Workshop Cheat Sheet

Deploy and Scale LLM Apps

Getting Started

Welcome to Anyscale! If you need any accommodations, let us know and we'll do our best to make you feel more comfortable in our space.

WiFi Network: Anyscale-Guest
WiFi Password: ProgramTheCloud

How to Participate

Join live polls, quizzes, and engage in Q&A by going to app.sli.do and enter code **#LLM**

You can also ask questions live by calling over an instructor.

How to Access Anyscale

Log-in to your cluster at console.anyscale.com

We have already made you an account. Check your email for your unique credentials for this workshop.

GitHub Repository

Everything you need will be mounted to your Anyscale cluster.

If you want to experiment locally after this event, all the materials live in github.com/ray-project/llms-in-prod-workshop-2023, and you can find bonus notebooks at github.com/ray-project/ray-educational-materials.

Free Resources

We'll be holding office hours next week. Come talk to us about extending your Anyscale cluster access to explore more with Ray!

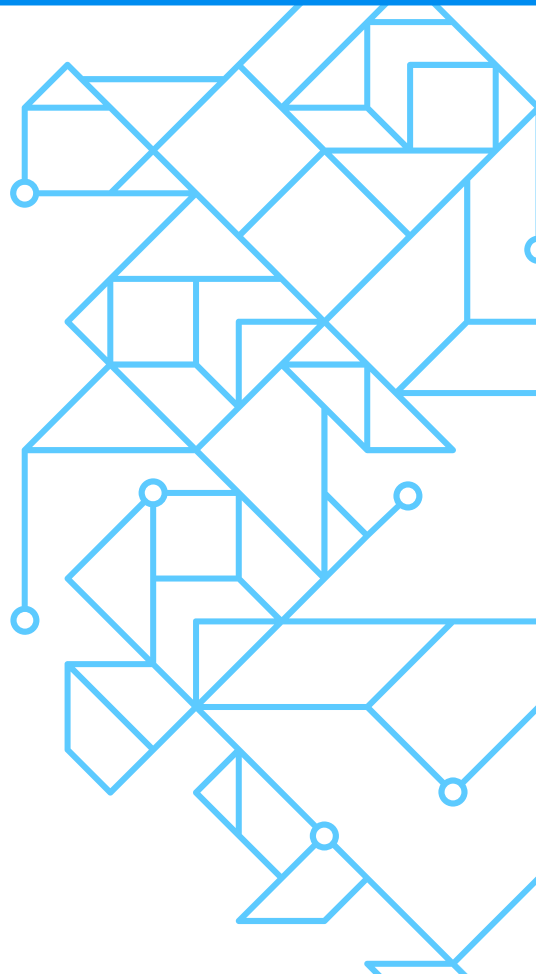
Docs - docs.ray.io
Blog - anyscale.com/blog
YouTube - youtube.com/anyscale

Ray Summit 2023

Join the Ray global community in San Francisco September 18-20 to explore the future of machine learning and scalable AI.

Register at raysummit.anyscale.com and get 15% off with code **RAYMEETUP.** Fill out the survey bit.ly/llms-feedback for 20% off!

anyscale

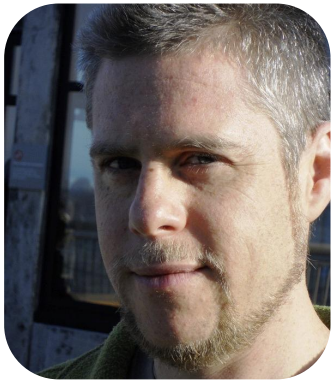


👋 Welcome!

We're happy to have you here.

👨‍💻 Meet the team!

Emmy



Adam

Kamil

🥅 Our goals.

📚 Bridge the gap between dev and prod.

Introduce Ray for production-grade LLM systems.

💻 Learn by doing.

Hands-on, relevant coding examples.

📣 Reinforce through discussion.

Polls, live Q&A, conversation at tables.

👋 Cultivate community.

Find friends, network, knowledge share.

The Plan

Here’s what to expect today.

Today's agenda.

3:30pm (25 min)	**Demo:** Run an LLM App in 15 Minutes
3:55pm (10 min)	**Talk:** Introduction to Ray and Anyscale
4:05pm (15 min)	**Coding Lab:** Ray Serve for Scalable Deployments
4:20pm (10 min)	Coffee Break
4:30pm (50 min)	**Coding Lab:** LLMOps - Launching an LLM QA App in Production
5:20pm (10 min)	**Talk:** Resources for Further Exploration

✅ Tech check.

Participating via app.sli.do

- Join with code **#LLM**
- Answer polls.
  - Enter your name to compete!
- Ask questions.
  - Pose your own and upvote others.
  - TAs will be answering questions on a rolling basis.

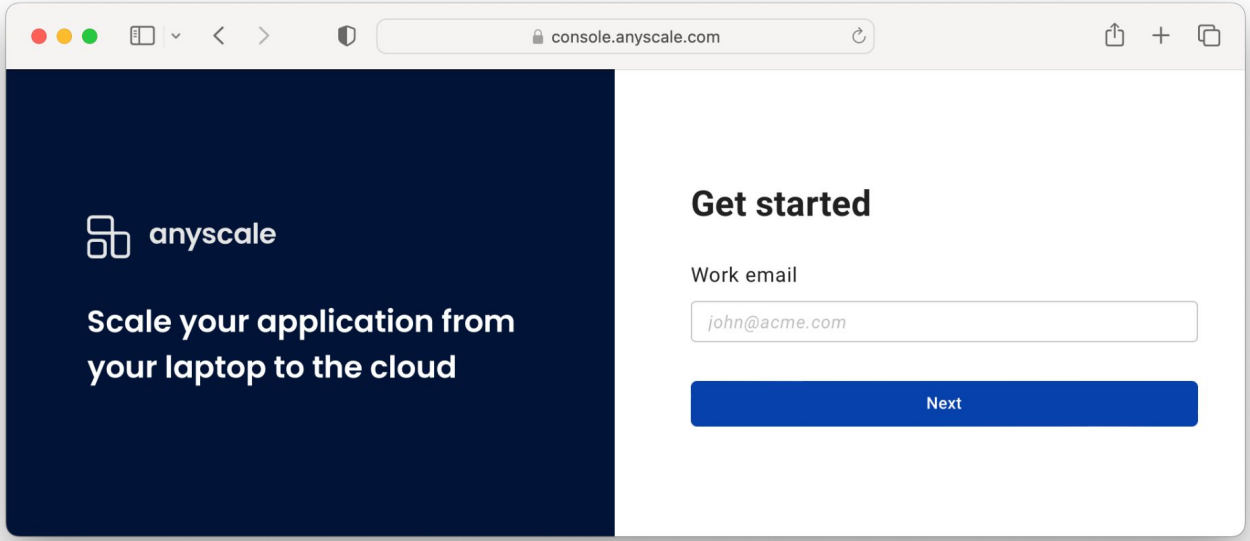
✅ Tech check.

 Accessing Anyscale clusters.

- All work will be in Anyscale provisioned clusters.
- Our GitHub repo will be mounted automatically.
- Access begins now.
  - Check your email for login information.
  - Step-by-step instructions to follow.

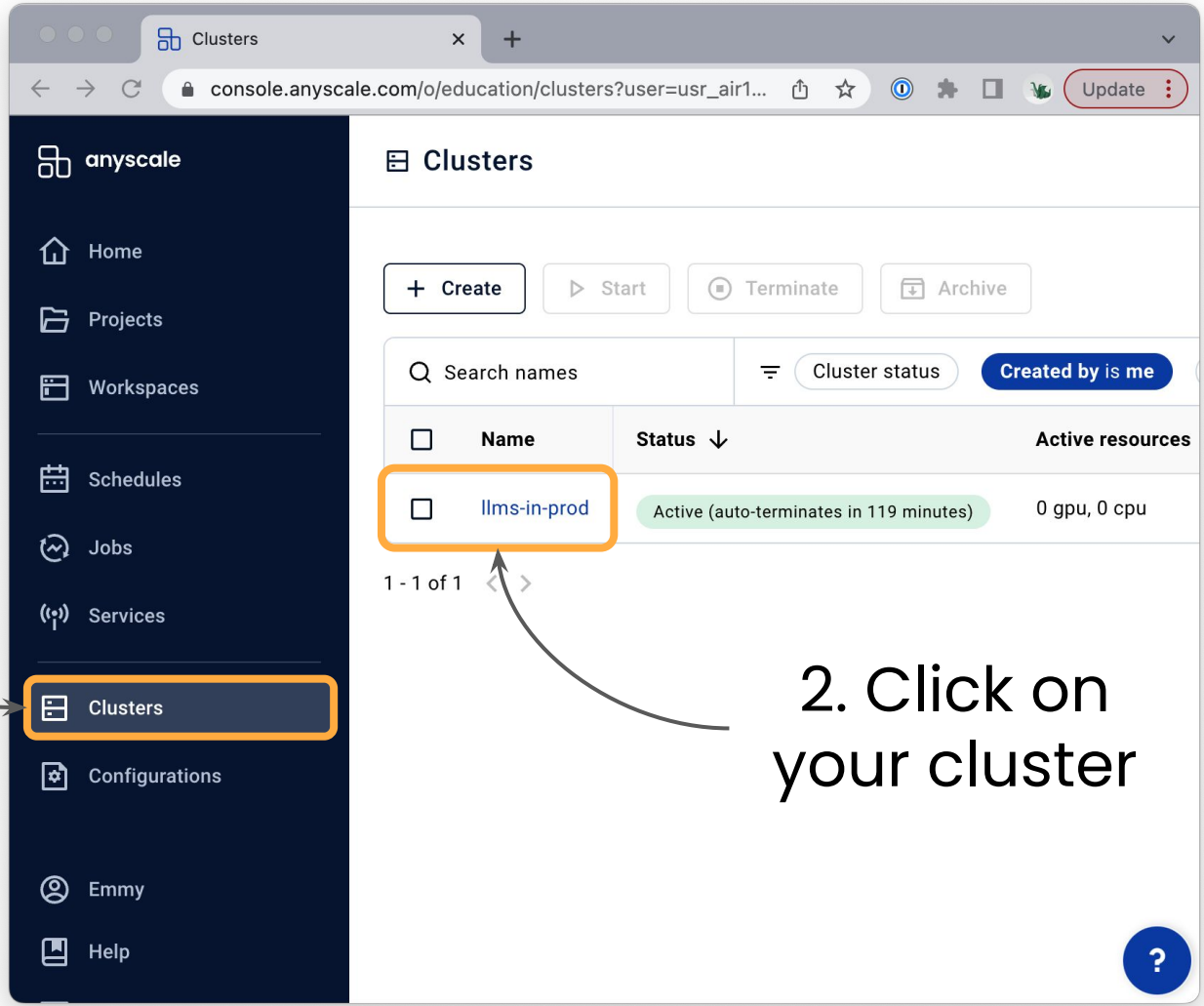
Anyscale login

Link to Anyscale cluster: console.anyscale.com



Enter the **unique credentials** sent to your email!

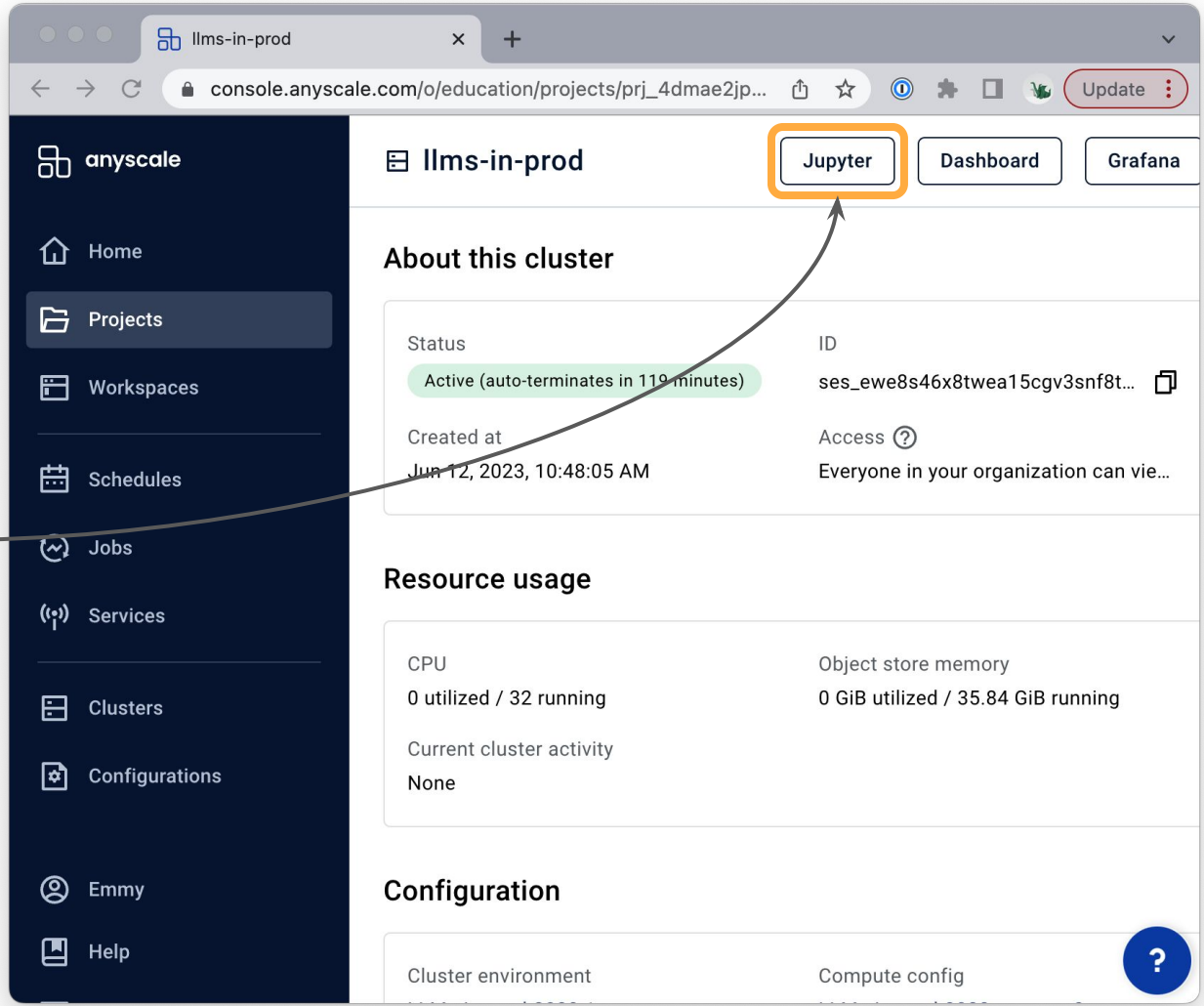
1. Select “Clusters”



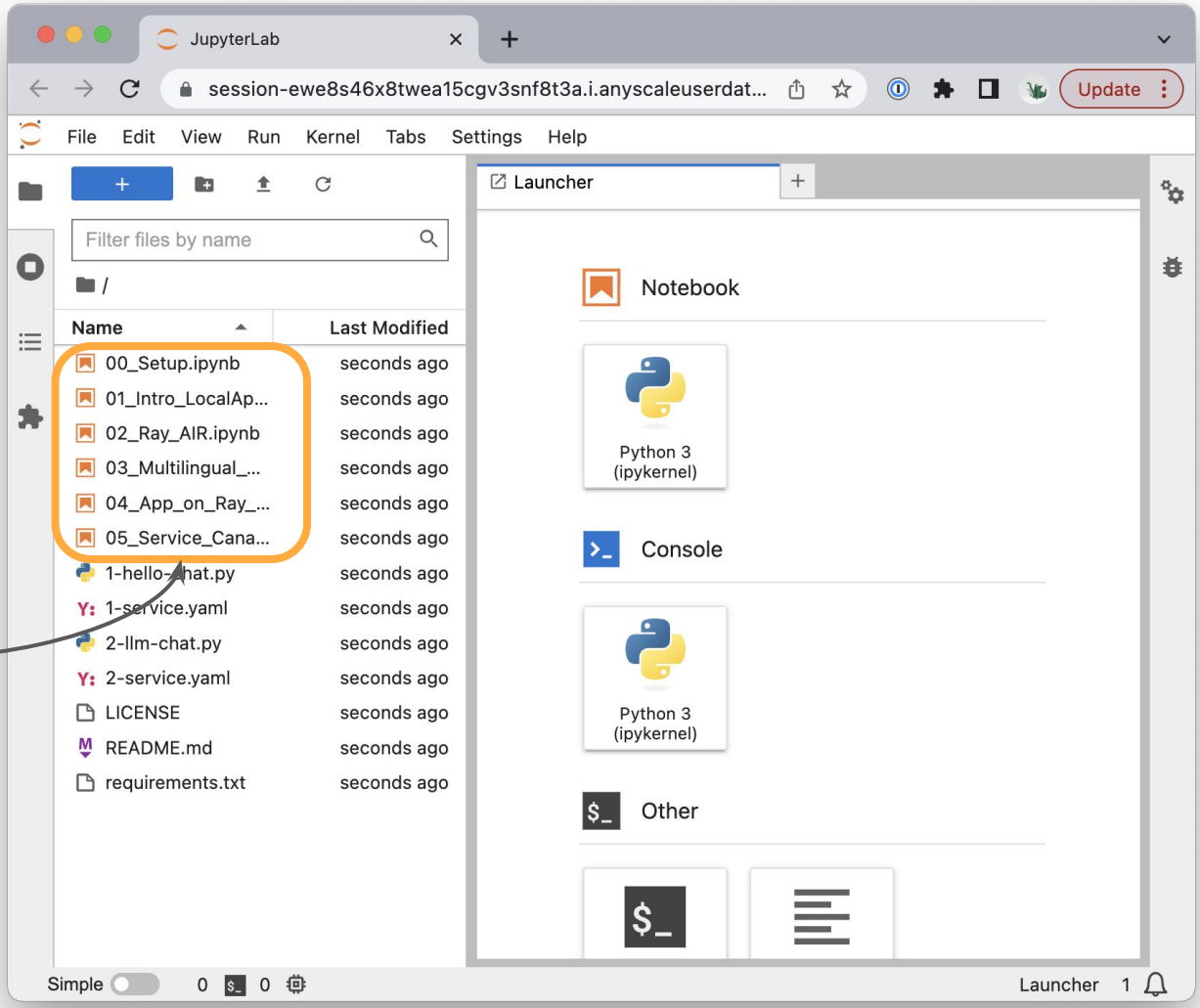
2. Click on your cluster

3. “Start” the cluster

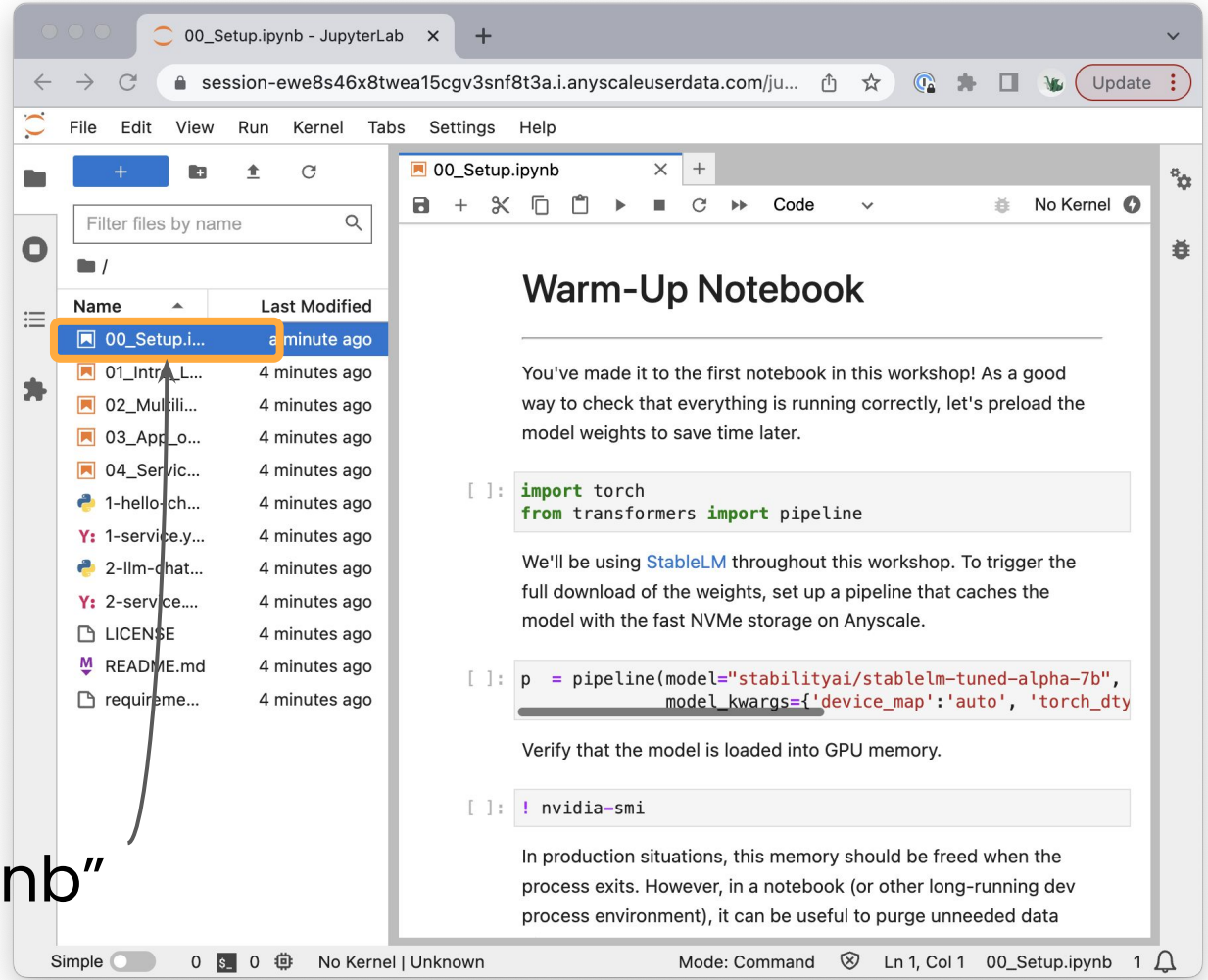
4. Then select “Jupyter”



5. View modules here



6. Run
“00_Setup.ipynb”



LLM App in 15 Min

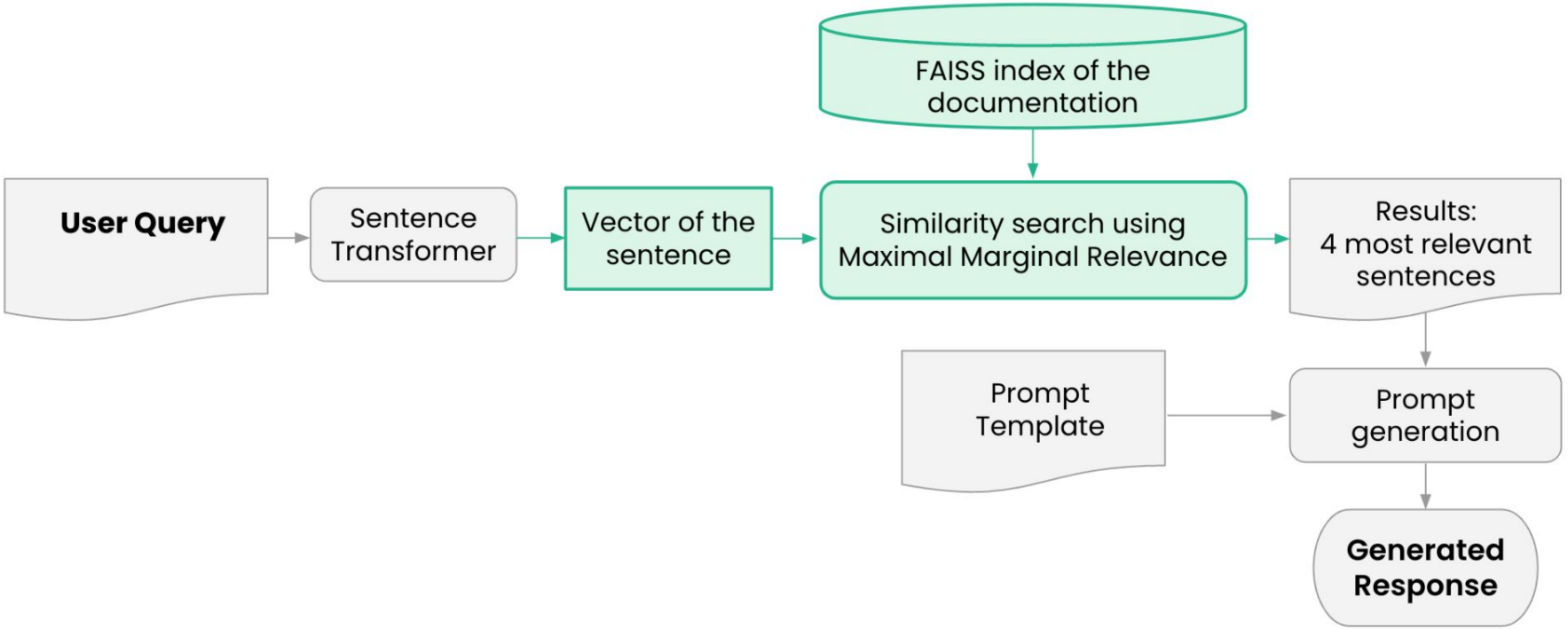
Run a question answering application locally.

The task.

Retrieval-based question answering.

- `question` - A prompt and/or question.
- `context` - Additional contextual information from documents

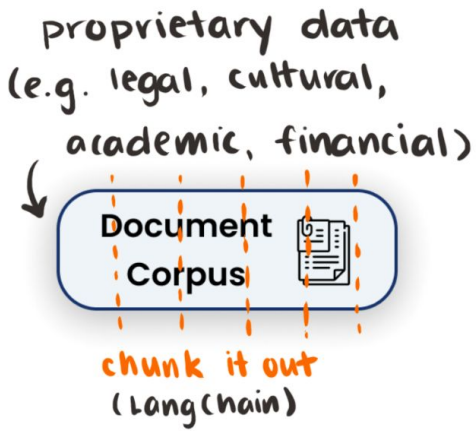
- `answer` - The AI generated response.


FAISS index of the documentation
User Query
Sentence Transformer
Vector of the sentence
Similarity search using Maximal Marginal Relevance
Results: 4 most relevant sentences
Prompt Template
Prompt generation
Generated Response

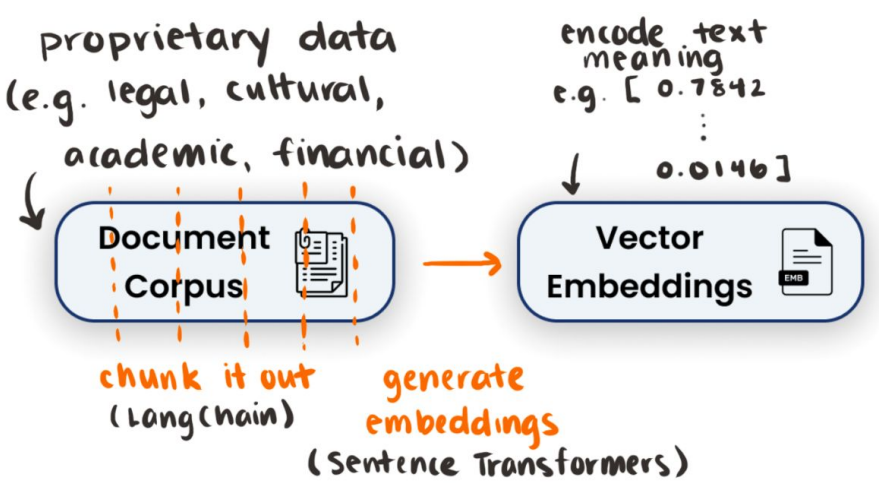
The Retrieval QA Pattern

proprietary data
(e.g. legal, cultural,
academic, financial)

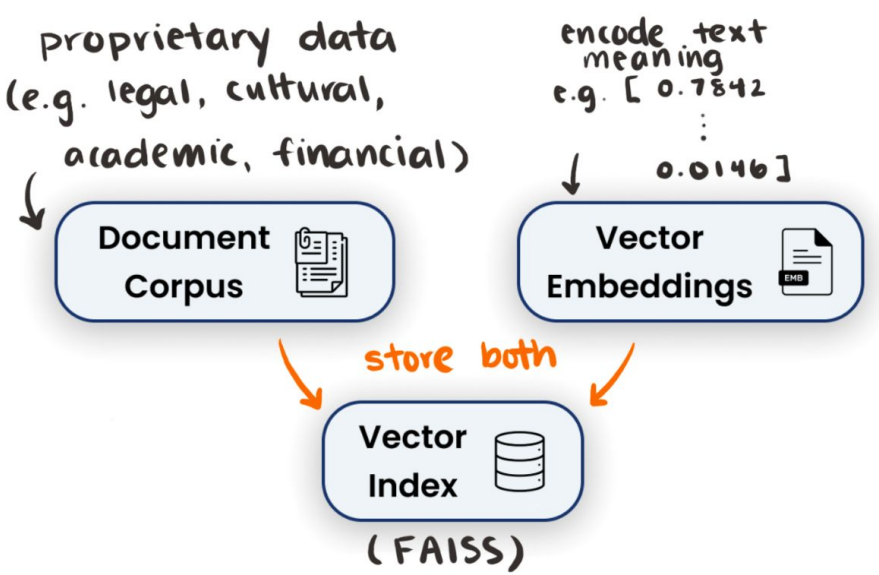
The Retrieval QA Pattern



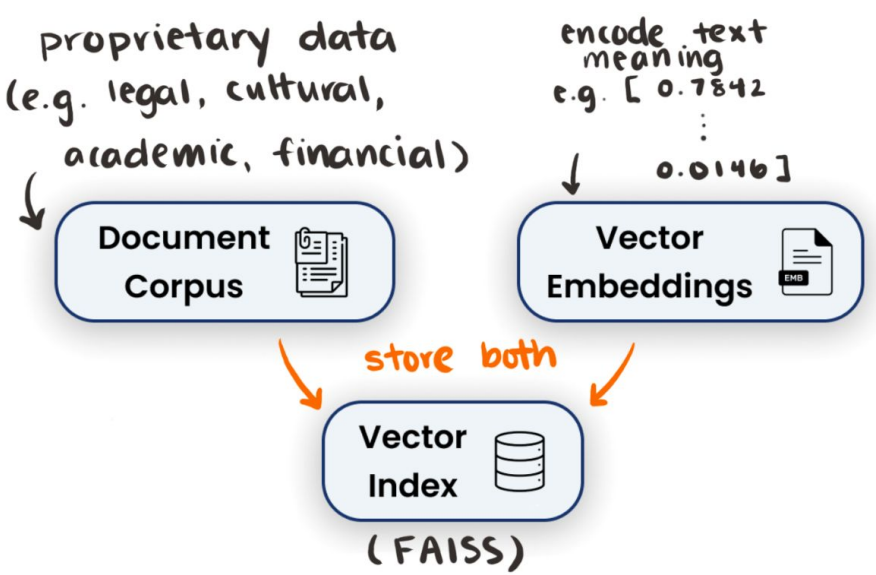
The Retrieval QA Pattern



The Retrieval QA Pattern

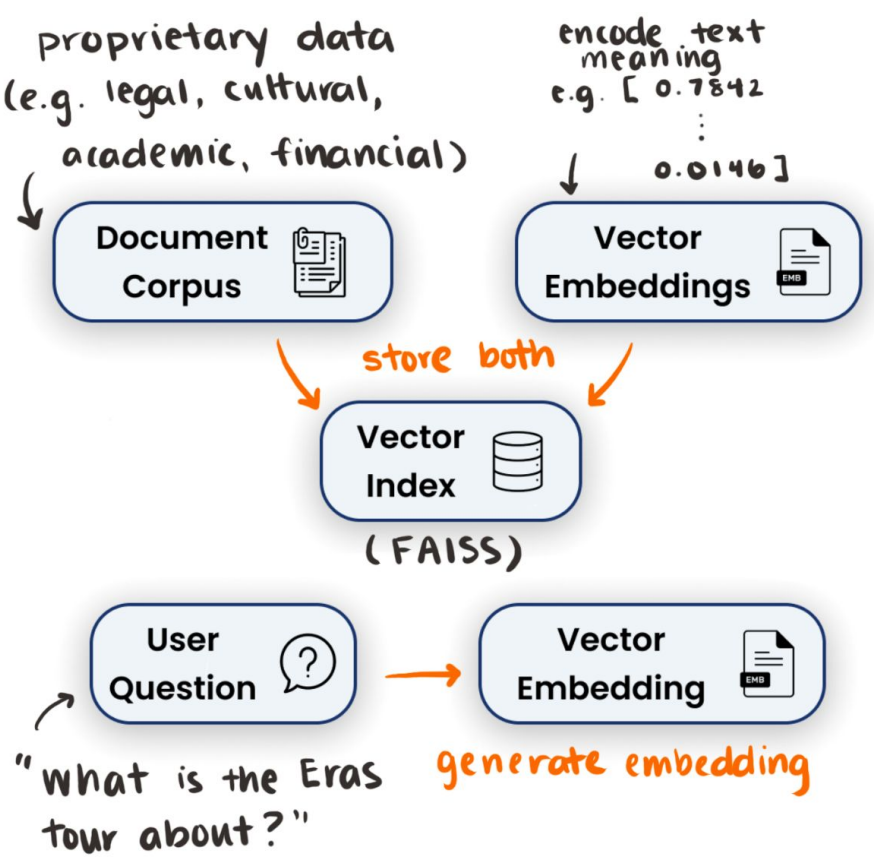


The Retrieval QA Pattern

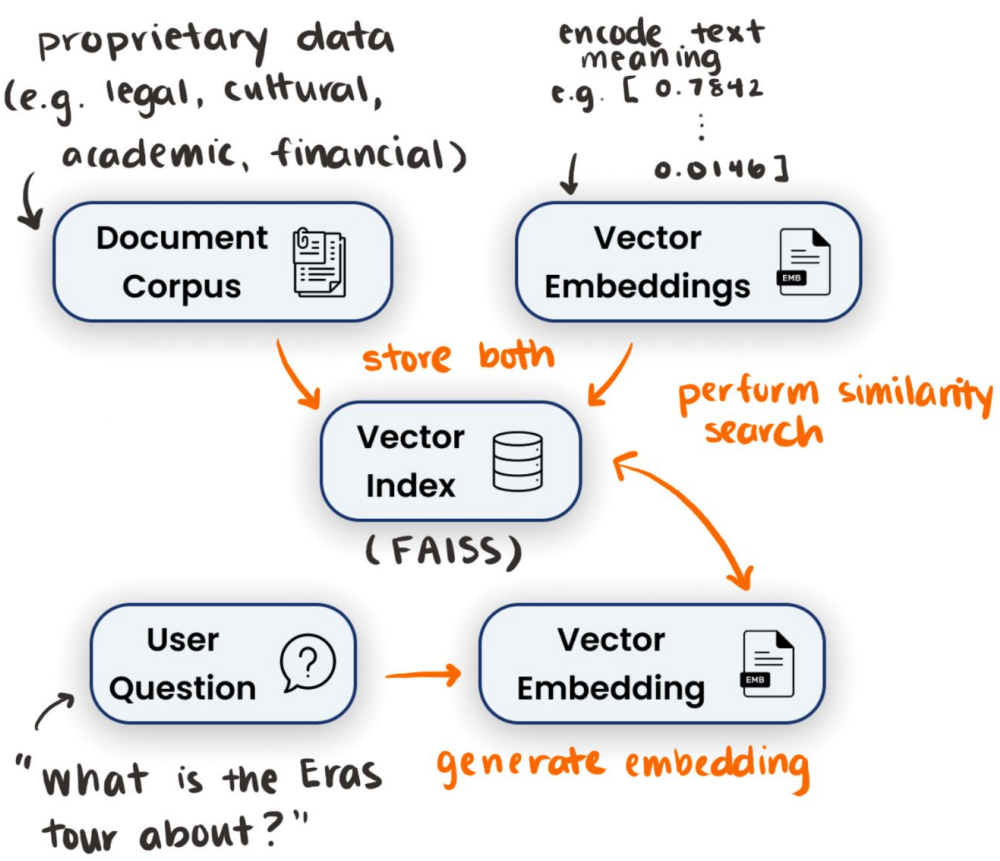


The Retrieval QA Pattern

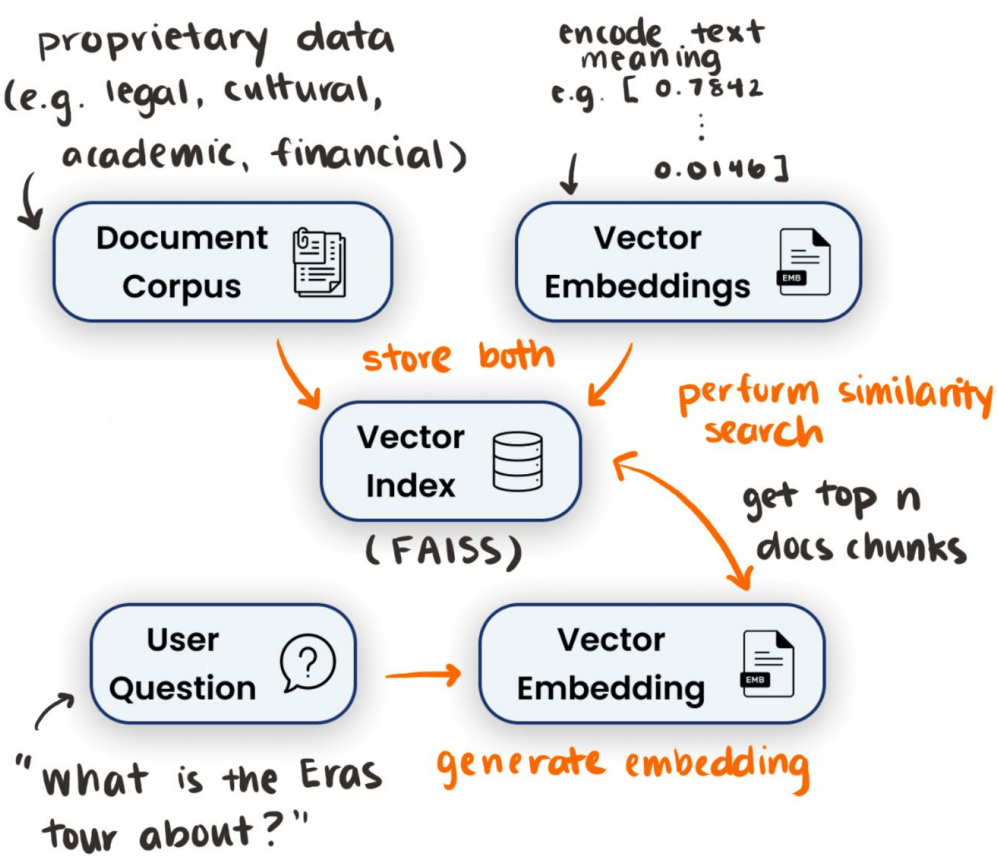
The Retrieval QA Pattern



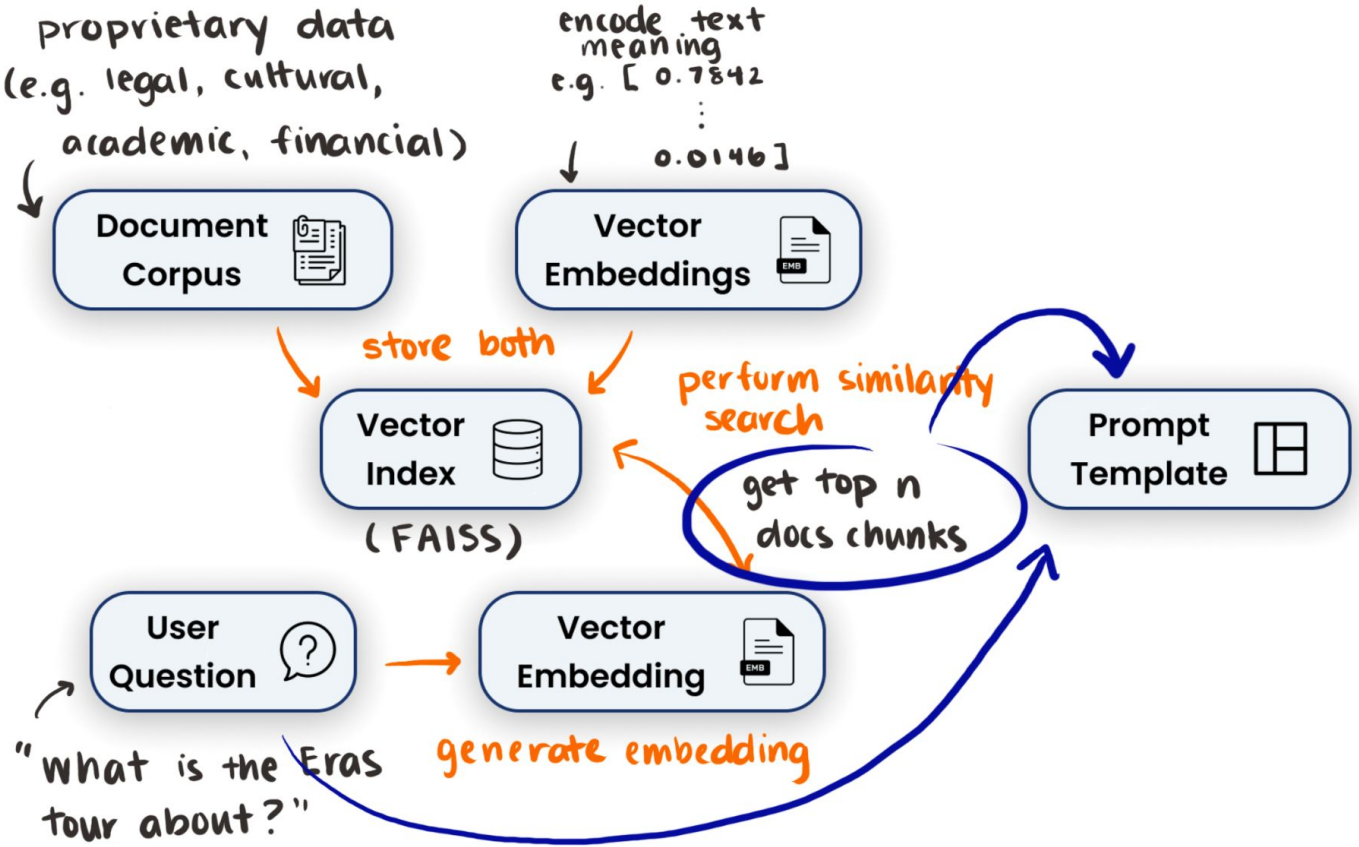
The Retrieval QA Pattern



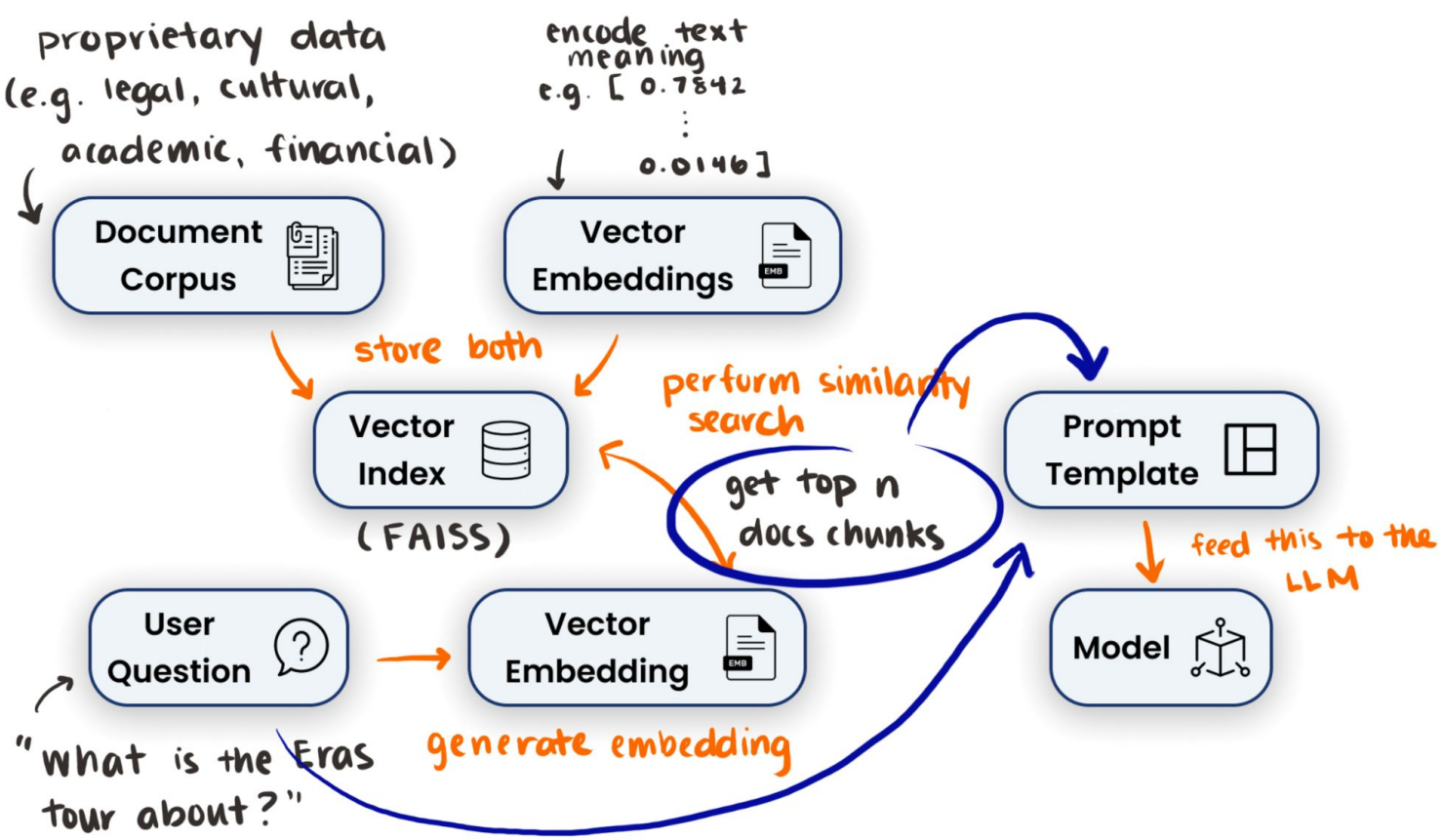
The Retrieval QA Pattern



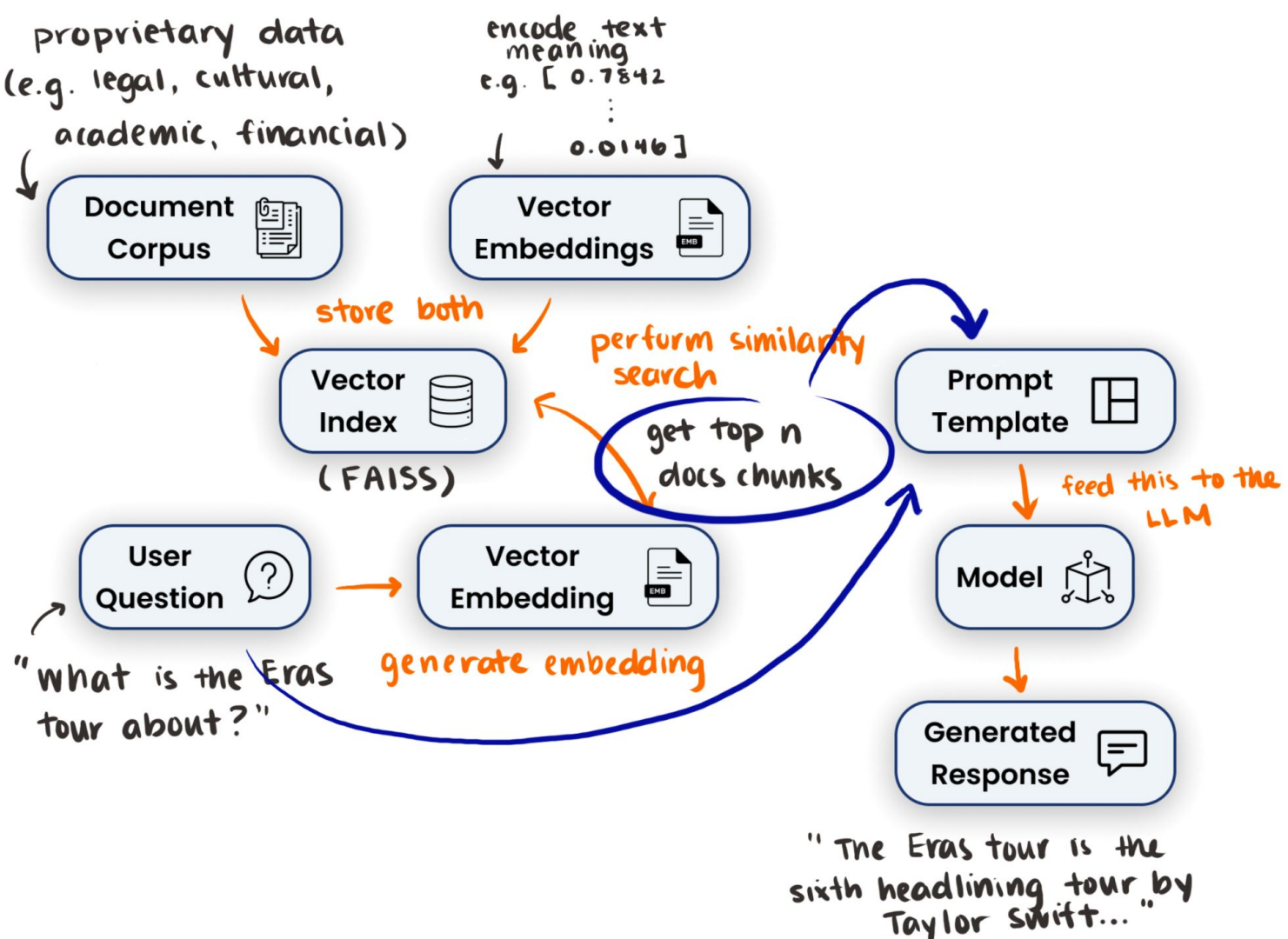
The Retrieval QA Pattern

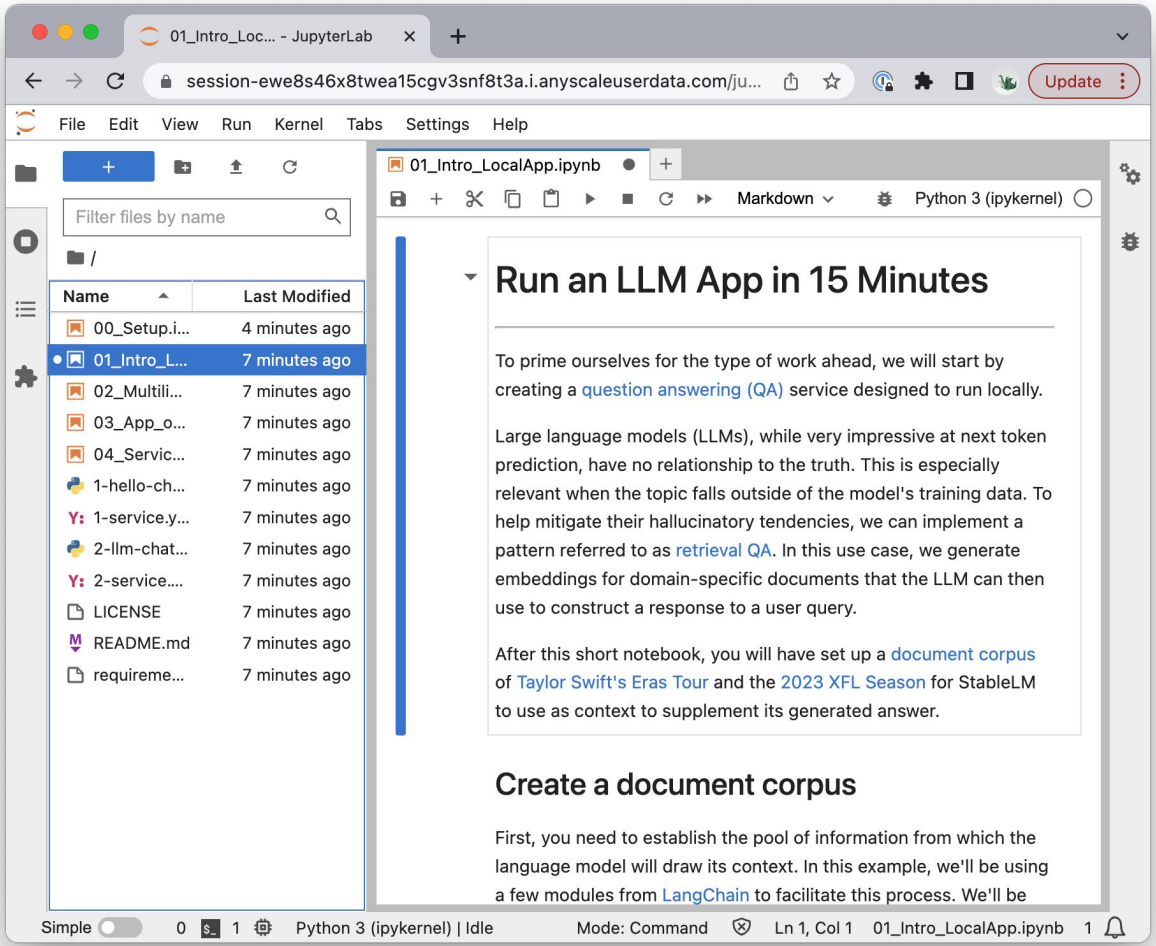


The Retrieval QA Pattern



The Retrieval QA Pattern





01_Intro_LocalApp.ipynb

slido

Audience Q&A Session

ⓘ Click **Present with Slido** or install our Chrome extension to show live Q&A while presenting.

From local to cloud

An introduction to Ray and Anyscale.

💼 Potential use cases.

Imagine...

- **Customer service** with long memory of previous tickets
- **Legal assistant** with access to a corpus of legal documents
- **Academic researcher** able to pull from papers and articles

Let’s move to production!

✅ **Tested thoroughly** on your local machine.

✅ Refactored from notebooks to a **reusable, encapsulated** format.

✅ Hit the **accuracy** and **latency** benchmarks we’re okay with.

What could go wrong?

Everything that went wrong.

Infrastructure

❌ Deployment strategy

Which cloud, how much storage, how much compute

❌ Load balancing

Making sure no surge in traffic breaks the entire system.

❌ Fault tolerance

Dealing with disaster and building in redundancy.

Everything that went wrong.

Maintenance

❌ Monitoring and logging

Inspecting performance, error tracking, metrics.

❌ Continual learning

Swapping in new data, model, and prompt versions.

❌ Dependency management

Ensuring consistent execution of complicated LLM systems.

Everything that went wrong.

Cost

❌ Scaling

Orchestrating large-scale deployments that adjust to traffic.

❌ Resource management

Precise resource allocation, using spot instances, batching

❌ Proprietary vs. OSS models

Pay through the teeth or go the self-hosted route.

😭 Everything that went wrong.

Trap Doors

❌ Security and privacy

Working with sensitive data, breaches, unauthorized access.

❌ Ethics and bias mitigation

Monitoring a non-deterministic app for problematic content.

❌ Inflexibility

Painting yourself into a corner with choices you made.

🌟 The wishlist.

✨ **Easy scaling and reliability**

"I got into this for ML, not for infrastructure management."

✨ **Efficiency and performance**

Built-in optimizations and ability to control when needed.

✨ **Extensibility**

Flexible integrations with other frameworks, clouds, and tools.

✨ **Observability tooling**

Inspect the infrastructure and ML application layers.

✨ **Intuitive cost control**

Clarity into $$$-eating resources and inefficiencies.

An introduction to Ray

A general-purpose, open source, distributed compute framework for scaling AI applications in native-Python.

An introduction to Ray

A **general-purpose**, open source, distributed compute framework for scaling AI applications in native-Python.

An introduction to Ray

A general-purpose, **open source**, distributed compute framework for scaling AI applications in native-Python.

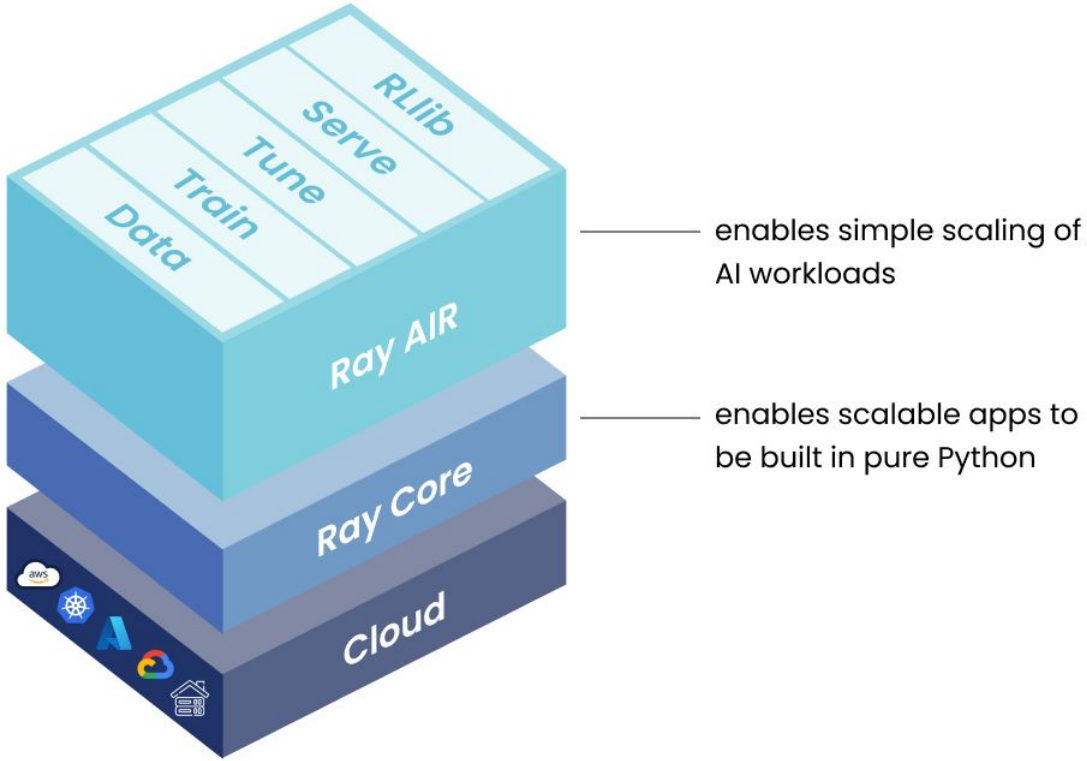
An introduction to Ray

A general-purpose, open source, **distributed compute framework** for scaling AI applications in native-Python.

An introduction to Ray

A general-purpose, open source, distributed compute framework for **scaling AI applications in native-Python.**

An introduction to Ray



↔ End-to-end ML scaling.

An introduction to Anyscale

Fully managed scalable compute platform built on Ray

- Easily move from development to production
- Abstract away setting up and managing clusters so developers can focus on the ML application layer
- Features like: workspaces, jobs, services, observability, access control

CATEGORY	FEATURE	OPEN SOURCE RAY	ANYSCALE PLATFORM™	DESCRIPTION
Development	Development experience	Basic	Workspaces with hosted VSCode and Jupyter Notebooks	Managed VSCode and Jupyter Notebooks with git integration. Workspaces provide a similar development experience on a large cluster as on a laptop.
Development	Sharing artifacts	None	Easily share application's code and environment.	Provide the ability to duplicate the application's code and environment from Workspace, Jobs and Services for development and debugging. Share and collaborate with other users.
Development	Dependency management	None	Dependency management	Anyscale manages dependencies for the latest Python release with CUDA. Anyscale also provides various dependency management capabilities for users to integrate with their existing infrastructure simply and effectively.
Development	Cluster Startup Time	Basic	Faster cluster startup time.	Anyscale optimizes cluster startup time, allowing for a quicker development cycle.
Production	Job Submission	Basic	Job submission	Anyscale provides major enhancements over Ray jobs, including cron support, retries, and persistent outputs along with email notifications for retries & failures
Production	Services	Basic	Services	Anyscale provides high availability service deployments with autoscaling and observability capabilities for workloads such as ML inference.
Production	Observability	Ray Dashboard	Grafana + Enhanced Ray Dashboard + Persistent logs	With OSS, the developer needs to instrument her own logging and monitoring stack using Prometheus or Grafana and build additional infrastructure for logs persistence and access controls. This is all managed for you on Anyscale.
Cost saving	Cluster auto suspension	None	Built-in	Anyscale provides automatic cluster termination to reduce wasted spend.

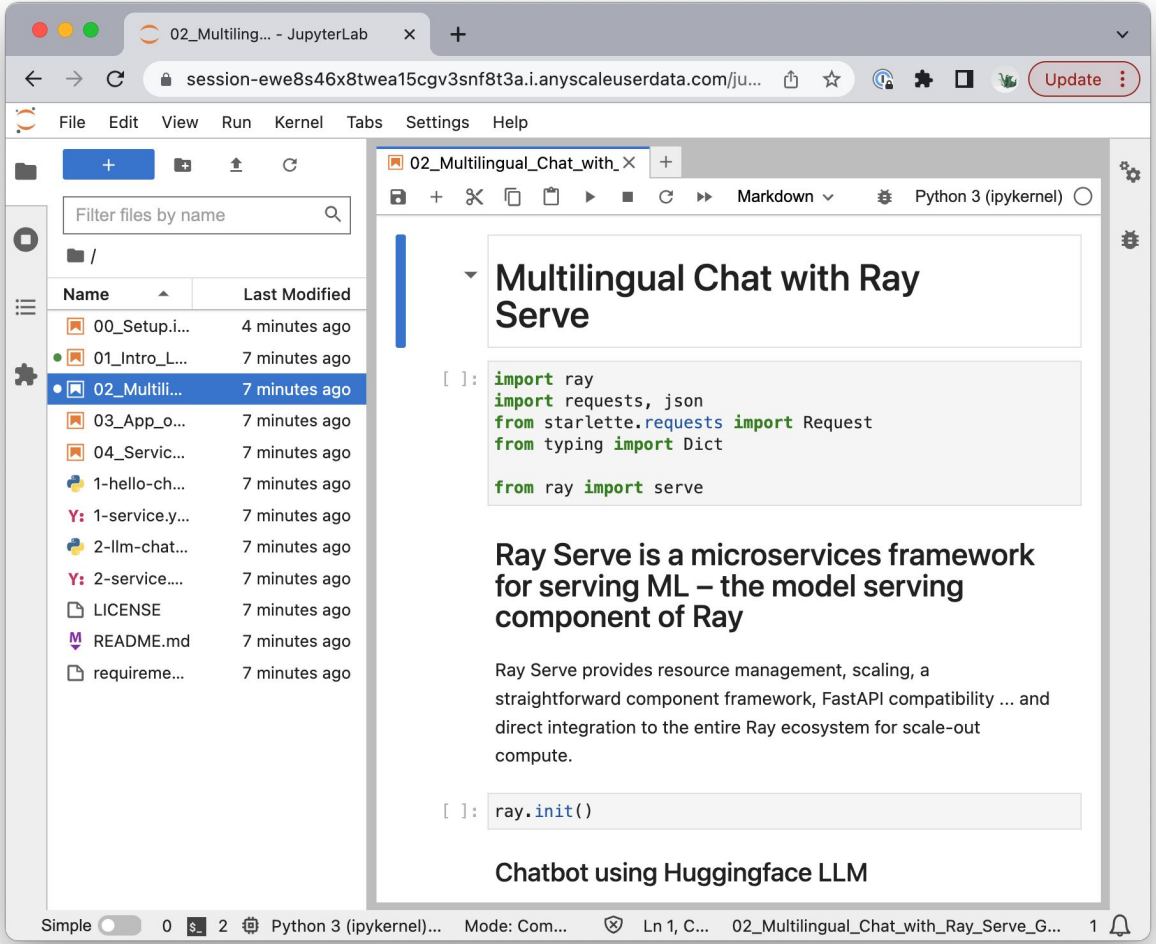
slido

Audience Q&A Session

ⓘ Click **Present with Slido** or install our Chrome extension to show live Q&A while presenting.

Scale Out

An introduction to Ray Serve for scalable deployments.



02_Multilingual_Chat_with_Ray_Serve_GPU.ipynb

slido

Audience Q&A Session

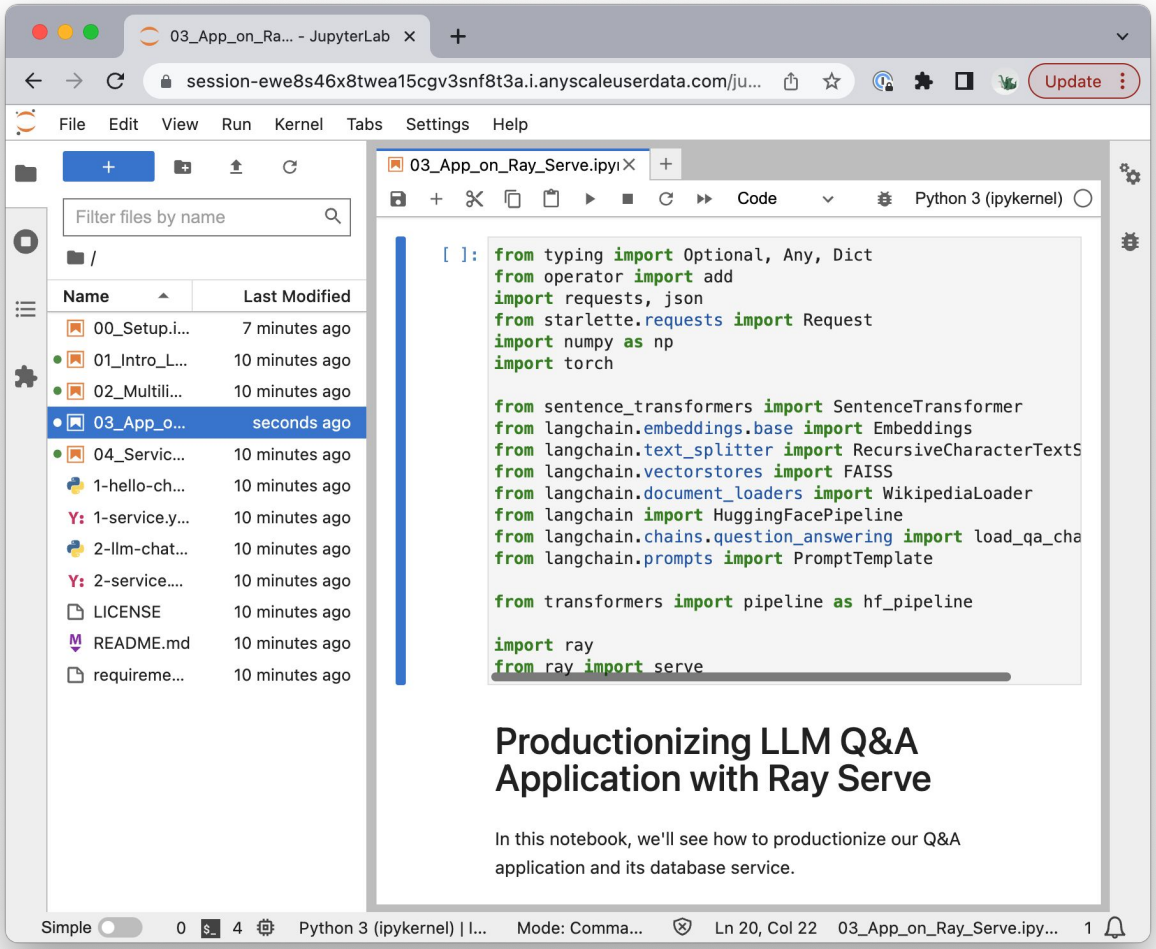
ⓘ Click **Present with Slido** or install our Chrome extension to show live Q&A while presenting.

Time for a Break!

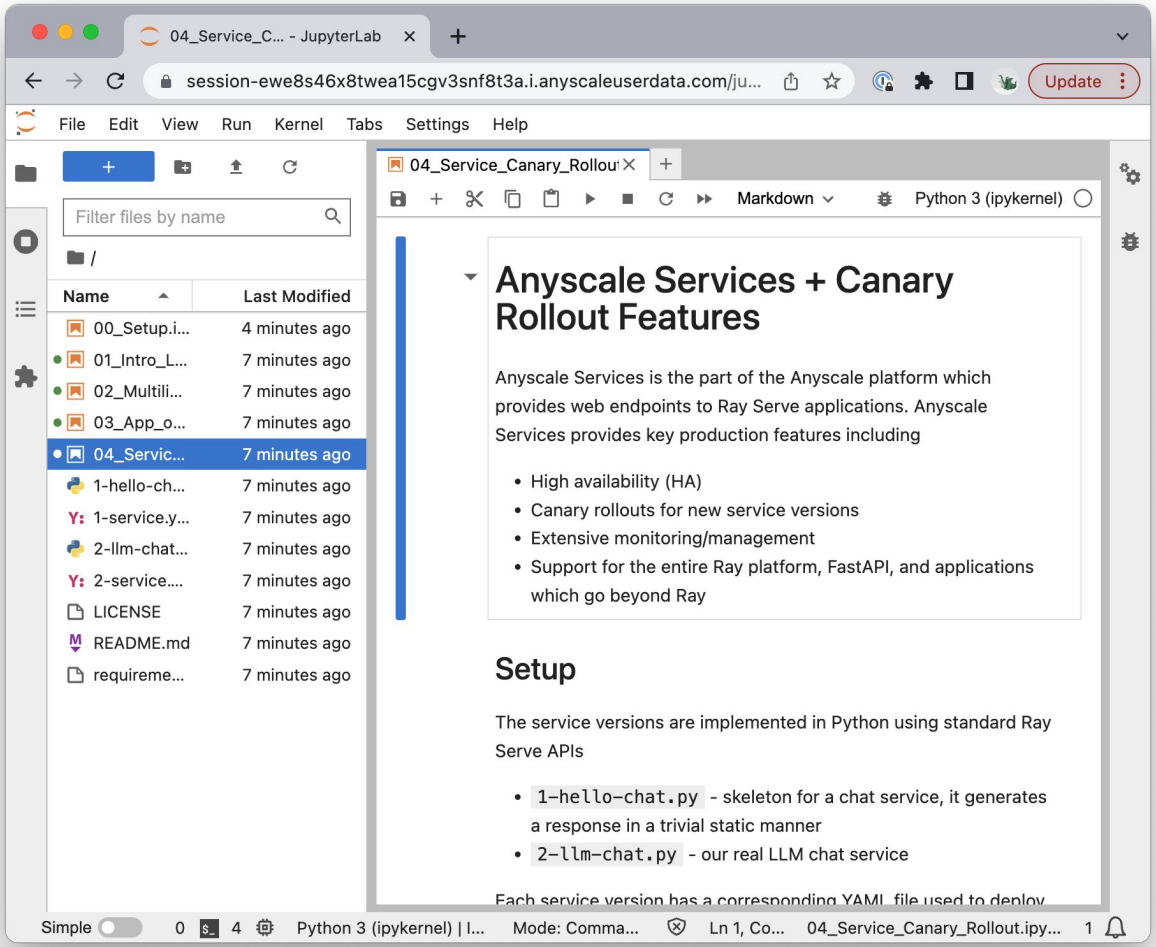
10 minutes.

LLMOps

Launching an LLM QA application in production



03_App_on_Ray_Serve.ipynb



04_Service_Canary_Rollout.ipynb

slido

Audience Q&A Session

ⓘ Click **Present with Slido** or install our Chrome extension to show live Q&A while presenting.

More Resources

For further exploration with Ray, Anyscale, and LLMs.

🍎 Today we learned...

🚂 Introduction to Ray and Anyscale

Why distributed compute is necessary for AI.

🔬 Ray Serve for scalable deployments

Building robust application services that scale.

🔍 LLMOps

Launching an LLM QA application in production.

Reading list.

Ray Education GitHub

Access bonus notebooks and scripts about Ray.

Ray documentation

API references and user guides.

Anyscale Blogs

Real world use cases and announcements.

YouTube Tutorials

Video walkthroughs about learning LLMs with Ray.

📅 Upcoming events

✋ Office Hours

June 21 at 4:00p.m. - follow-up for today's participants.

🌉 Ray Meetup

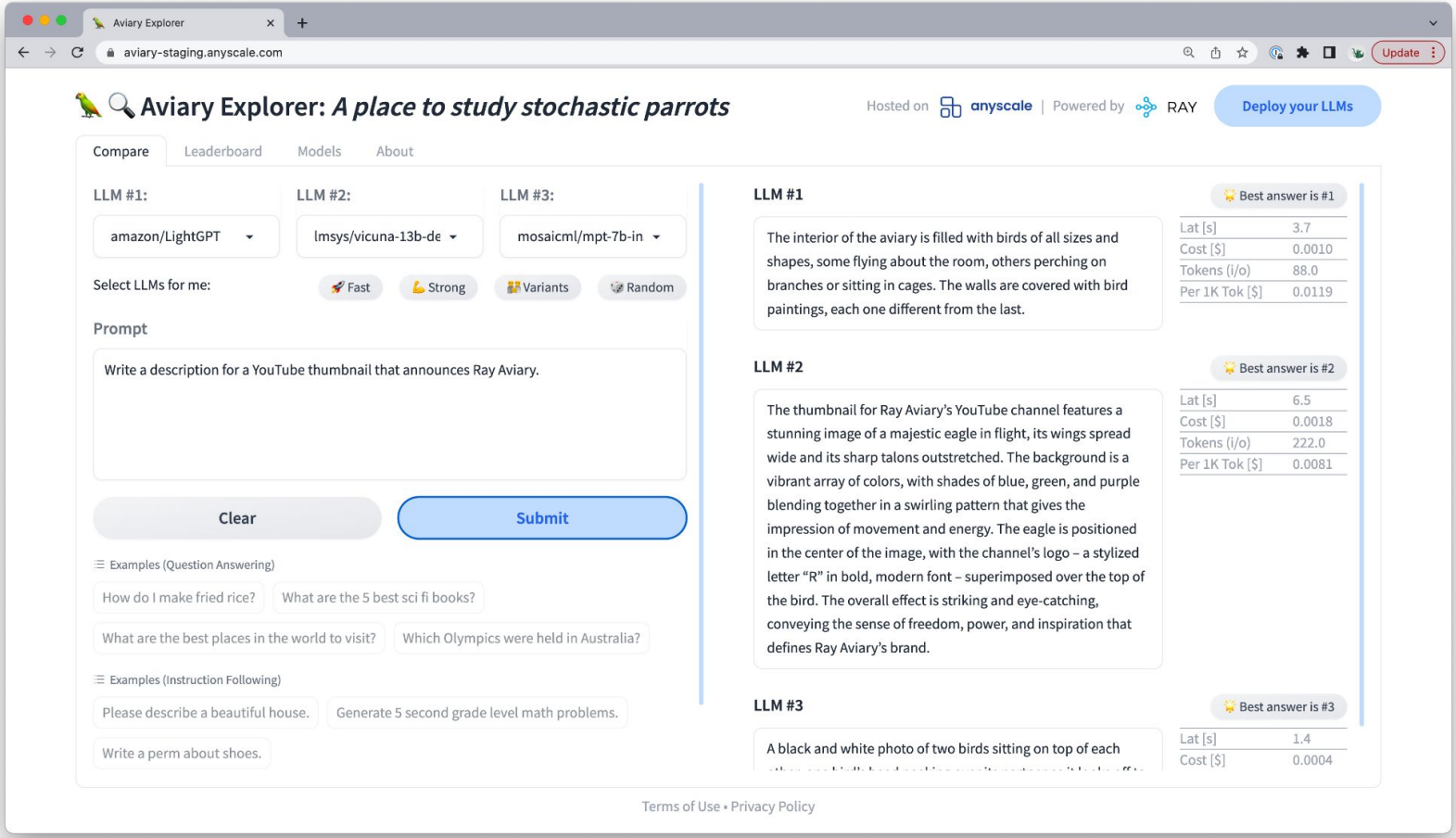
June 21 at 5:30p.m. ft. Ray data streaming & Pinterest ML

🦜 LangChain x Ray Meetup

June 22 at 5:30 ft. Robert Nishihara, Harrison Chase, Michel Tricot, Lianmin Zheng, & Charles Frye

Ray Aviary

Serving open source LLMs in production.



aviary.anyscale.com

RAY SUMMIT 23

THE PLACE FOR EVERYTHING RAY

SEPT. 18–19 + SEPT. 20 TRAINING DAY

SAN FRANCISCO, CA

presented by anyscale

Keynote Speakers for Ray Summit 2023

Albert Greenberg
VP Engineering
Uber

Brian McClendon
SVP Engineering
Niantic

Robert Nishihara
CEO
Anyscale

Ya Xu
Head of Data & AI
VP Engineering
LinkedIn

Ion Stoica
Co-Founder & President
Anyscale
Professor, U.C. Berkeley

Aidan Gomez
Co-founder and CEO
Cohere

John Schulman
Co-founder
OpenAI

raysummit.anyscale.com

Connect with the community.

👋 Join the community

Attend events, *subscribe to newsletter*, *follow on Twitter*.

🙋 Get support

Join Ray Slack, *ask questions on forum*, *open an issue*.

🔭 Contribute to Ray

Read contributor guide, *create a pull request*.

Fill out the survey.

🔗 Go to bit.ly/llms-feedback

We'll send all survey submitters a 20% discount code for Ray Summit tickets.

Thank you!

We hope to meet again.

slido

Audience Q&A Session

ⓘ Click **Present with Slido** or install our Chrome extension to show live Q&A while presenting.