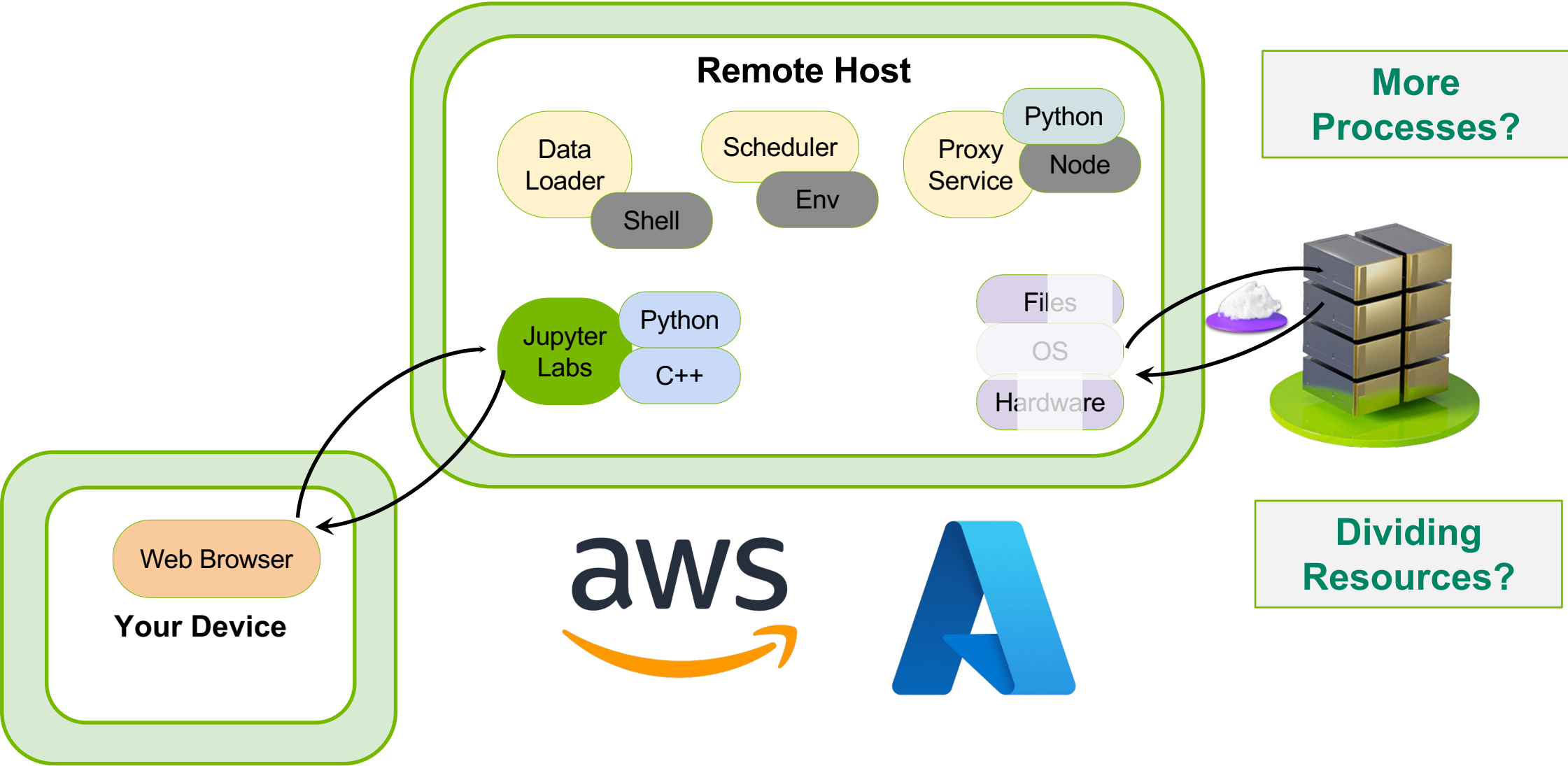
DLI Jupyter Labs Interface



DLI Jupyter Labs Interface

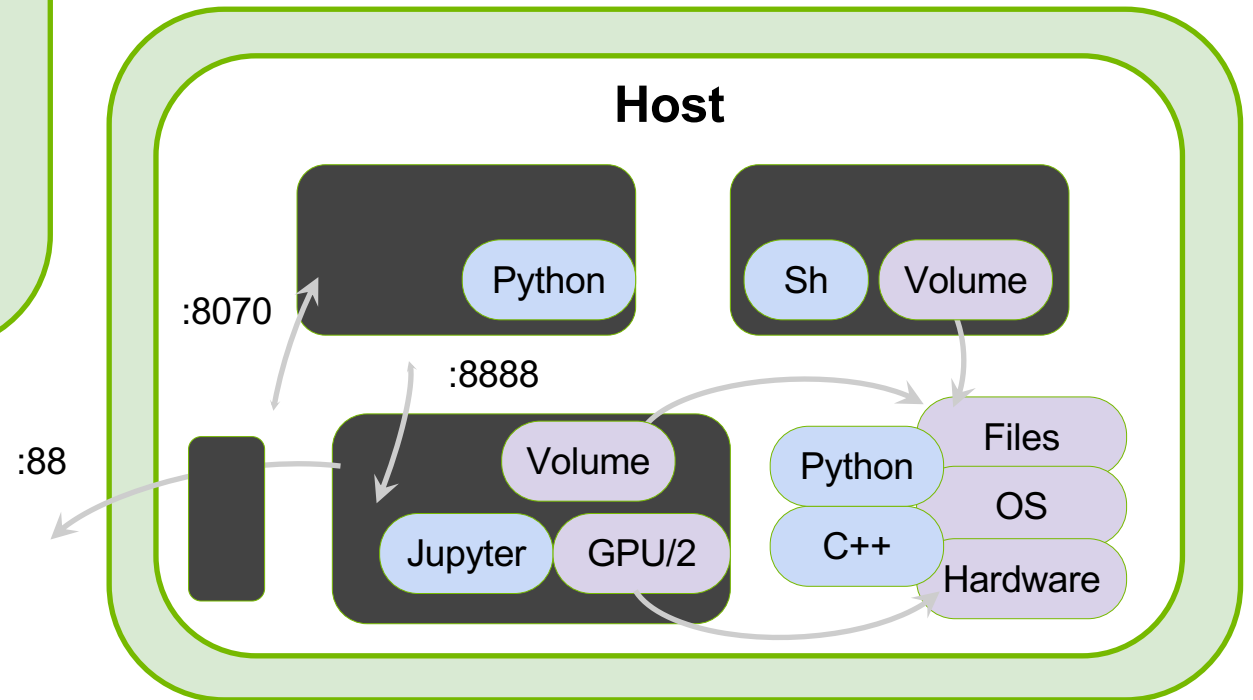
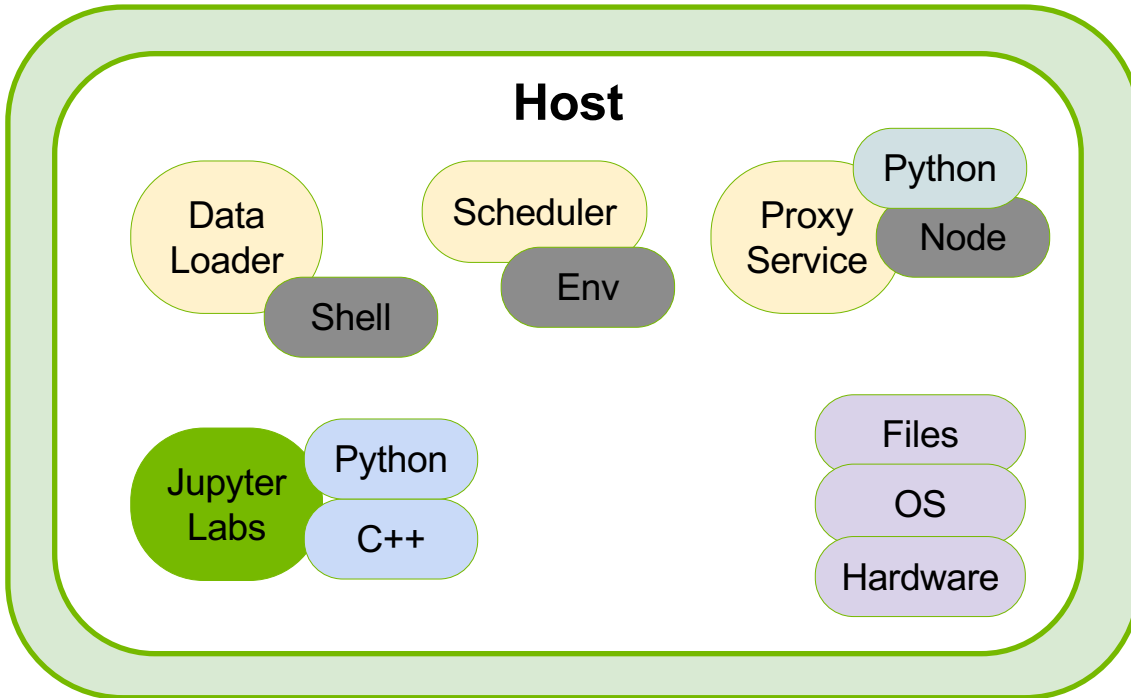
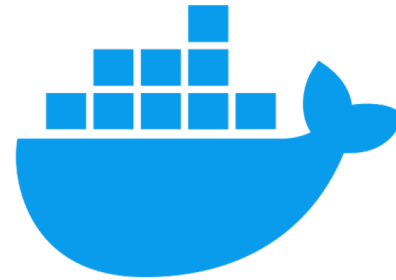


DLI Jupyter Labs Interface



Containerization with Docker

Compartmentalizing Functionality Into Microservices



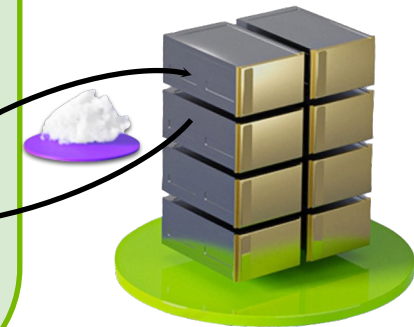
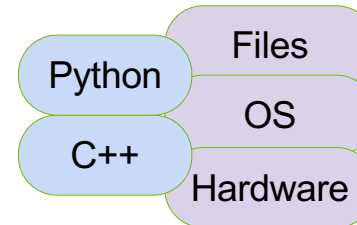
Microservices Workflow



Microservices Workflow

1. Allocate Resources
2. Define Services
3. Construct Containers
4. Start Processes

Remote Host

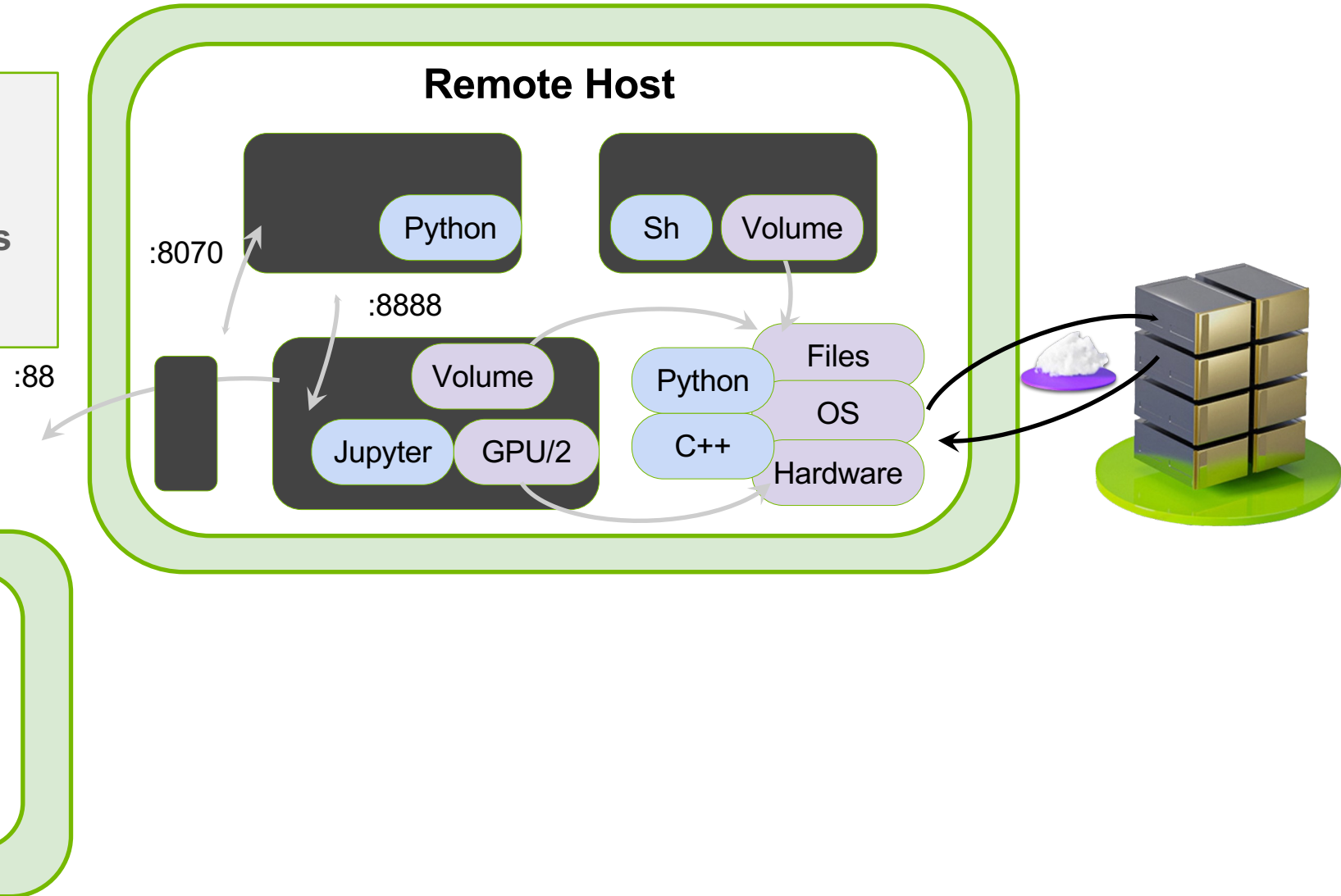


Web Browser

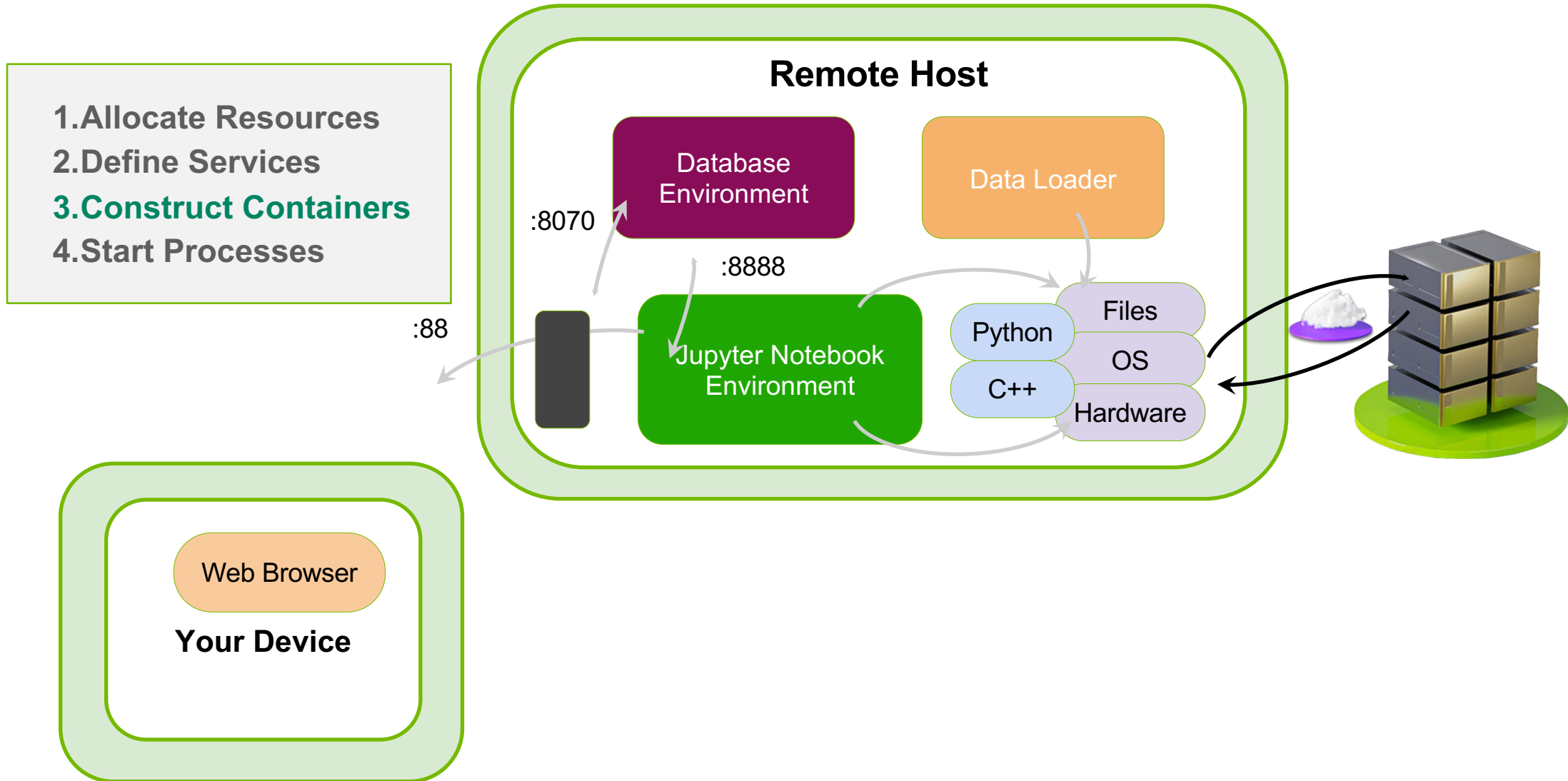
Your Device

Microservices Workflow

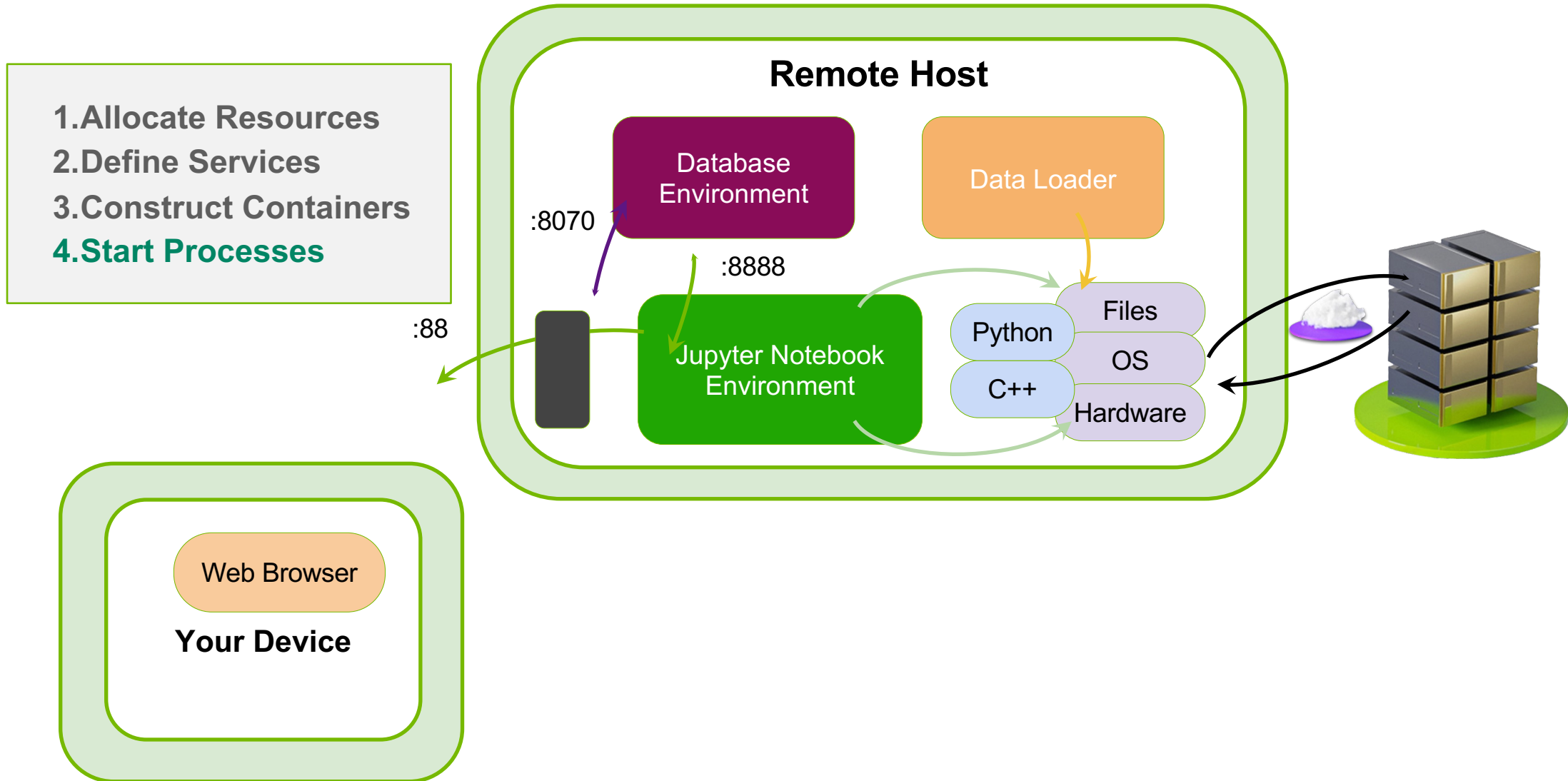
1. Allocate Resources
2. Define Services
3. Construct Containers
4. Start Processes



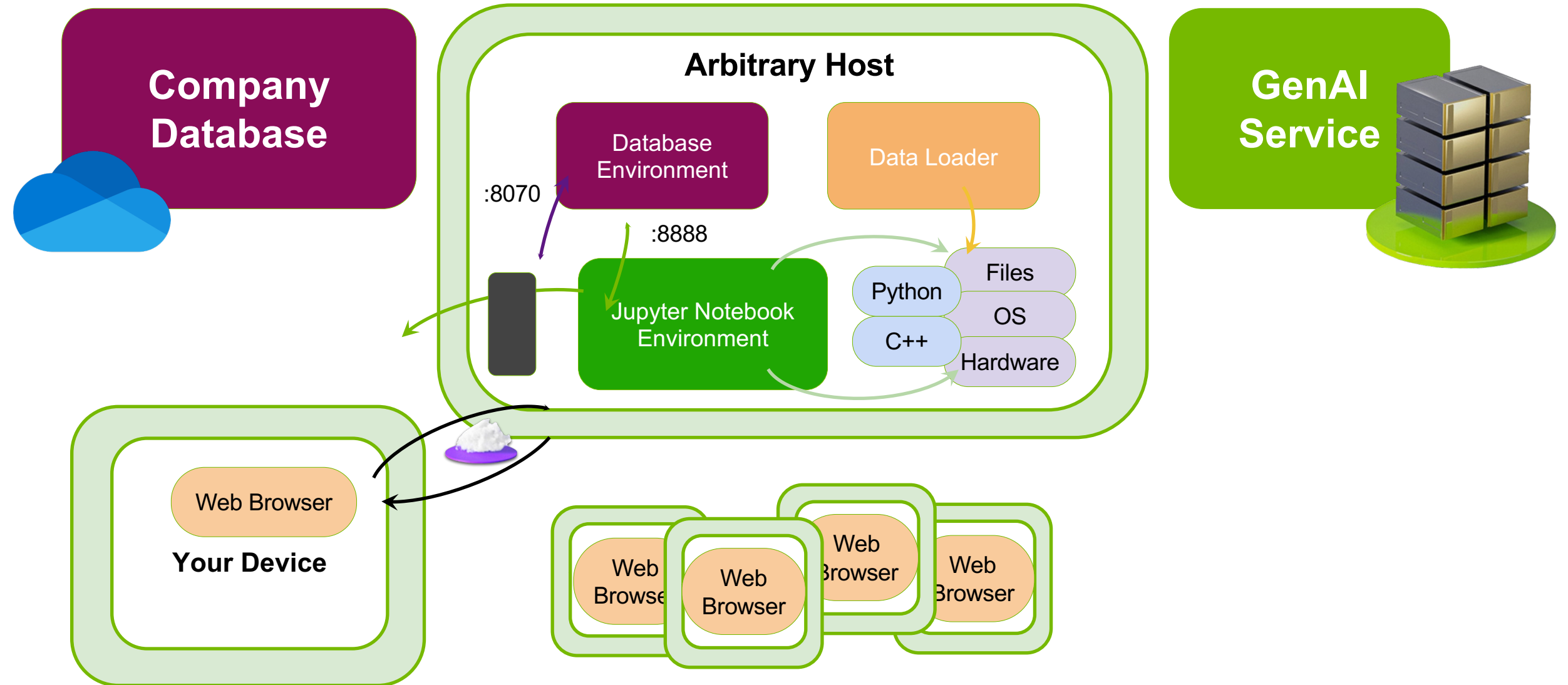
Microservices Workflow



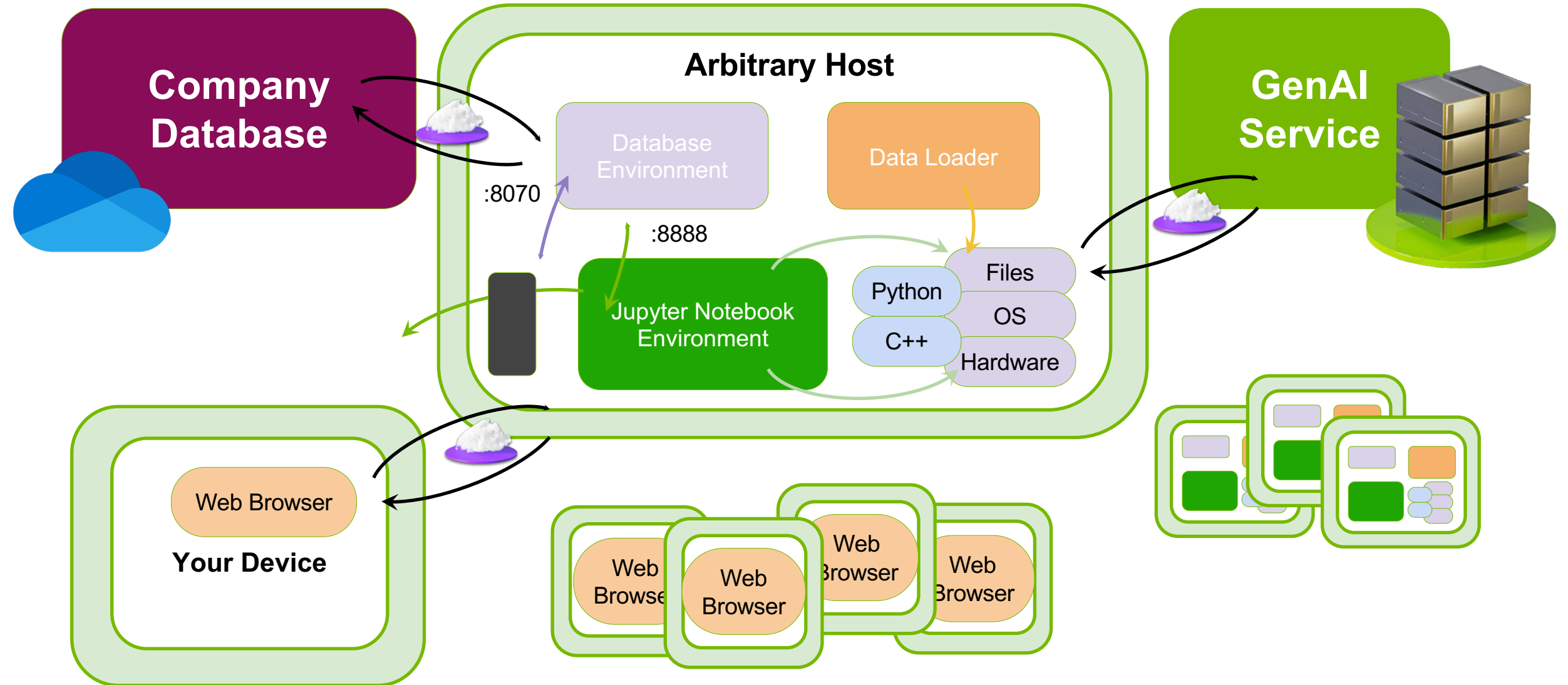
Microservices Workflow



Scaling Containerized Applications

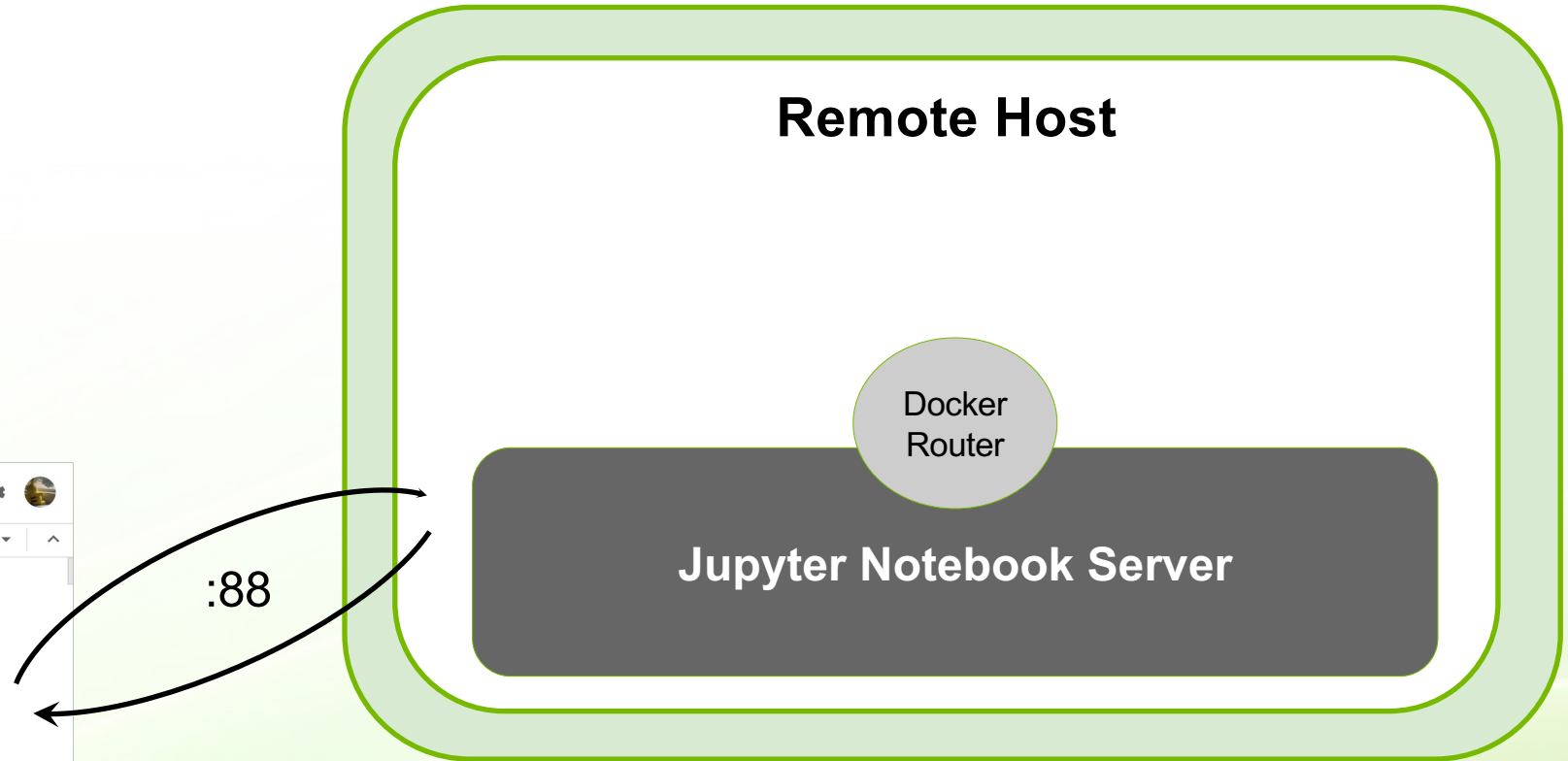
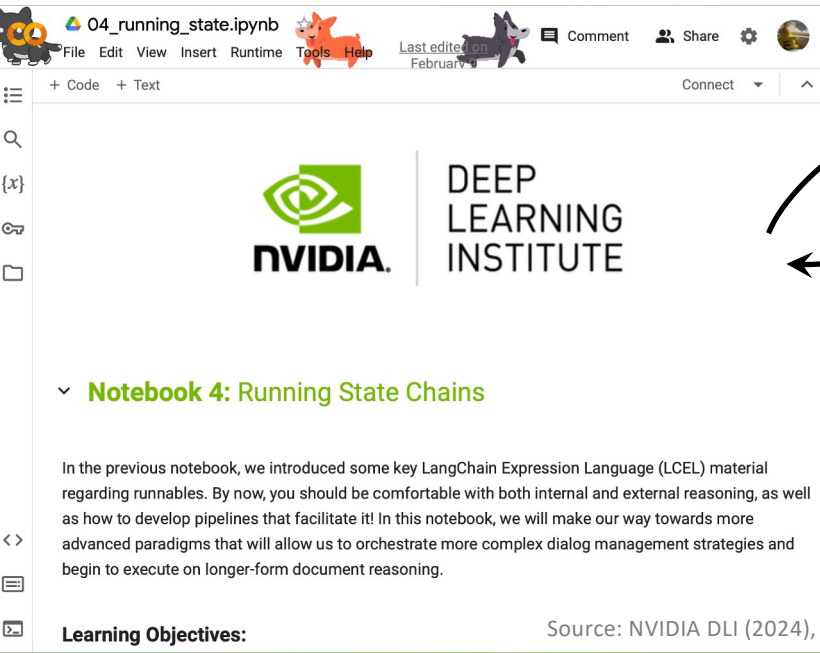


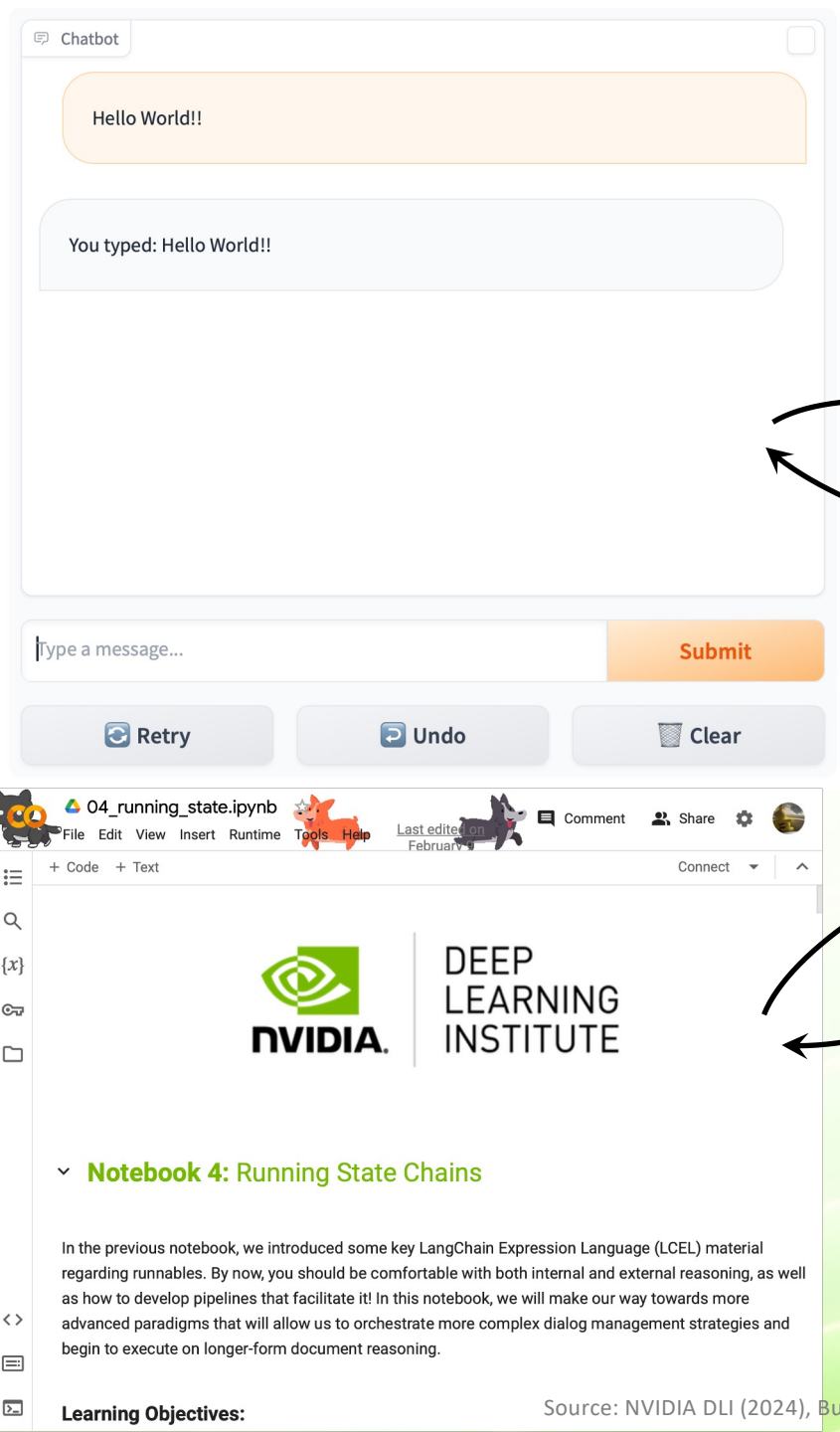
Scaling Containerized Applications



Our Environment

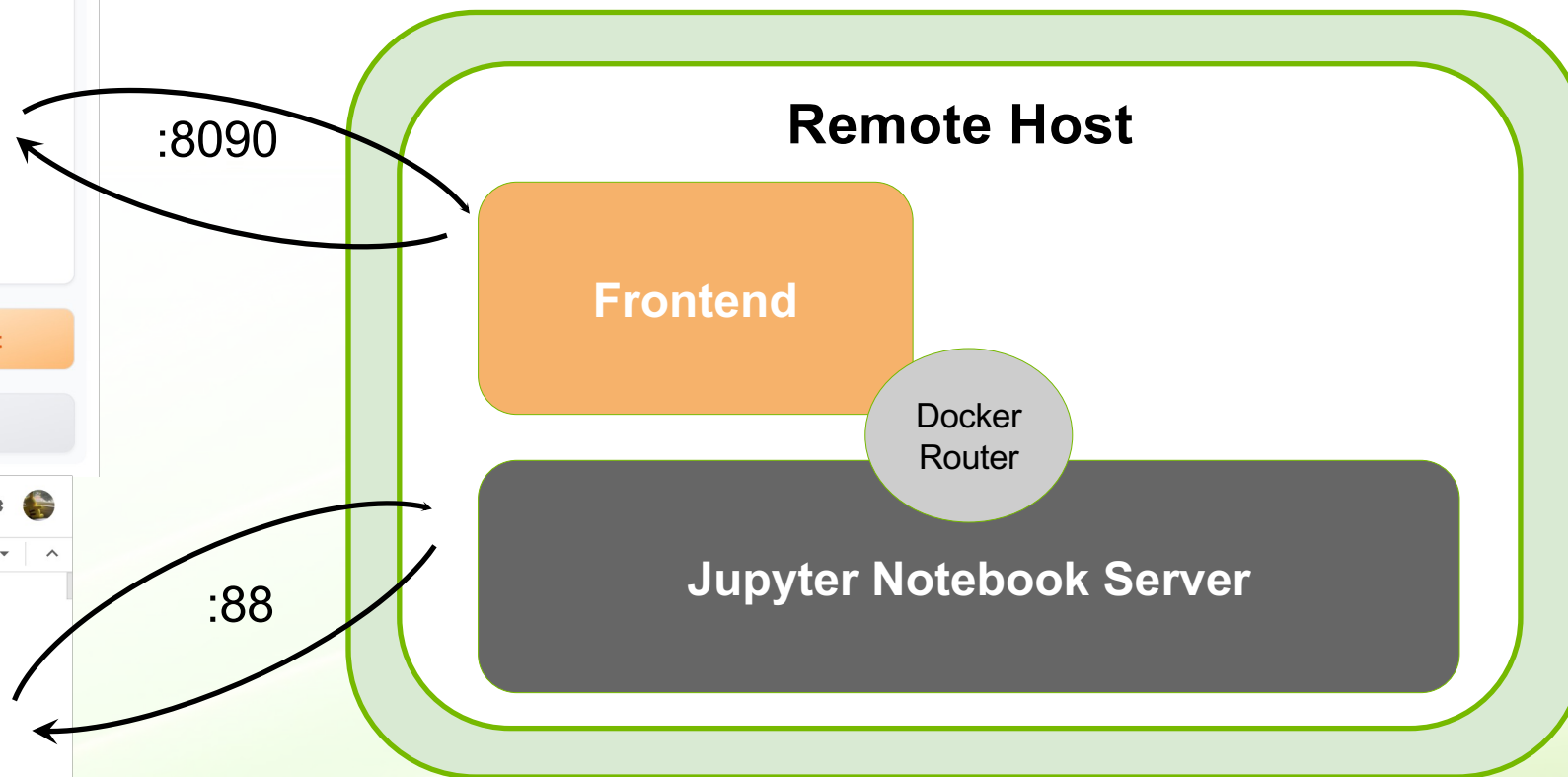
Jupyter Notebooks



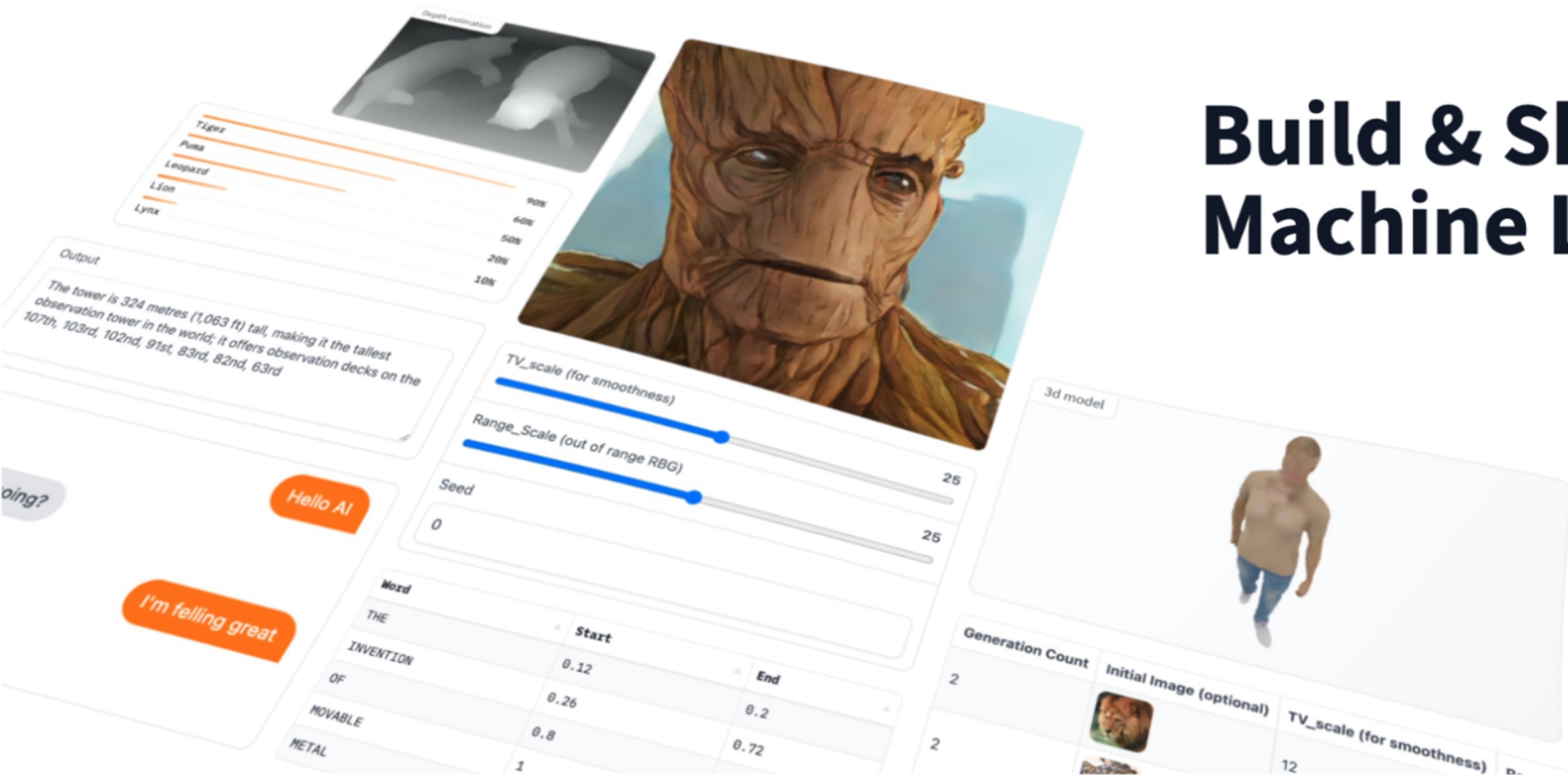


Our Environment

Jupyter Notebooks + Frontend



Gradio

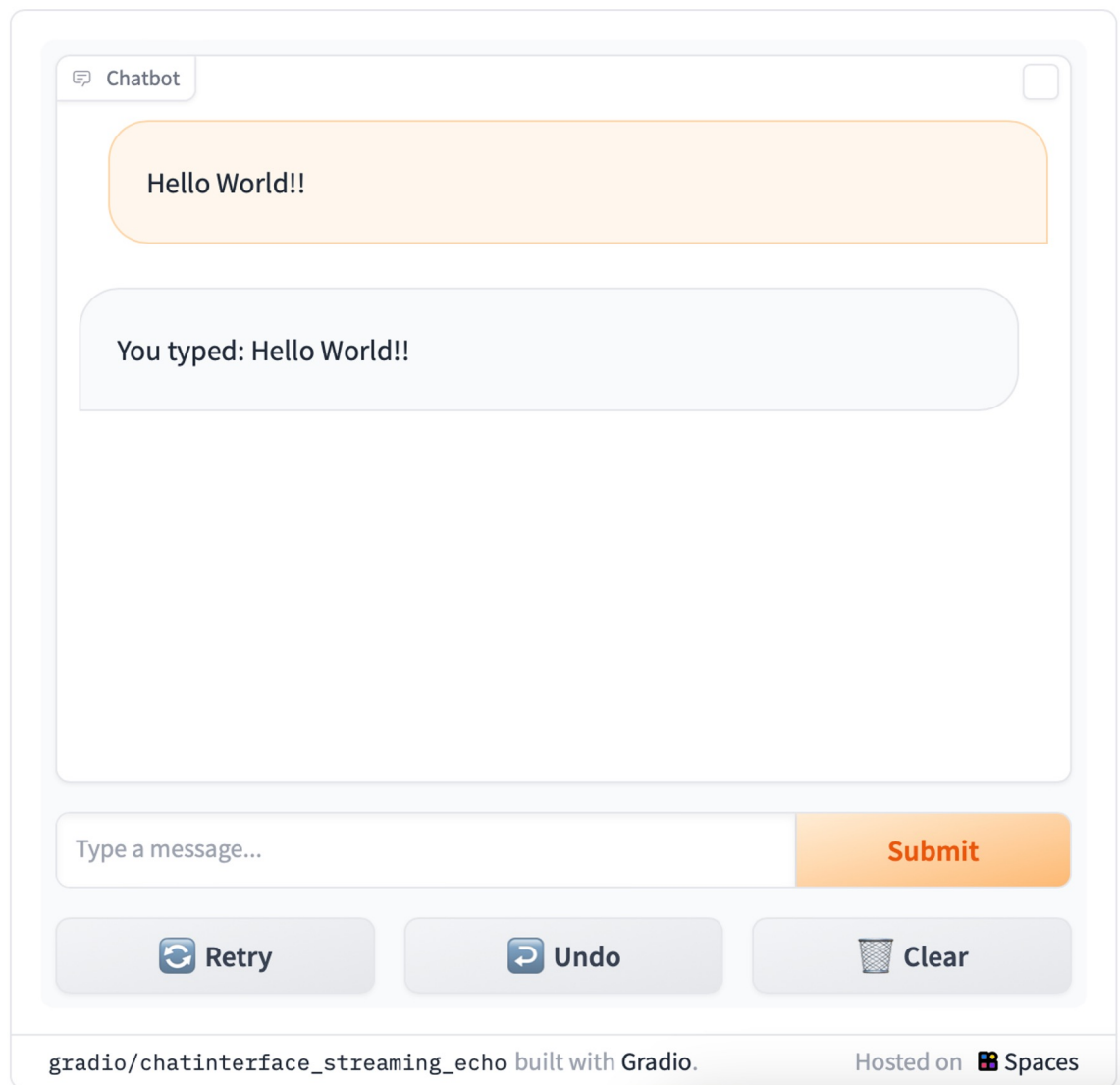
[⚡ Quickstart](#)[📝 Docs](#)[💡 Guides](#)[📊 Playground ^{NEW}](#)[👋 Community ▾](#)

Build & Share Delightful Machine Learning Apps

Gradio is the fastest way to demo your machine learning model with a friendly web interface so that anyone can use it, anywhere!

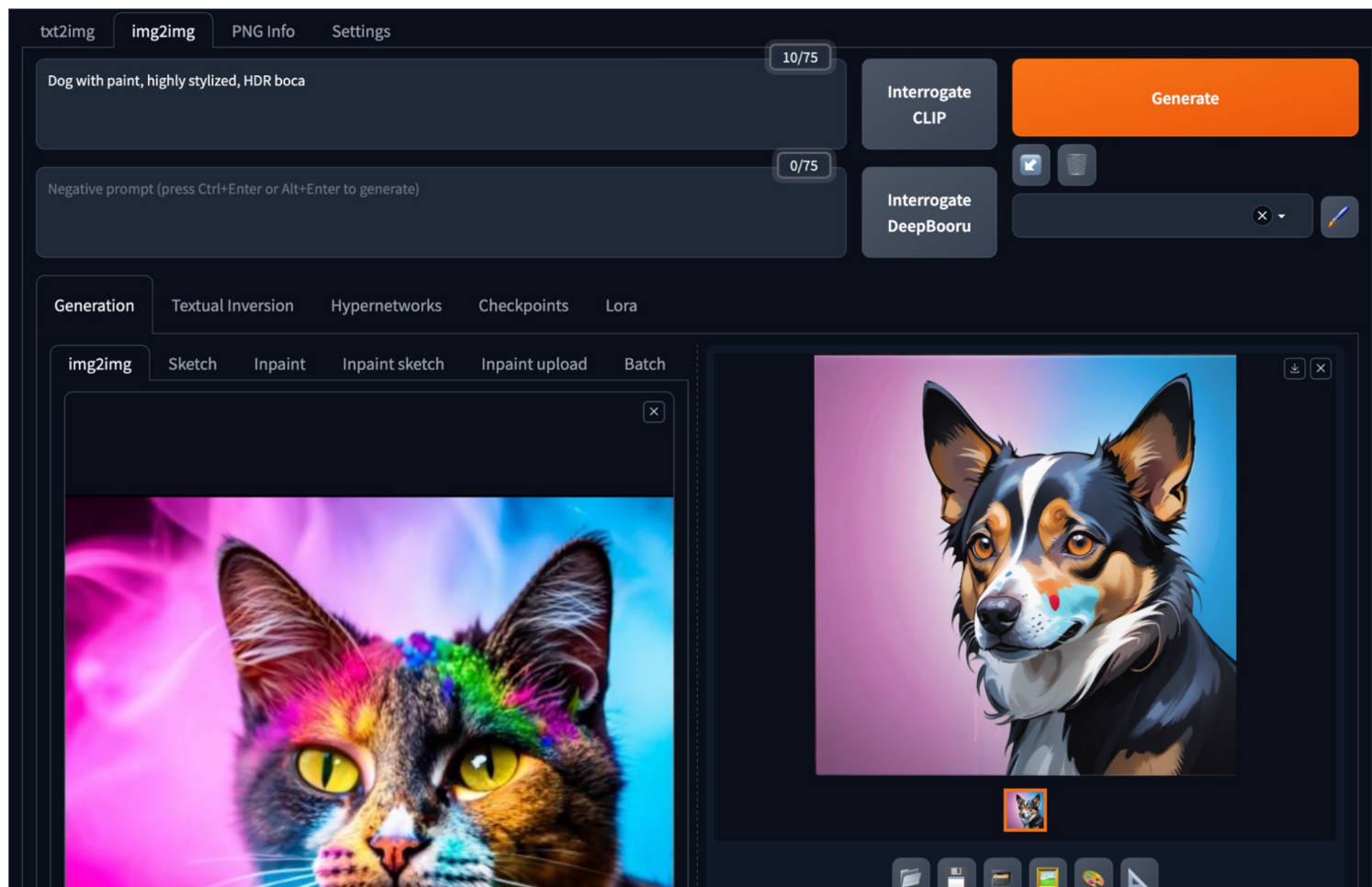
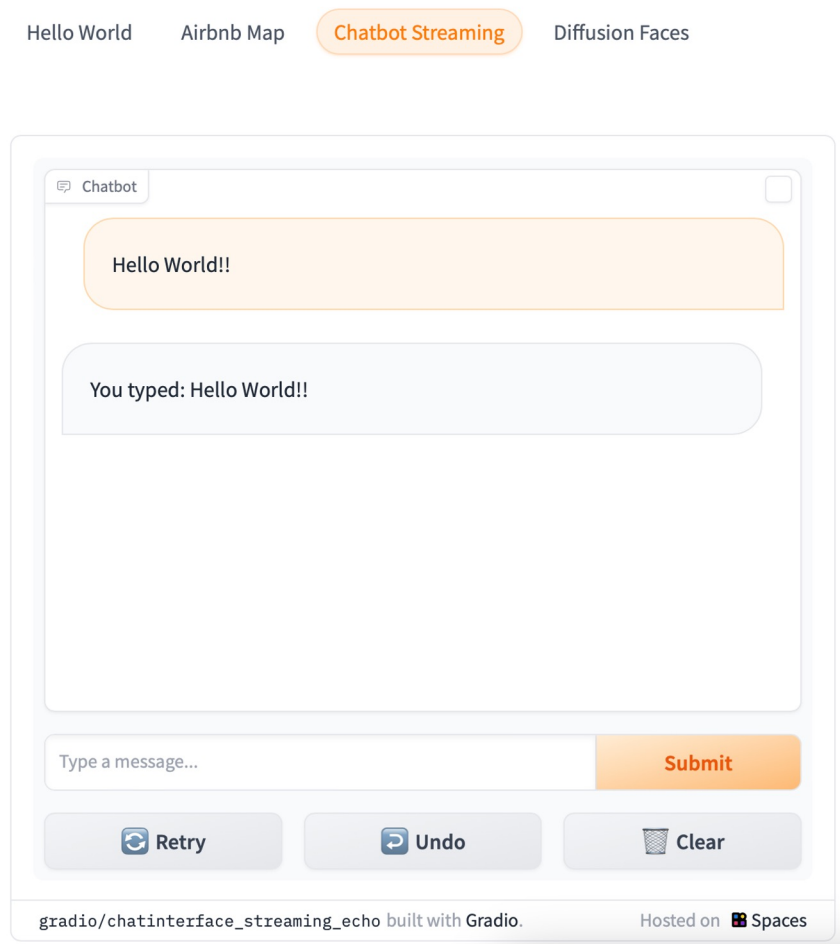
[Get Started](#)[🌟 Star](#)[23997](#)

Simple Gradio ChatInterface

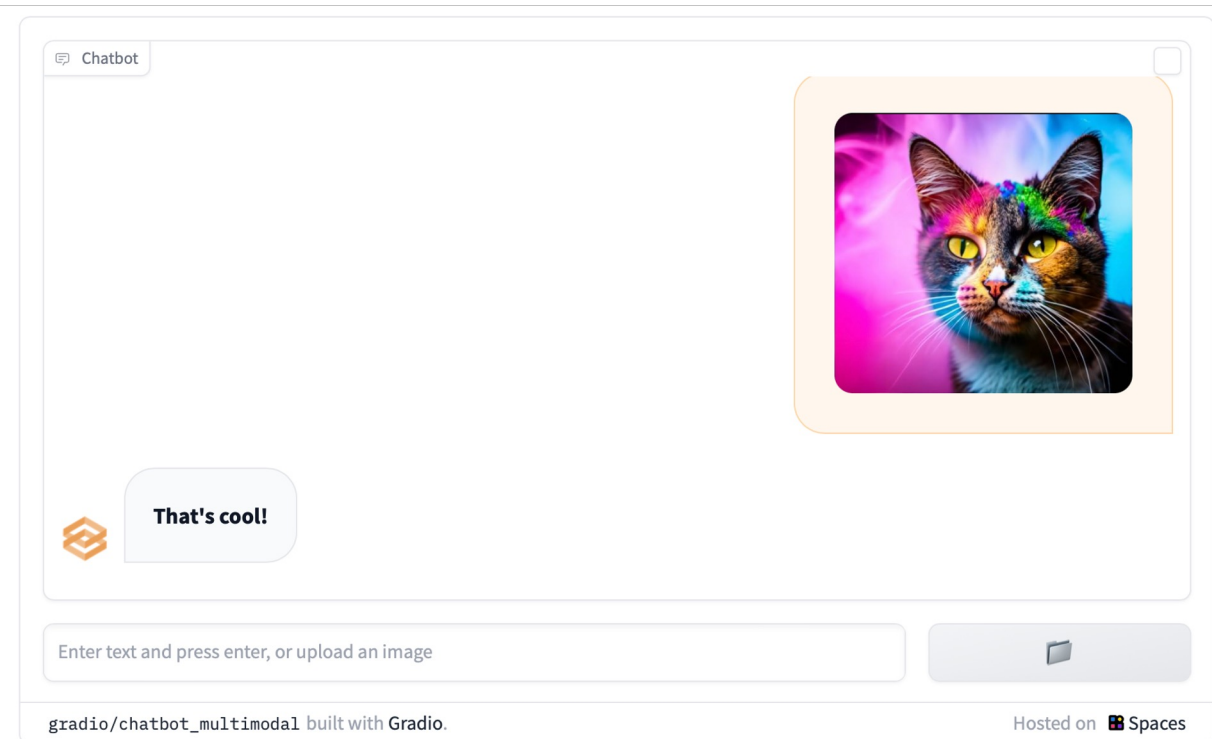


```
1  import time
2  import gradio as gr
3
4  def slow_echo(message, history):
5      for i in range(len(message)):
6          time.sleep(0.05)
7          yield "You typed: " + message[: i+1]
8
9  demo = gr.ChatInterface(slow_echo).queue()
10
11  if __name__ == "__main__":
12      demo.launch()
13
```


Gradio in HuggingFace Spaces



Custom Gradio Block Interface

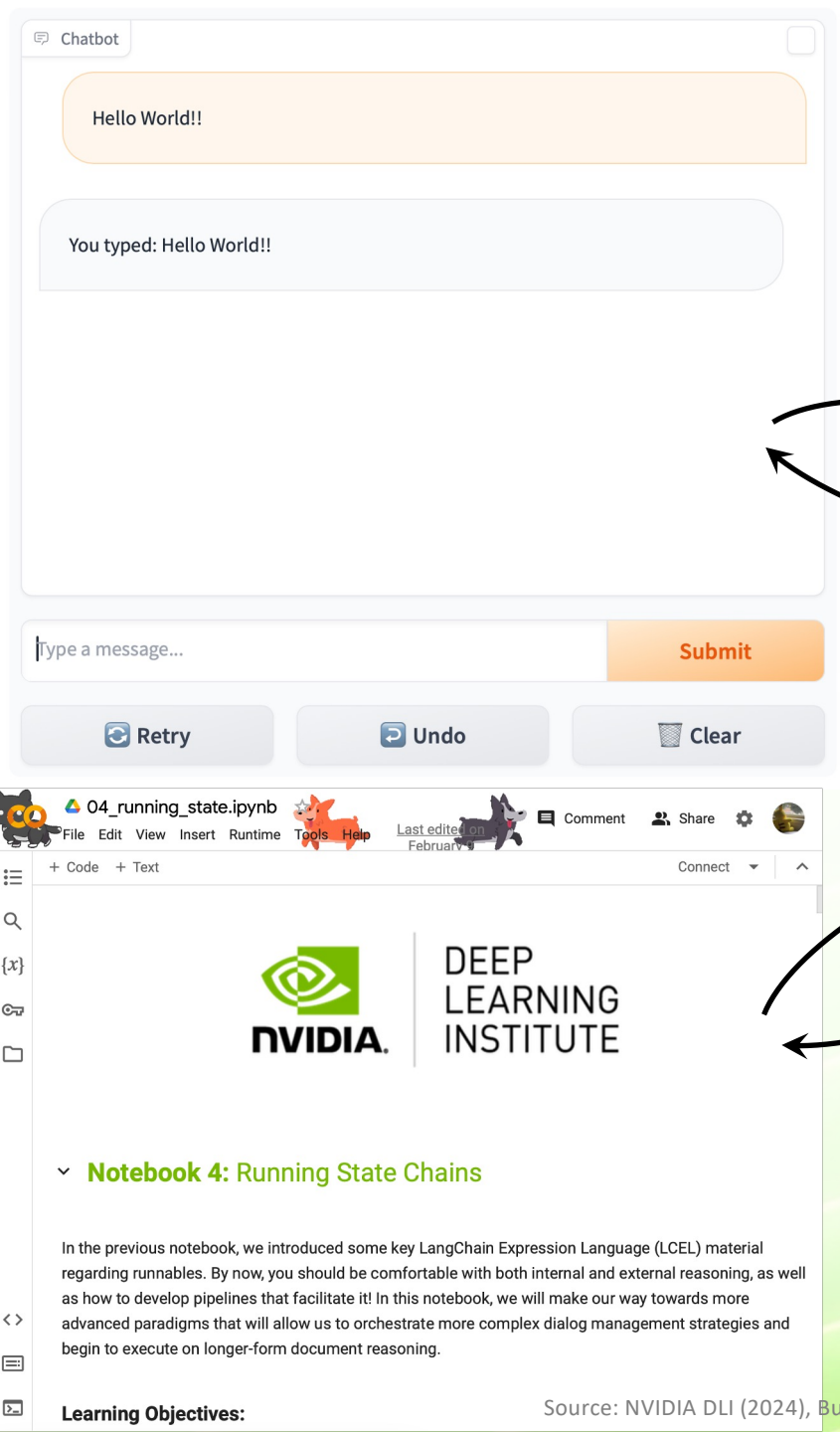


```
with gr.Blocks() as demo:
    chatbot = gr.Chatbot(
        [],
        elem_id="chatbot",
        bubble_full_width=False,
        avatar_images=(None, (os.path.join(os.path.dirname(__file__), "avatar.png"))),
    )

    with gr.Row():
        txt = gr.Textbox(
            scale=4,
            show_label=False,
            placeholder="Enter text and press enter, or upload an image",
            container=False,
        )
        btn = gr.UploadButton("📁", file_types=["image", "video", "audio"])

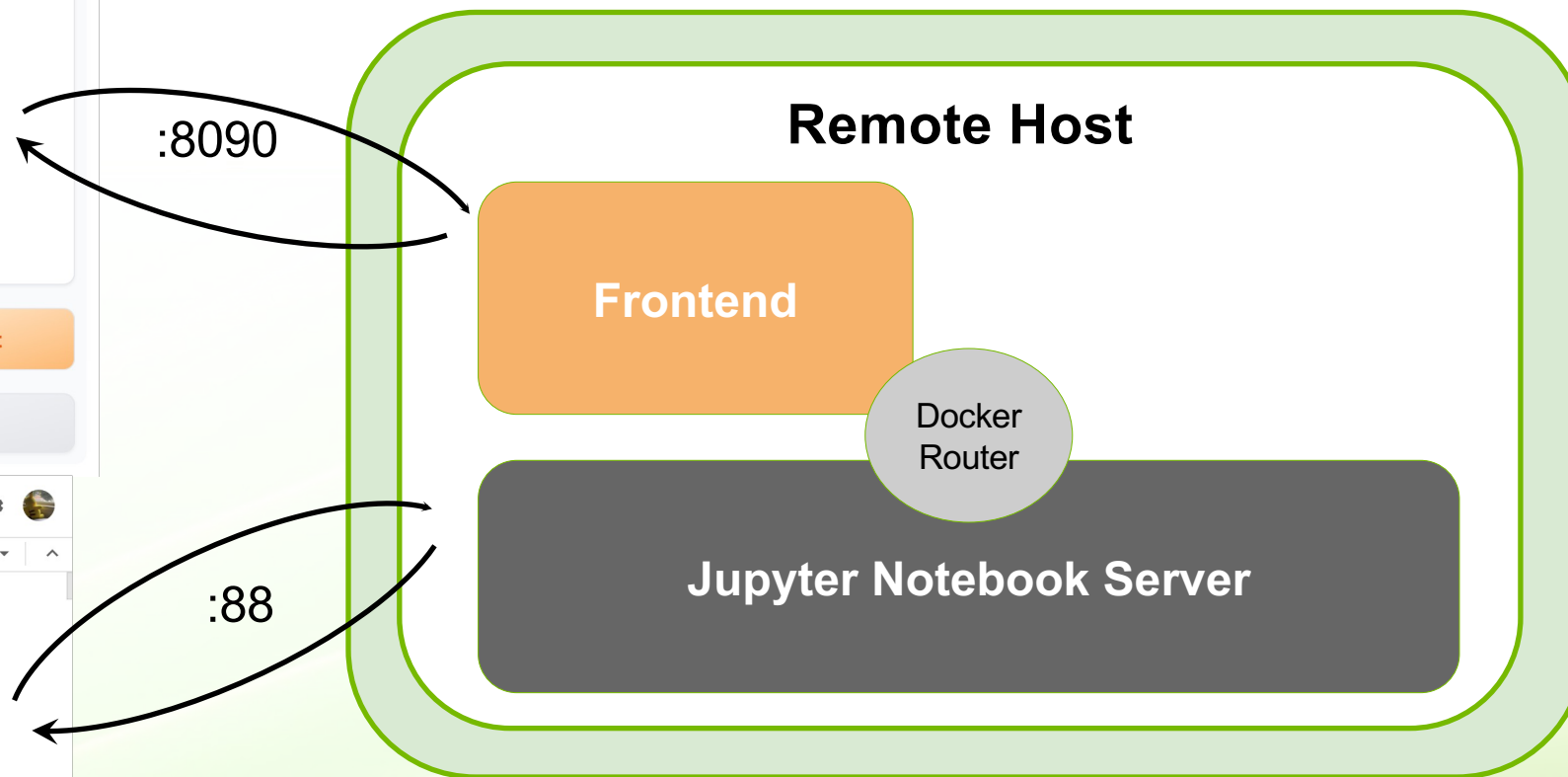
    txt_msg = txt.submit(add_text, [chatbot, txt], [chatbot, txt], queue=False).then(
        bot, chatbot, chatbot, api_name="bot_response"
    )
    txt_msg.then(lambda: gr.Textbox(interactive=True), None, [txt], queue=False)
    file_msg = btn.upload(add_file, [chatbot, btn], [chatbot], queue=False).then(
        bot, chatbot, chatbot
    )

demo.queue()
demo.launch()
```



Our Environment

Jupyter Notebooks + Frontend



Building RAG Agents with LLMs

Part 1: Your Course Environment

Part 2: LLM Services

Part 3: Intro to LangChain

Part 4: Running State Chain

Part 5: Working with Documents

Part 6: Embedding Model for Retrieval

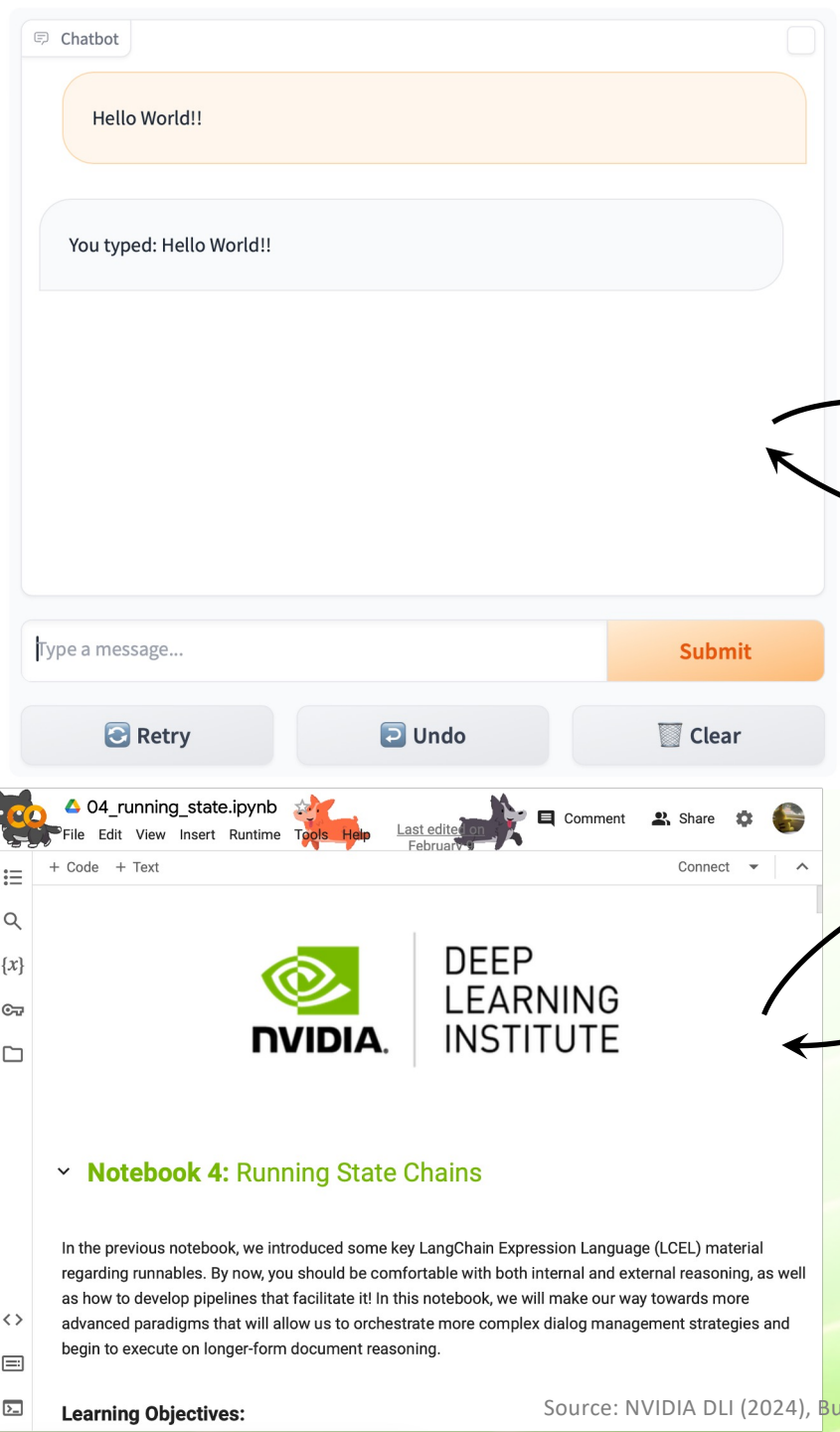
Part 7: Document Retrieval with Vector Databases

Part 8: RAG Evaluation



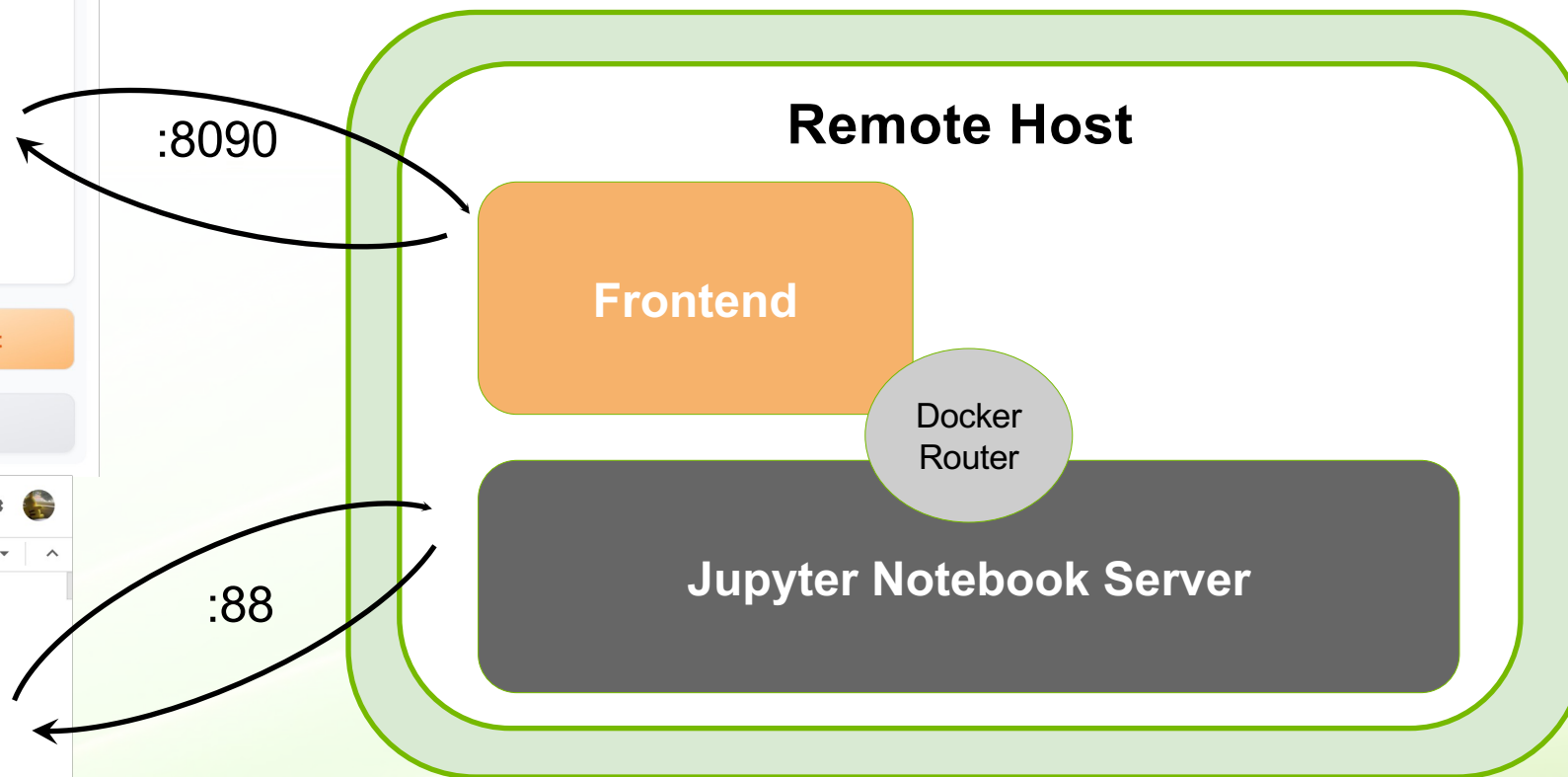
Building RAG Agents with LLMs

Part 2: LLM Services



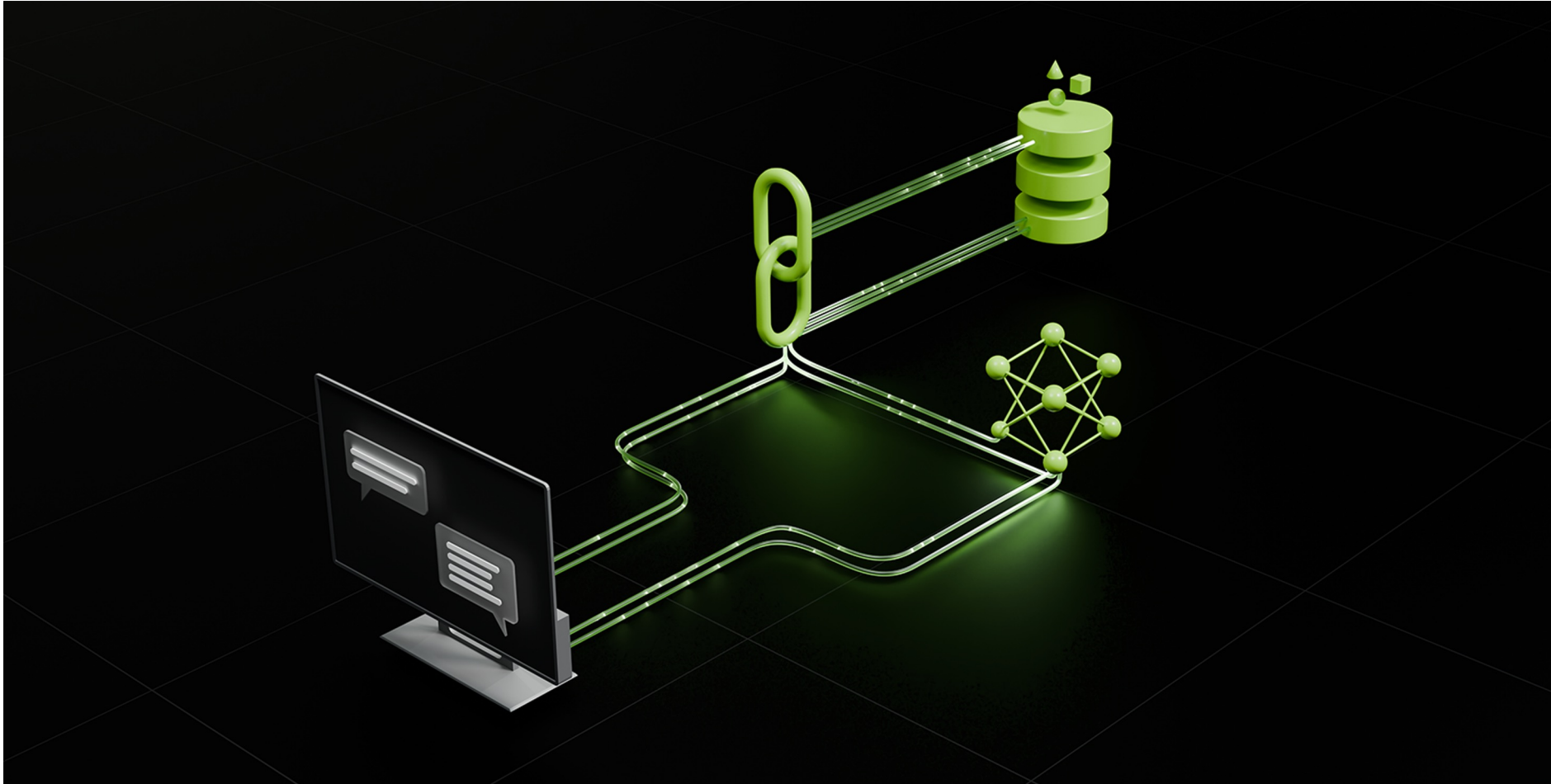
Our Environment

Jupyter Notebooks + Frontend



Dialog/Retrieval Agents

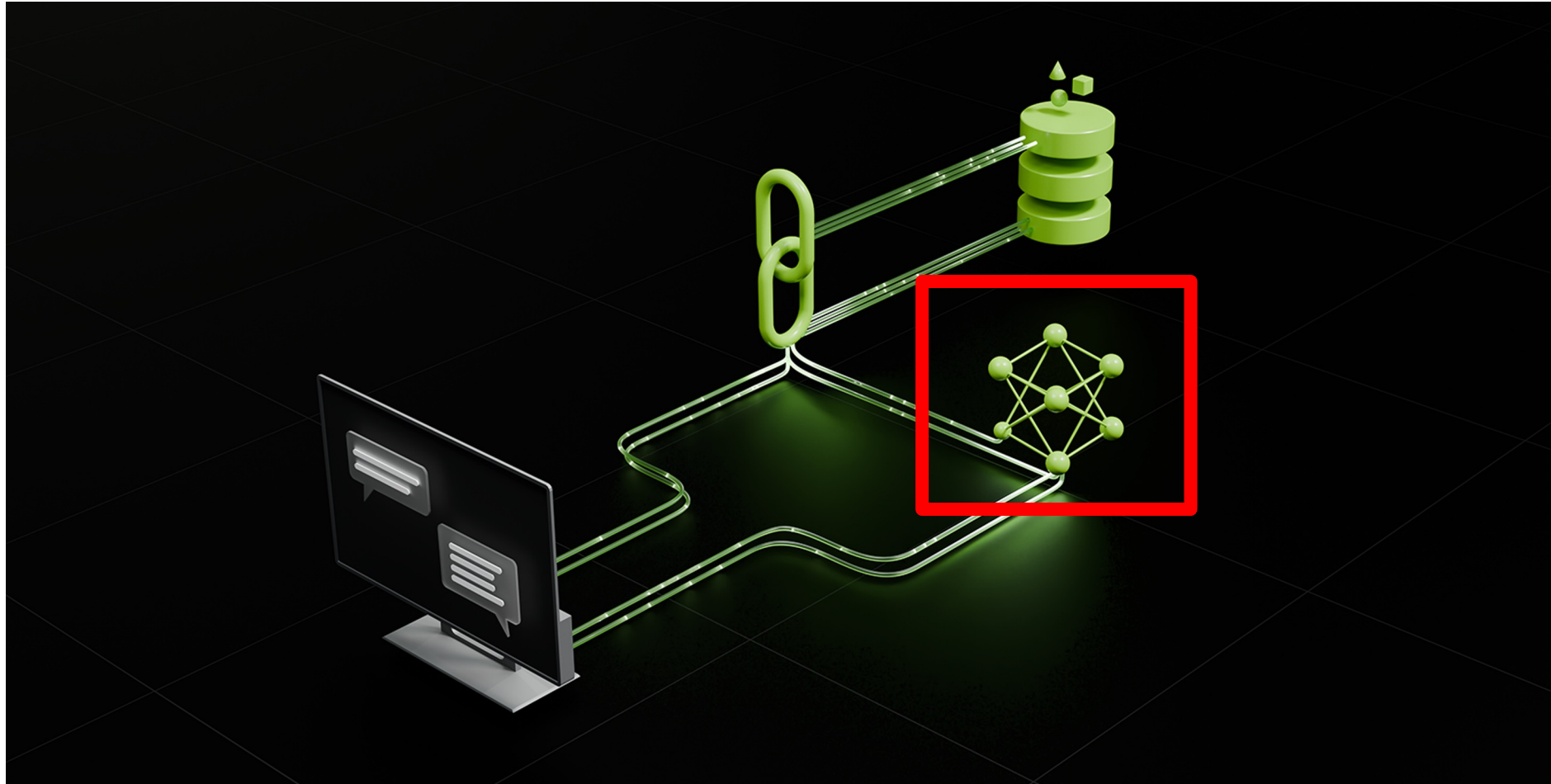
LLMs with Context and Control



Execute with dialog management and information retrieval in production

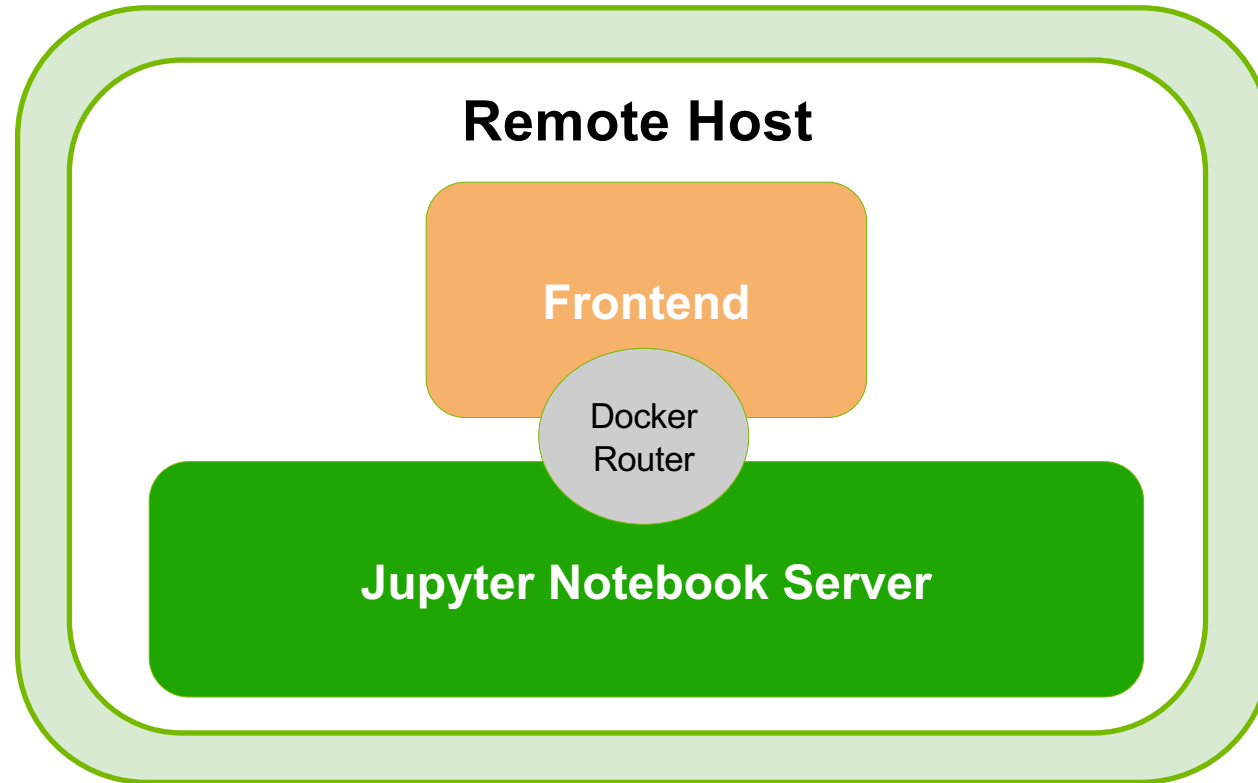
Dialog/Retrieval Agents

LLMs with Context and Control

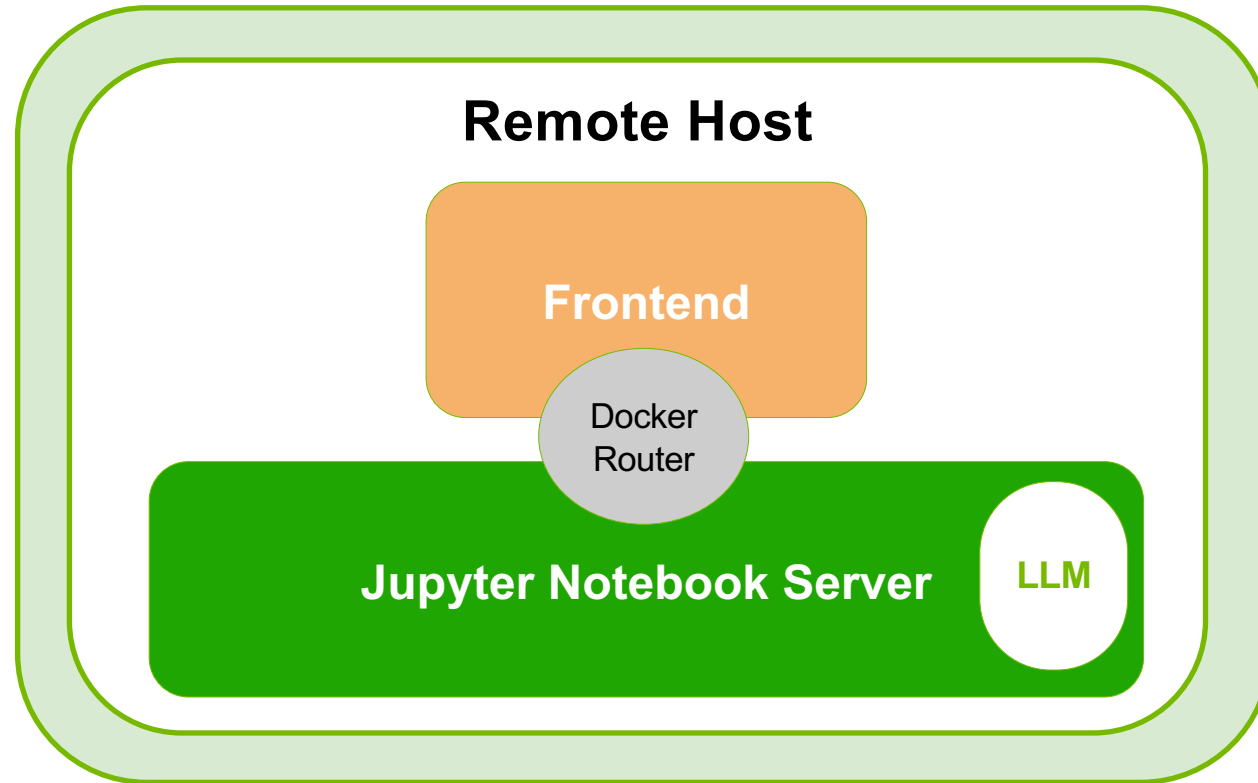


Execute with dialog management and information retrieval in production

Standalone Environment LLM



Standalone Environment LLM



Standalone Environment LLM

Jupyter Notebook

H200

A100

Jupyter Notebook

A10

4070

Jupyter Notebook

VRAM-bound

CPU-only



```
<s>[INST]<<SYS>>
{{system_message}}
<</SYS>>
```

```
{{instruction}} [/INST]
{{primer}}
```

```
<s>[INST]<<SYS>>
You are a code generator.
Please provide Python code
per the instruction.
<</SYS>>
```

```
Write a Fibonacci method [/INST]
```python
Implementation of Fibonacci w/
```

# Standalone Environment LLM

Jupyter Notebook

H200

A100

Jupyter Notebook

A10

4070

Jupyter Notebook

VRAM-bound

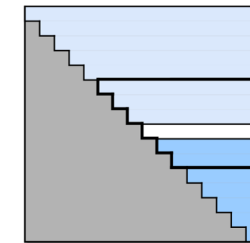
CPU-only



64	134	217
76	119	21
3	81	99

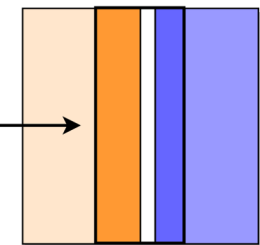
INT8

Inverse Layer Hessian  
(Cholesky Form)

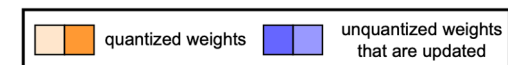


computed initially

Weight Matrix / Block



block i quantized recursively  
column-by-column



# Standalone Environment LLM

Jupyter Notebook

H200

A100

Jupyter Notebook

A10

4070

Jupyter Notebook

VRAM-bound

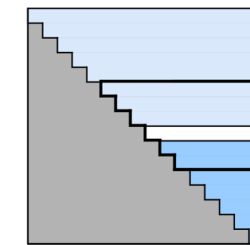
CPU-only



64	134	217
76	119	21
3	81	99

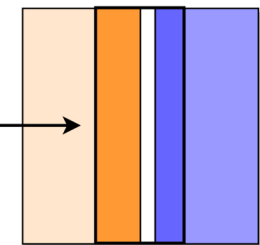
INT8

Inverse Layer Hessian  
(Cholesky Form)

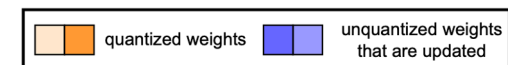


computed initially

Weight Matrix / Block



block  $i$  quantized recursively  
column-by-column





# Remote LLM Access

## Remote Host

Frontend

Jupyter Notebook

VRAM-bound

CPU-only

```
<s>[INST]<<SYS>>
{{system_message}}
<</SYS>>
```

```
{{instruction}} [/INST]
{{primer}}
```

```
<s>[INST]<<SYS>>
You are a code generator.
Please provide Python code
per the instruction.
<</SYS>>
```

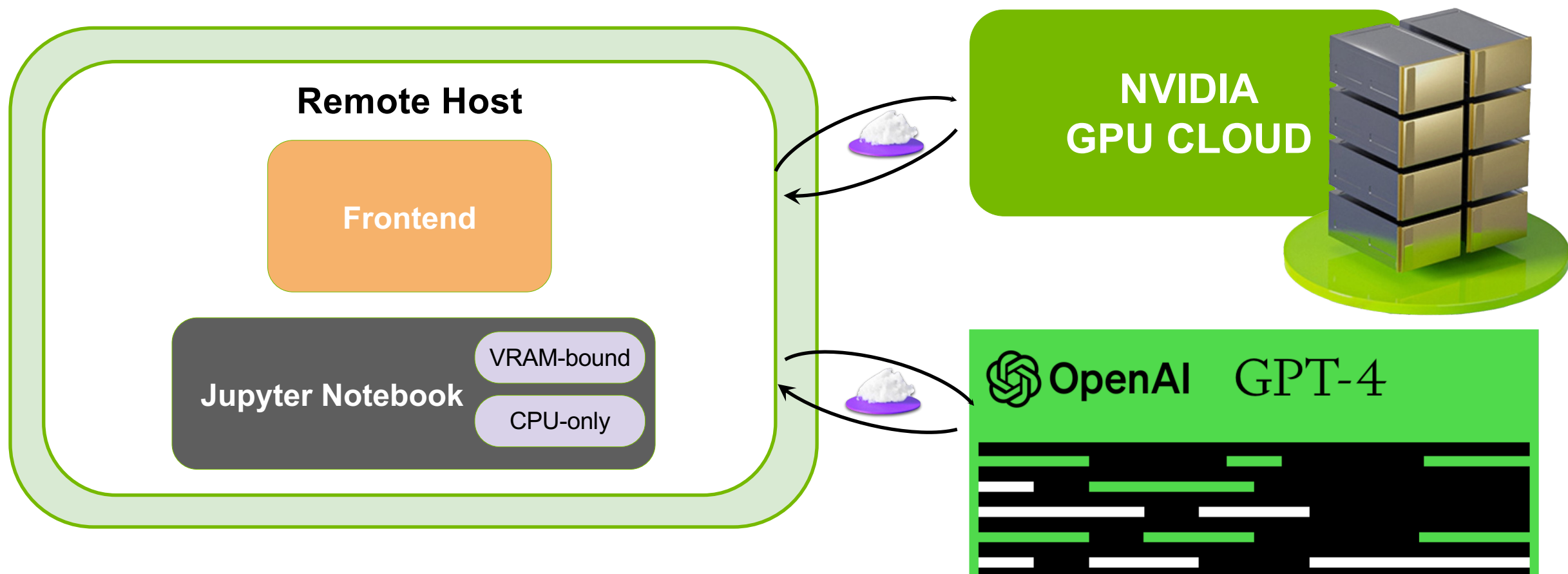
```
Write a Fibonacci method [/INST]
```python
## Implementation of Fibonacci w/
```

Llama

LLM



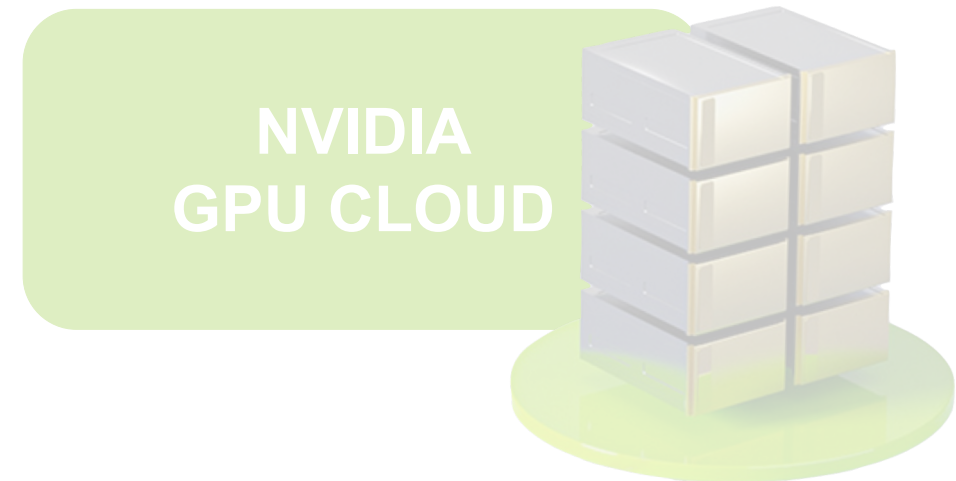
Large Model Hosting Platforms



Large Model Hosting Platforms

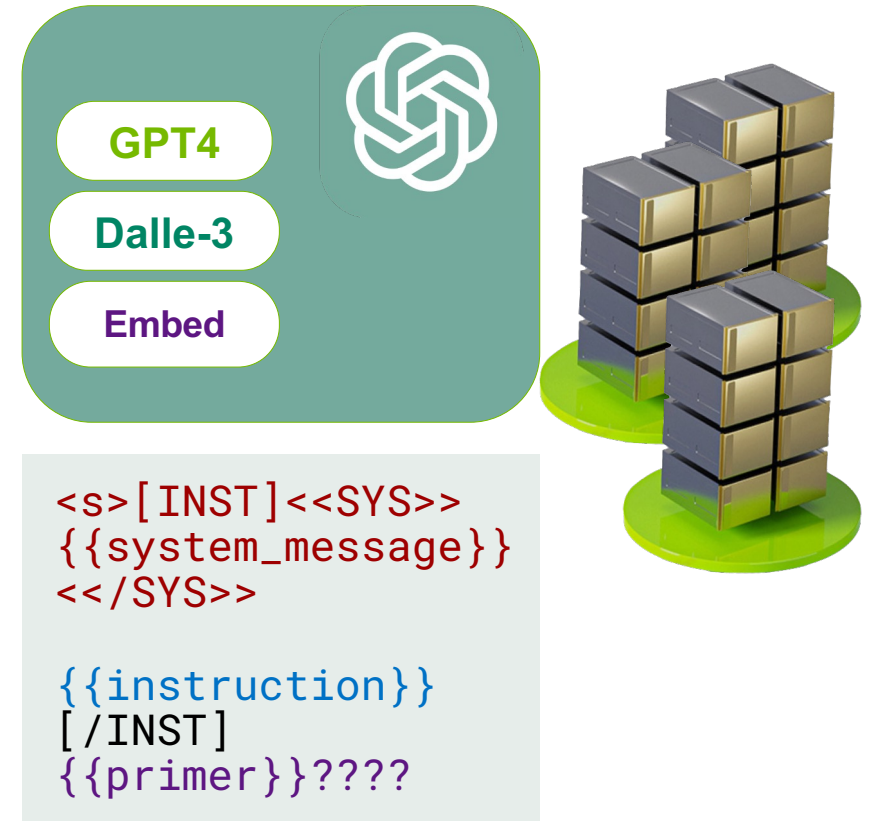


- 1. GPT4/3.5
- 2. Integrated Tools/Multimodal Support
- 3. Models/Query Router Internal
- 4. Path to Local Deployment Less Clear



- 1. Publicly Available Models
- 2. Models/Query Router Accessible
- 3. Path to Local Deployment/Self-Hosting
- 4. Pieces to Compose Complex Systems

Query Router Access



Query Router Access

```
{  
  "messages": [{  
    "content": "...",  
    "role": "system"  
  }, {  
    "content": "...",  
    "role": "user"  
  }],  
  "model": "gpt-4",  
  "temperature": 0.2,  
  "top_p": 0.7,  
  "max_tokens": 1024,  
  "stream": True  
}
```

/chat/completions

Facilitate
Monitor
Optimize
Load Balance

GPT4

Dalle-3

Embed

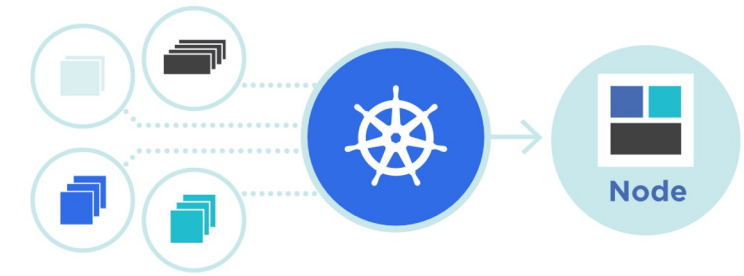


```
<s>[INST]<SYS>  
{{system_message}}  
<</SYS>>
```

```
{{instruction}}  
[/INST]  
{{primer}}????
```



Query Router Access



```
{  
  "messages": [{  
    "content": "...",  
    "role": "system"  
  }, {  
    "content": "...",  
    "role": "user"  
  }],  
  "model": "gpt-4",  
  "temperature": 0.2,  
  "top_p": 0.7,  
  "max_tokens": 1024,  
  "stream": True  
}
```

/chat/completions

Facilitate
Monitor
Optimize
Load Balance

GPT4

Dalle-3

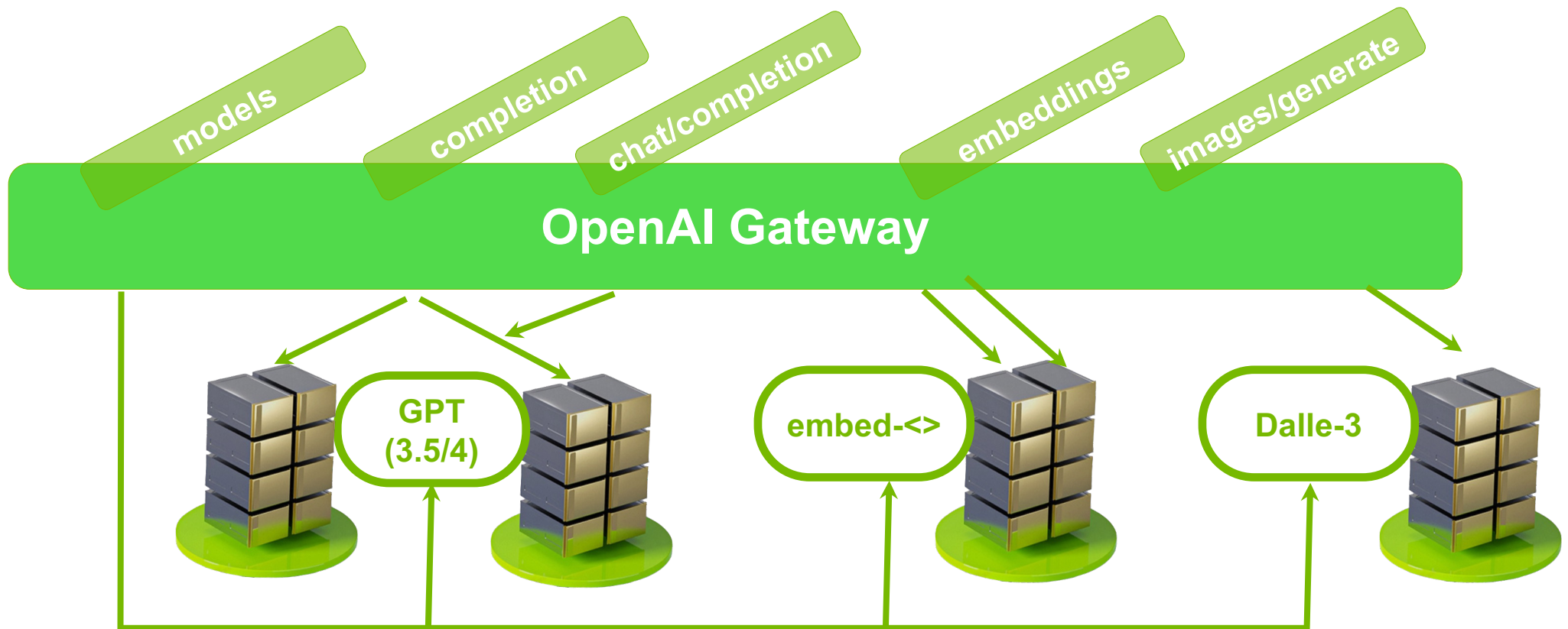
Embed

```
<s>[INST]<<SYS>>  
{{system_message}}  
<</SYS>>
```

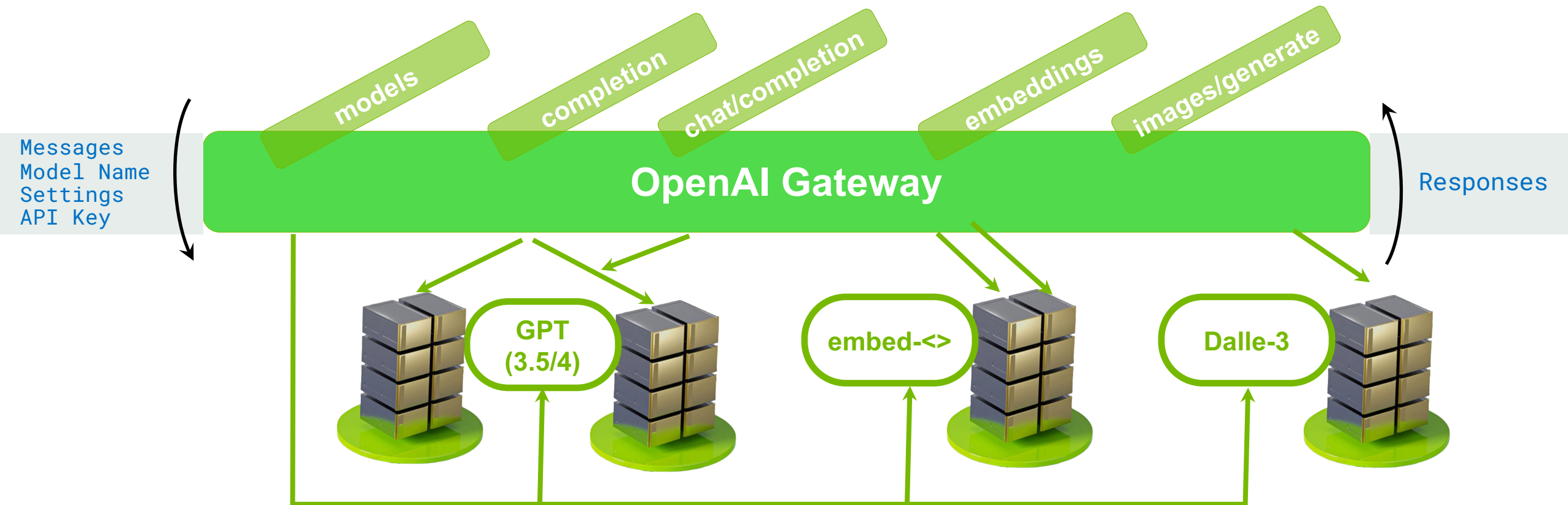
```
{{instruction}}  
[/INST]  
{{primer}}????
```



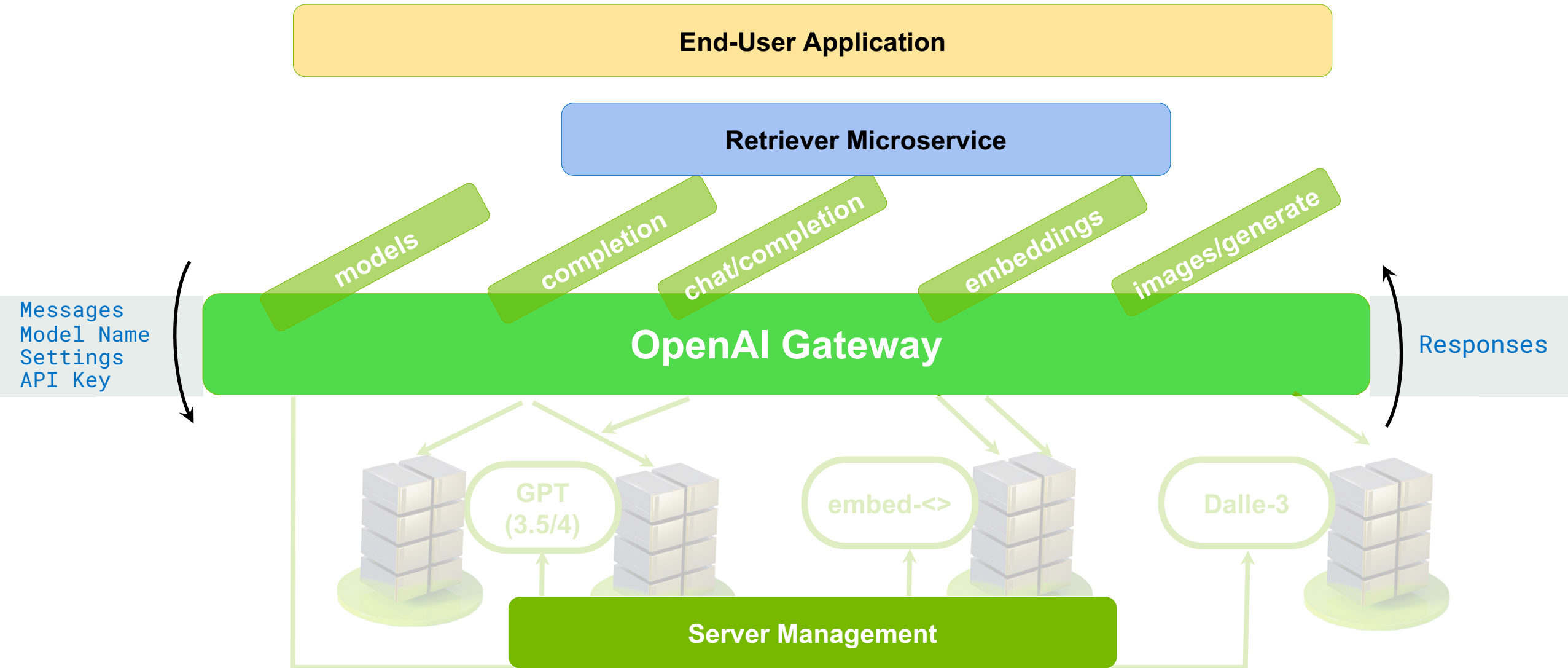
Large Model Hosting Platforms - OpenAI



Large Model Hosting Platforms - OpenAI



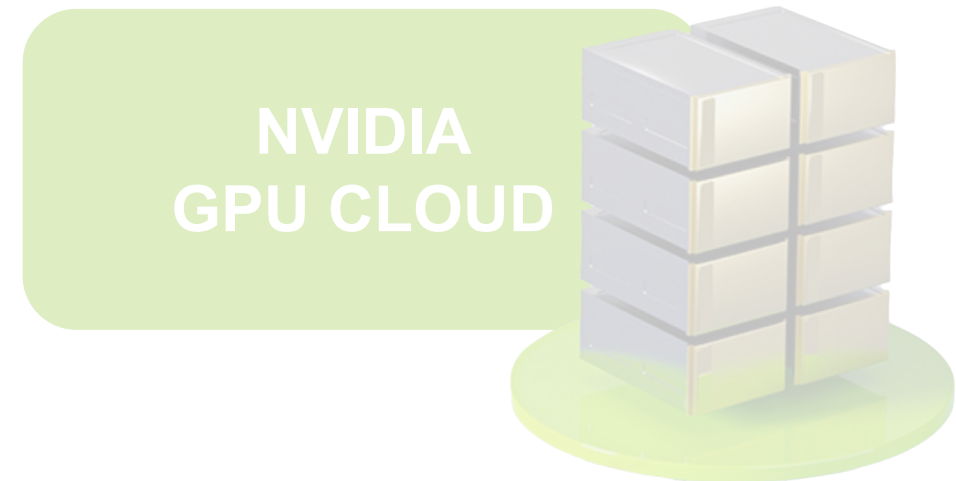
Large Model Hosting Platforms - OpenAI



Large Model Hosting Platforms



- 1. GPT4/3.5
- 2. Integrated Tools/Multimodal Support
- 3. Models/Query Router Internal
- 4. Path to Local Deployment Less Clear

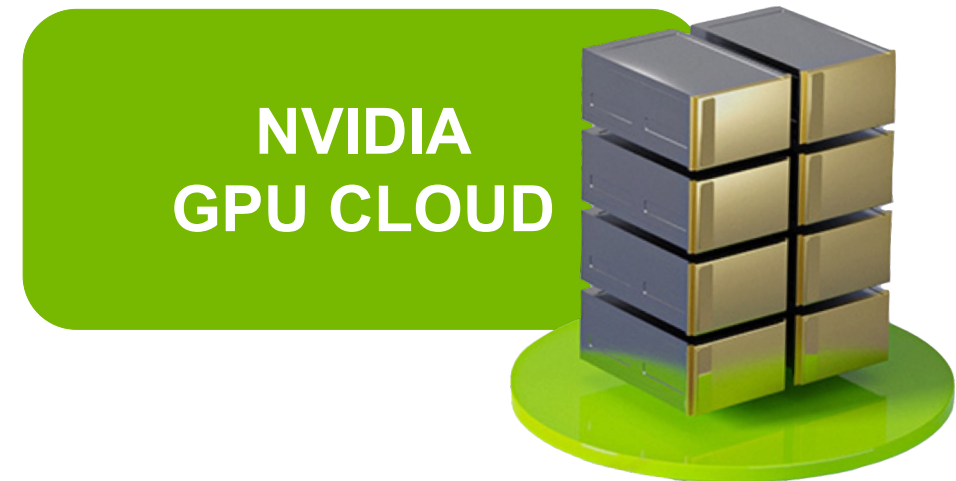


- 1. Publicly Available Models
- 2. Models/Query Router Accessible
- 3. Path to Local Deployment/Self-Hosting
- 4. Pieces to Compose Complex Systems

Large Model Hosting Platforms



- 1. GPT4/3.5
- 2. Integrated Tools/Multimodal Support
- 3. Models/Query Router Internal
- 4. Path to Local Deployment Less Clear



- 1. Publicly Available Models
- 2. Models/Query Router Accessible
- 3. Path to Local Deployment/Self-Hosting
- 4. Pieces to Compose Complex Systems

Query Router Access

```
{  
  "messages": [{  
    "content": "...",  
    "role": "system"  
  }, {  
    "content": "...",  
    "role": "user"  
  }],  
  "model": "mixtral",  
  "temperature": 0.2,  
  "top_p": 0.7,  
  "max_tokens": 1024,  
  "stream": True  
}
```

/chat/completions

Facilitate
Monitor
Optimize
Load Balance

```
<s>[INST]<SYS>>  
{{system_message}}  
<</SYS>>
```

```
{{instruction}} [/INST]  
{{primer}}
```

```
{  
  "context": {{context}}  
  "model": "query"/"doc"  
}
```

GPT4

Dalle-3

Embed

Mixtral

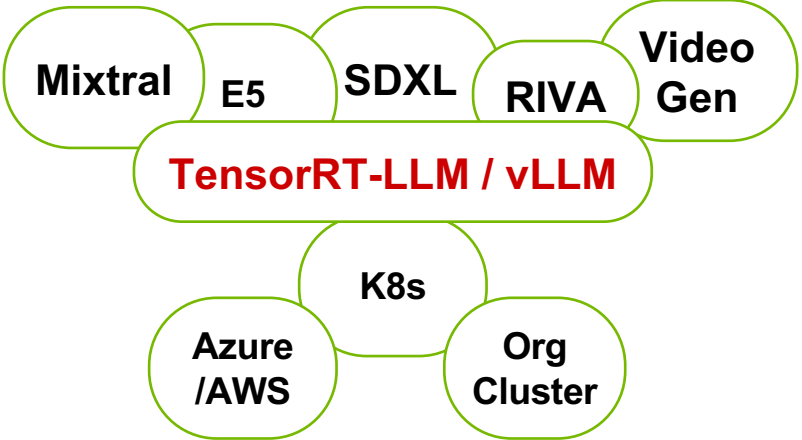
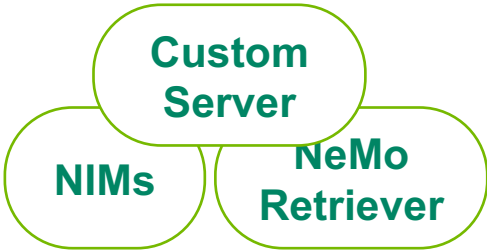
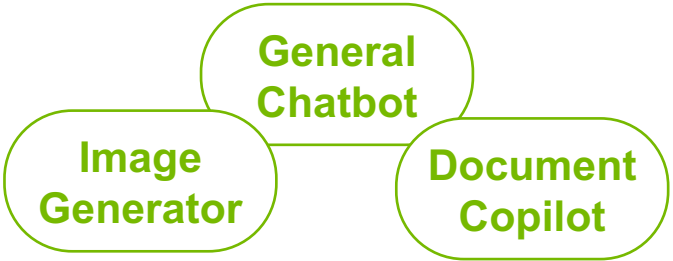
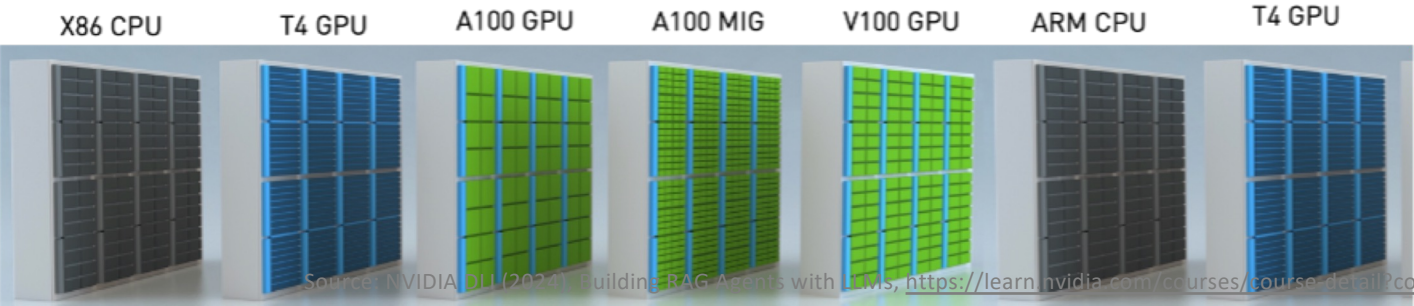
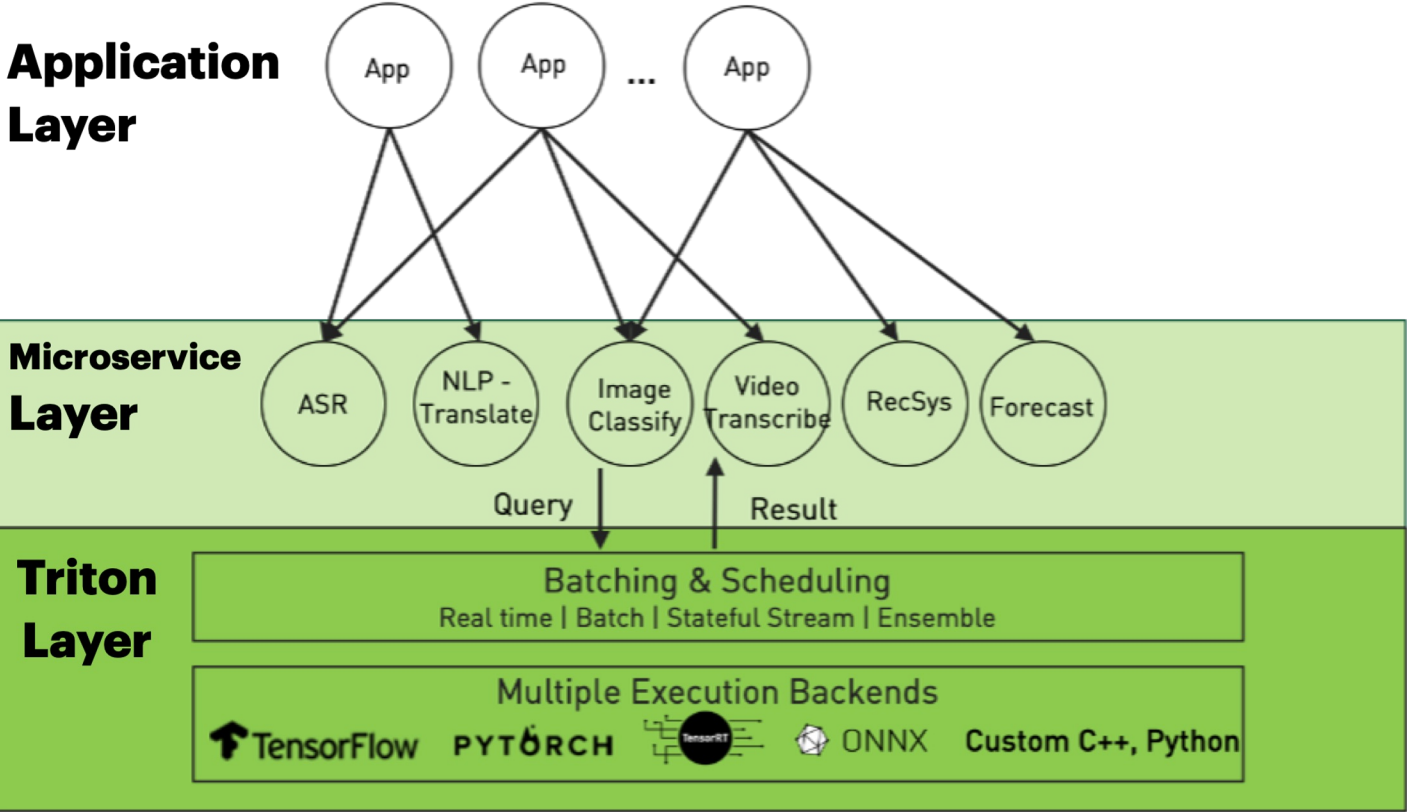
SDXL

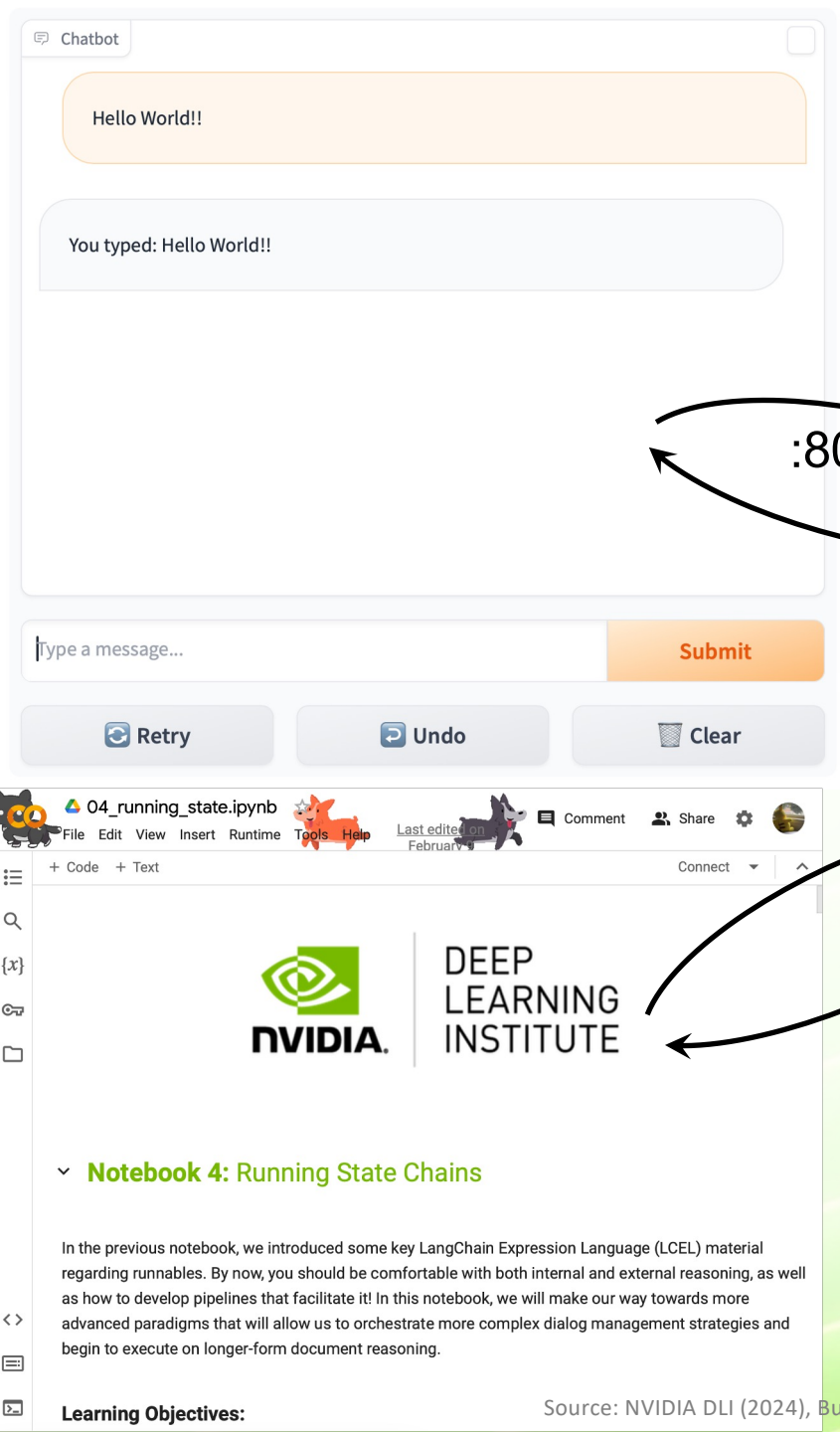
LLM

E5

Full Deployment Stack

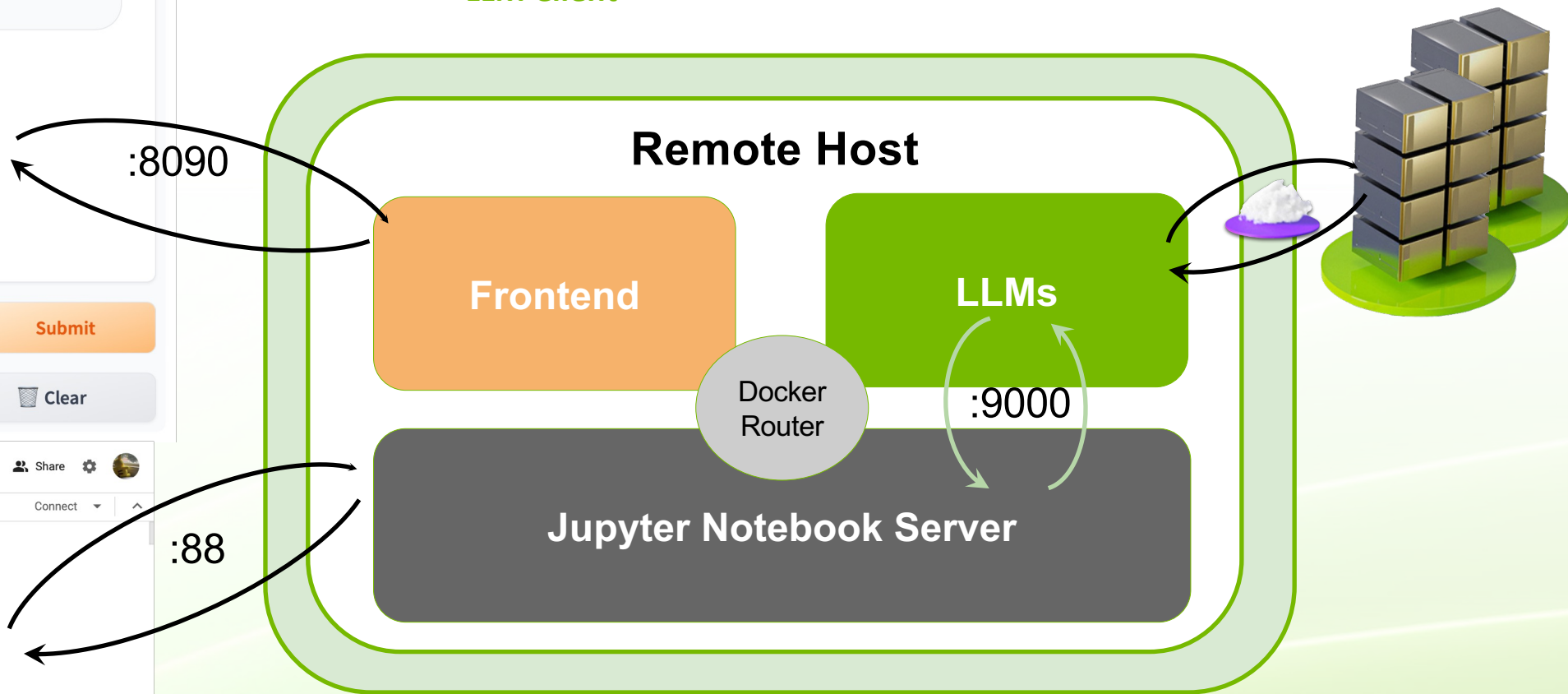
Millions of Apps | Billions of Users | Trillions of Queries



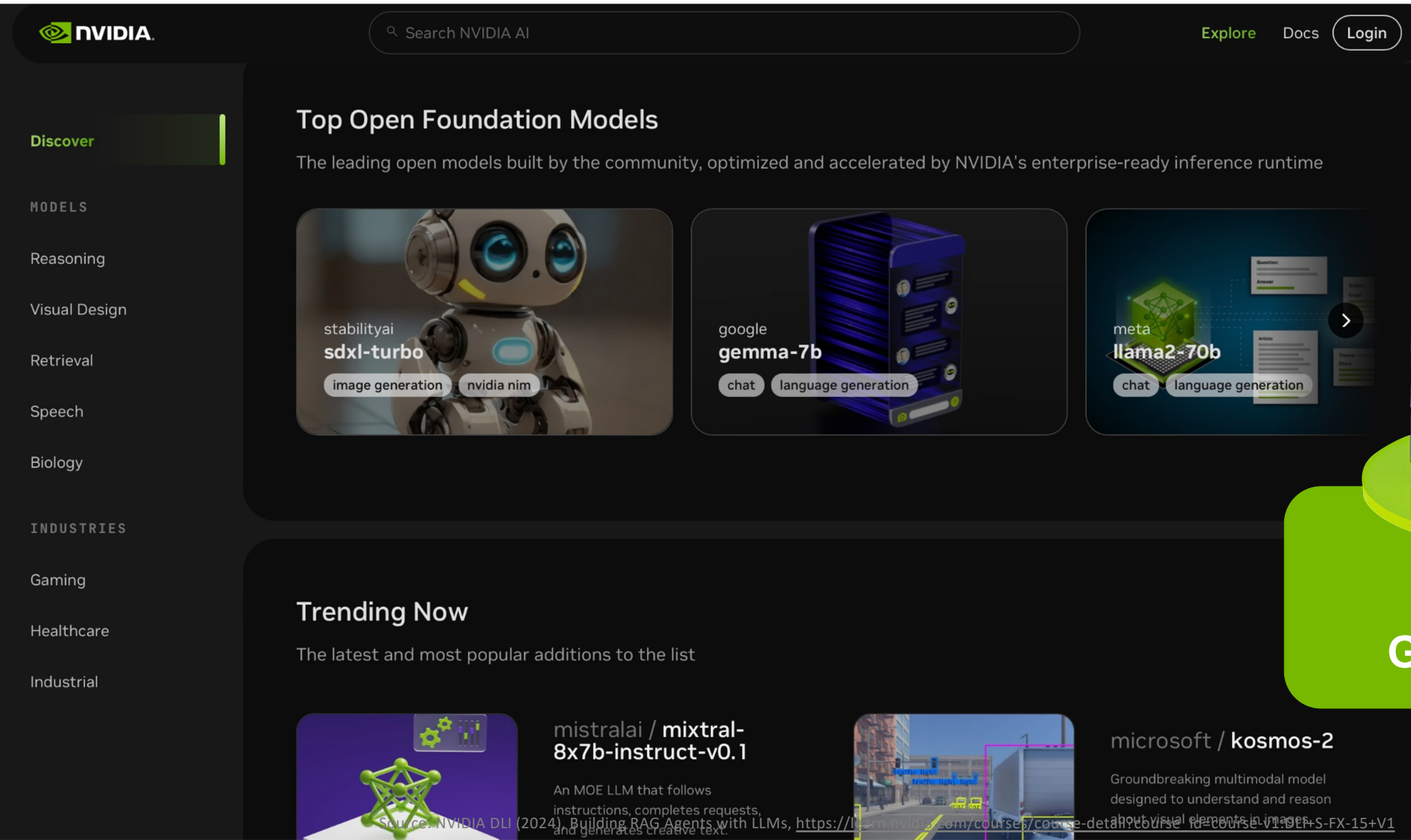


Our Environment

Jupyter Notebooks + Frontend
+ LLM Client



NVIDIA Foundation Model Endpoints



The screenshot displays the NVIDIA AI website interface. At the top, the NVIDIA logo is on the left, a search bar labeled 'Search NVIDIA AI' is in the center, and 'Explore', 'Docs', and 'Login' links are on the right. A left sidebar contains a 'Discover' button and a 'MODELS' section with categories: Reasoning, Visual Design, Retrieval, Speech, and Biology. Below these are 'INDUSTRIES' with categories: Gaming, Healthcare, and Industrial. The main content area features a 'Top Open Foundation Models' section with the subtitle 'The leading open models built by the community, optimized and accelerated by NVIDIA's enterprise-ready inference runtime'. It lists three models: stabilityai/sd-xl-turbo (image generation, nvidia nim), google/gemma-7b (chat, language generation), and meta/llama2-70b (chat, language generation). Below this is a 'Trending Now' section with the subtitle 'The latest and most popular additions to the list', featuring mistralai/mixtral-8x7b-instruct-v0.1 and microsoft/kosmos-2. A large green 3D graphic of a Rubik's cube sits on a green pedestal to the right of the models section.

Discover

MODELS

- Reasoning
- Visual Design
- Retrieval
- Speech
- Biology

INDUSTRIES

- Gaming
- Healthcare
- Industrial

Top Open Foundation Models

The leading open models built by the community, optimized and accelerated by NVIDIA's enterprise-ready inference runtime

stabilityai
sd-xl-turbo

image generation nvidia nim

google
gemma-7b

chat language generation

meta
llama2-70b

chat language generation

Trending Now

The latest and most popular additions to the list

mistralai / **mixtral-8x7b-instruct-v0.1**

An MOE LLM that follows instructions, completes requests, and generates creative text.

microsoft / **kosmos-2**

Groundbreaking multimodal model designed to understand and reason about visual elements in images.

**NVIDIA
GPU CLOUD**

NVIDIA Foundation Model Endpoints



Search NVIDIA AI

< Back

mistralai / **mixtral-8x7b-instruct-v0.1**

An MOE LLM that follows instructions, completes requests, and generates creative text.

advanced reasoning

chat

large language models

text-to-text



**NVIDIA
GPU CLOUD**

NVIDIA Foundation Model Endpoints

mistralai / **mixtral-8x7b-instruct-v0.1**

An MOE LLM that follows instructions, completes requests, and generates creative text.

advanced reasoning

chat

large language models

text-to-text

Preview

JSON

Reset Chat



NVIDIA's GPU Technology Conference (GTC) is an annual event that showcases the latest developments in graphics processing units (GPUs) and artificial intelligence (AI). Here are some of the things you can expect to see at GTC:

1. Keynote speeches: GTC typically features keynote speeches from NVIDIA's CEO, Jensen Huang, as well as other industry leaders and experts. These speeches cover the latest trends and breakthroughs in AI, data science, high-performance computing, and graphics.
2. Technical sessions: GTC offers a wide range of technical sessions, including

Type text here...

Python

Node

Shell

Get API Key Copy Code

```
from openai import OpenAI

client = OpenAI(
    base_url = "https://integrate.api.nvidia.com/v1",
    api_key = "$API_KEY_REQUIRED_IF_EXECUTING_OUTSIDE_NGC"
)

completion = client.chat.completions.create(
    model="mistralai/mixtral-8x7b-instruct-v0.1",
    messages=[{"role": "user", "content": ""}],
    temperature=0.5,
    top_p=1,
    max_tokens=1024,
    stream=True
)

for chunk in completion:
    if chunk.choices[0].delta.content is not None:
        print(chunk.choices[0].delta.content, end="")
```



**NVIDIA
GPU CLOUD**

NVIDIA Foundation Model Endpoints

mistralai / **mixtral-8x7b-instruct-v0.1**

An MOE LLM that follows instructions, completes requests, and generates creative text.

advanced reasoning

chat

large language models

text-to-text

Python Node Shell

```
curl https://integrate.api.nvidia.com/v1/chat/completions \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $API_KEY_REQUIRED_IF_EXECUTING_OUTSIDE_NGC" \
-d '{
  "model": "mistralai/mixtral-8x7b-instruct-v0.1",
  "messages": [{"role": "user", "content": ""}],
  "temperature": 0.5,
  "top_p": 1,
  "max_tokens": 1024,
  "stream": true
}'
```

Get API Key Copy Code

```
from openai import OpenAI

client = OpenAI(
    base_url = "https://integrate.api.nvidia.com/v1",
    api_key = "$API_KEY_REQUIRED_IF_EXECUTING_OUTSIDE_NGC"
)

completion = client.chat.completions.create(
    model="mistralai/mixtral-8x7b-instruct-v0.1",
    messages=[{"role": "user", "content": ""}],
    temperature=0.5,
    top_p=1,
    max_tokens=1024,
    stream=True
)

for chunk in completion:
    if chunk.choices[0].delta.content is not None:
        print(chunk.choices[0].delta.content, end="")
```



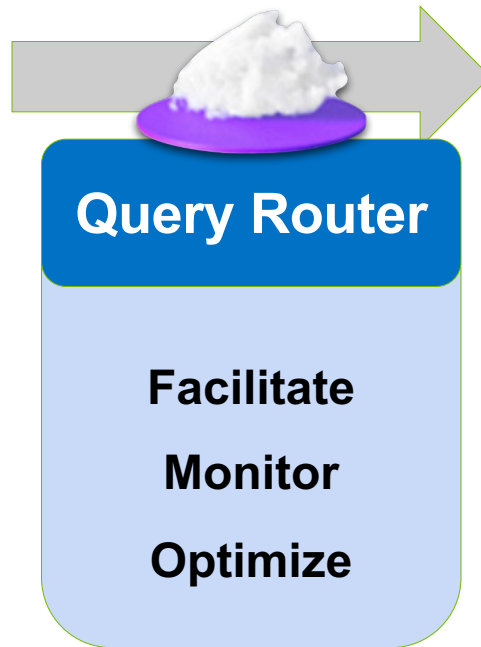
**NVIDIA
GPU CLOUD**

From Raw Requests to LangChain Model



From Raw Requests to LangChain Model

```
> llm("Hello world")  
  ChatMessage(content="hello")  
  
> embedder("Hello")  
  [0.4535437, 0.0800435, ...]
```



```
<s>[INST]<<SYS>>  
  {{system_message}}  
<</SYS>>
```

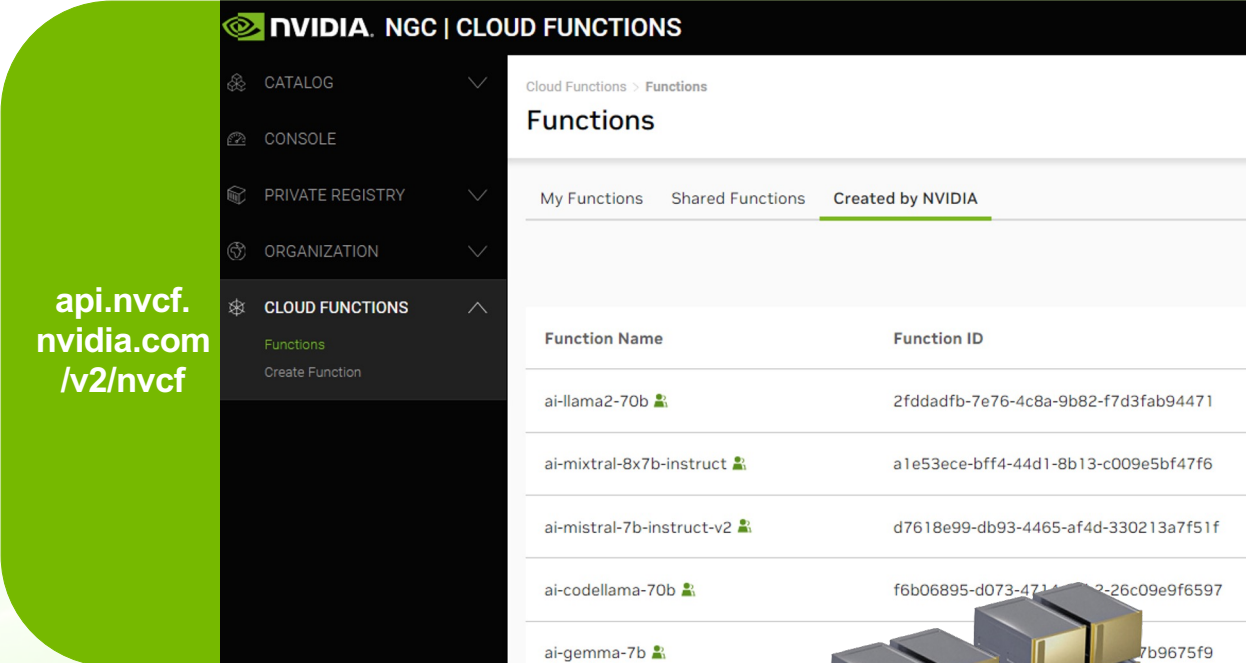
```
{{instruction}} [/INST]  
{{primer}}
```

```
{  
  "context": {{context}}  
  "model": "query"/"doc"  
}
```



LLM Interfaces

The Whole Stack



The screenshot displays the NVIDIA NGC Cloud Functions interface. On the left is a dark sidebar with a green vertical bar containing the URL `api.nvcf.nvidia.com/v2/nvcf`. The sidebar menu includes: CATALOG, CONSOLE, PRIVATE REGISTRY, ORGANIZATION, CLOUD FUNCTIONS (selected), Functions, and Create Function. The main content area shows the 'Functions' page with tabs for 'My Functions', 'Shared Functions', and 'Created by NVIDIA'. A table lists several functions:

Function Name	Function ID
ai-llama2-70b	2fddadfb-7e76-4c8a-9b82-f7d3fab94471
ai-mixtral-8x7b-instruct	a1e53ece-bff4-44d1-8b13-c009e5bf47f6
ai-mistral-7b-instruct-v2	d7618e99-db93-4465-af4d-330213a7f51f
ai-codellama-70b	f6b06895-d073-4711-82-26c09e9f6597
ai-gemma-7b	7b9675f9

Below the screenshot is a teal rounded rectangle containing the OpenAI logo on the left and a white circle with the text 'Scaled Function Deployment Solution' in the center. To the right of the circle is a 3D illustration of two stacks of server hardware.

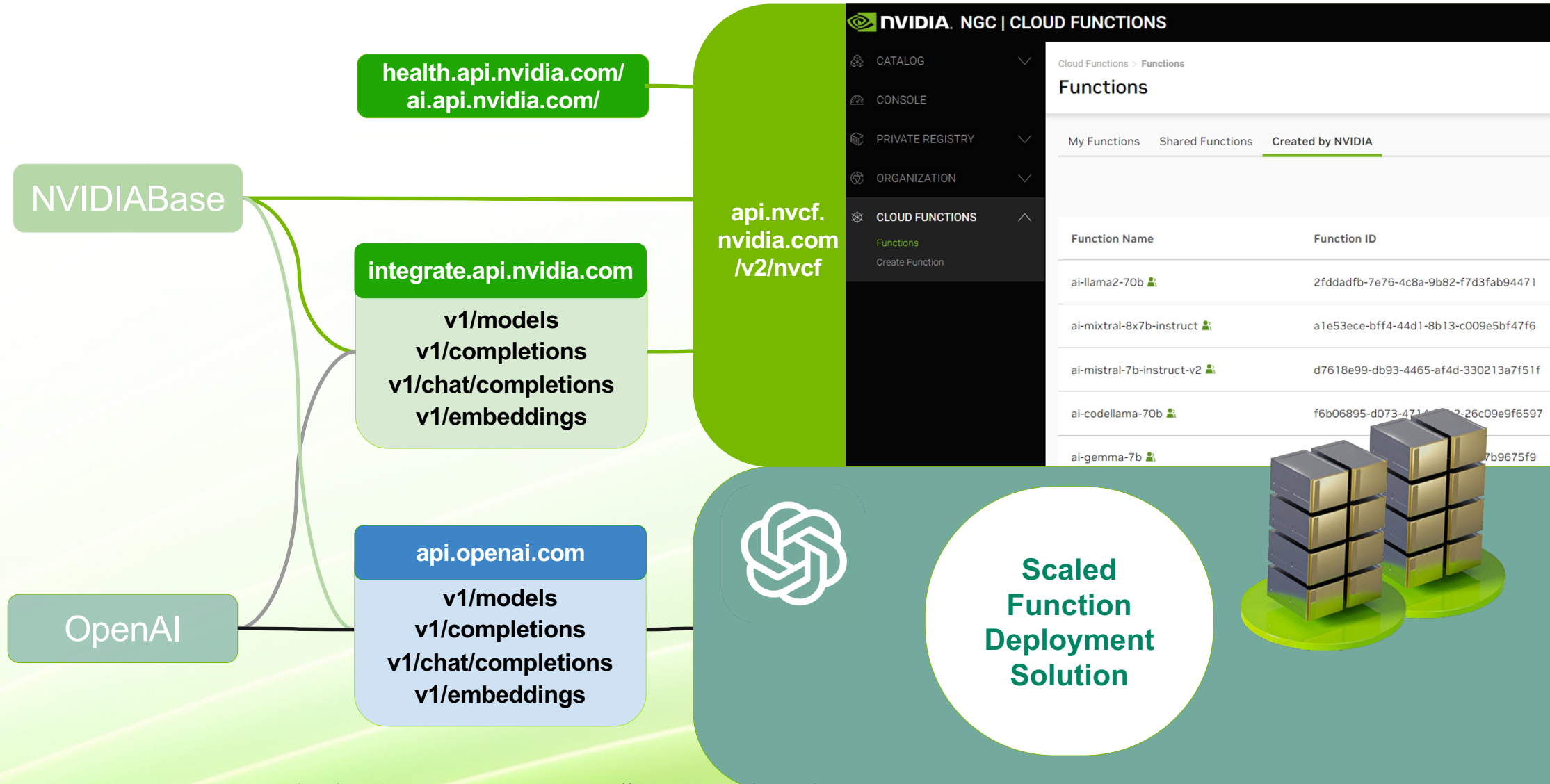
LLM Interfaces

The Whole Stack



LLM Interfaces

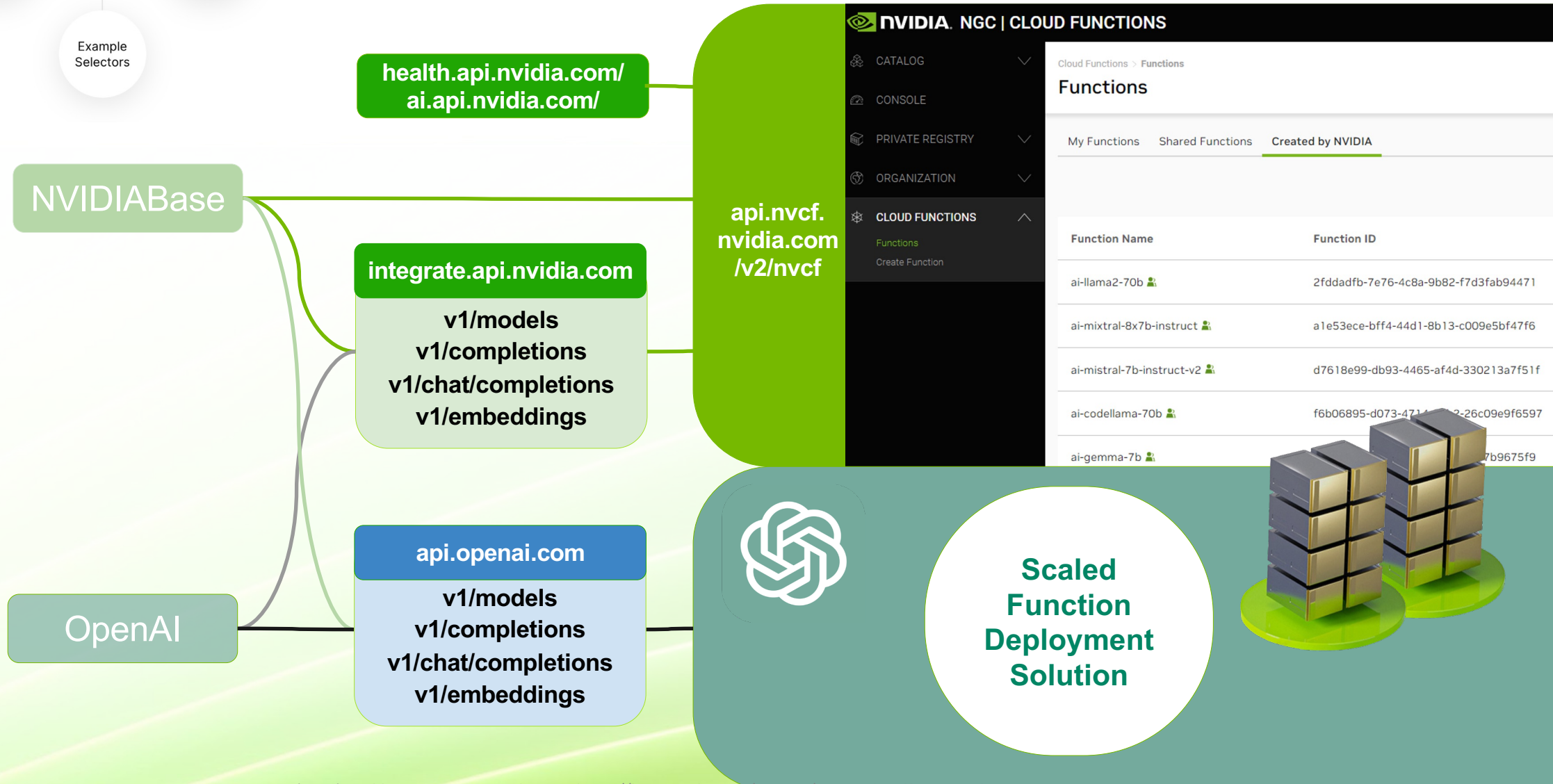
The Whole Stack





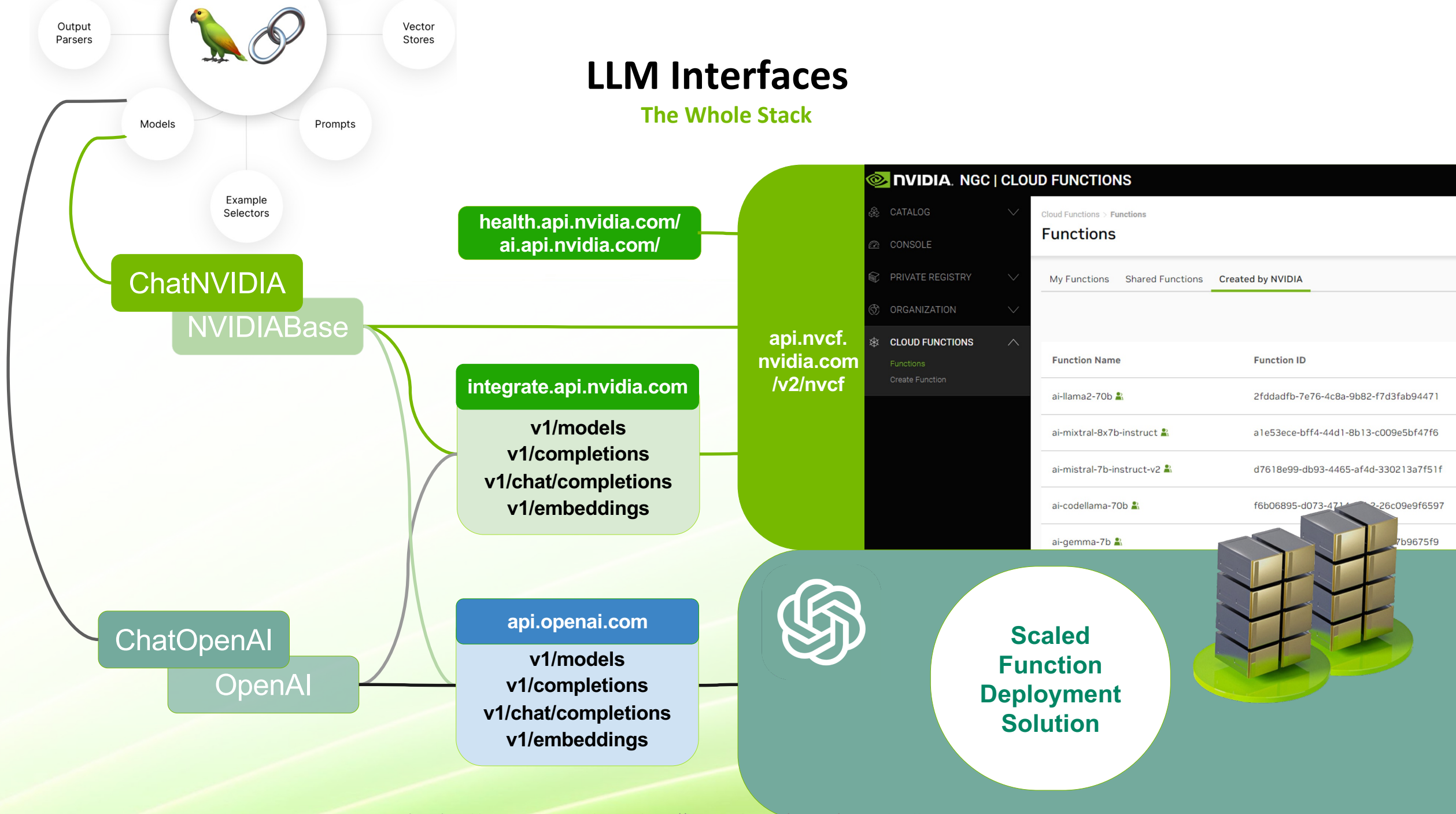
LLM Interfaces

The Whole Stack



LLM Interfaces

The Whole Stack



NVIDIA Building RAG Agents with LLMs

Part I

[0, 1, 2]

Building RAG Agents with LLMs



DEEP
LEARNING
INSTITUTE

This Lab 0 : 05 : 59 / 2 : 00 : 00

Course 1 : 47 : 00 / 32 : 00 : 00



LAUNCH



STOP TASK

Course Slides

Course Slides

Environment and LLMs

Environment [0, 1, 2]

- chatbot
- composer
- docker_router
- frontend
- imgs
- llm_client
- slides

- 00_jupyterlab.ipynb
- 01_microservices.ipynb
- 02_llms.ipynb
- 99_table_of_contents.ipynb

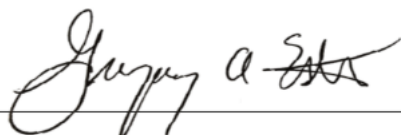
Certificate of Completion

This certificate is awarded to

Min-Yuh Day

for successfully completing

Building RAG Agents with LLMs



Greg Estes

Vice President, NVIDIA

Issue Date: : December 8, 2024

Certification ID: ed-qOCIMQaatzU8SNUNxgw |

https://learn.nvidia.com/certificates?id=ed-qOCIMQaatzU8SNUNxgw/courses/course?course_id=course-v1:DLI+S-FX-15+V1

<https://learn.nvidia.com/certificates?id=ed-qOCIMQaatzU8SNUNxgw>

References

- NVIDIA DLI (2025), Building RAG Agents with LLMs,
https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-15+V1
- NVIDIA DLI (2025), Generative AI with Diffusion Models,
https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-14+V1
- NVIDIA DLI (2025), Getting Started with Deep Learning,
https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-01+V1