aws marketplace

AWS DEV DAYS

Accelerate generative AI and LLM development with purpose-built ML developer tools

October 17, 2023
10 AM to 11:30 AM PT | 1 PM to 2:30 PM ET

In collaboration with

Weights & Biases

© 2023, Amazon Web Services, Inc. or its affiliates.

Agenda

- Challenges with developing generative AI
- Why customers choose AWS to build with generative AI?
- Features for LLM fine-tuning and prompt engineering
- Weights & Biases (W&B) integrations with AWS ML and AI services
- Better together: AWS and W&B
- Hands-on workshop
- Questions and answers

© 2023, Amazon Web Services, Inc. or its affiliates.

Today’s hosts:

James Yi
Senior Partner Solutions Architect in AI/ML, AWS

Thomas Capelle
Machine Learning Engineer, Weights & Biases

© 2023, Amazon Web Services, Inc. or its affiliates.

Challenges with developing generative AI

James Yi
Senior Partner Solutions Architect in AI/ML, AWS

aws marketplace

© 2023, Amazon Web Services, Inc. or its affiliates.

Challenges in developing generative AI

Data Diversity & Quality

Complex Model Training

Evaluation & Iteration

Interpretability & Control

aws marketplace

© 2023, Amazon Web Services, Inc. or its affiliates.

Why customers choose AWS to build with generative AI?

AWS and Amazon bring 20+ years of AI and ML experience

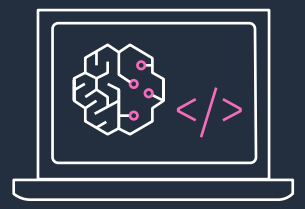
Easiest place to build with FMs

Most price-performant infrastructure for ML

Get started right away with game-changing generative applications such as Amazon CodeWhisperer

© 2023, Amazon Web Services, Inc. or its affiliates.

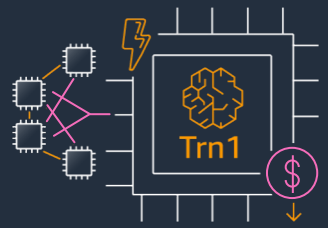
AWS generative AI services and infrastructure



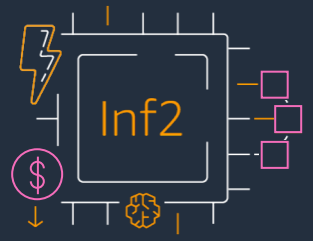
Amazon CodeWhisperer

Amazon Bedrock

Amazon SageMaker



AWS Trainium



AWS Inferentia2

© 2023, Amazon Web Services, Inc. or its affiliates.

Features for LLM fine-tuning and prompt engineering

Thomas Capelle
Machine Learning Engineer,
Weights & Biases

aws marketplace

© 2023, Amazon Web Services, Inc. or its affiliates.

Better together: AWS and W&B

- W&B for AWS is listed on AWS Marketplace, making it easy for customers to access and use

- W&B has achieved AWS Partner Machine Learning competency for technology partners.

- W&B focuses on differentiating the product in the AWS ecosystem and is a launch partner for the Amazon SageMaker Service Ready program.

© 2023, Amazon Web Services, Inc. or its affiliates.

W&B Developer-first MLOps Platform

SOFTWARE 1.0 - WRITE CODE

Design interface	Version code	Manage projects	Deploy to Infra	CI/CD	Application monitoring
Figma Sketch	GitHub GitLab	Jira VS Code	HashiCorp puppet	circleci Jenkins	PagerDuty DATADOG

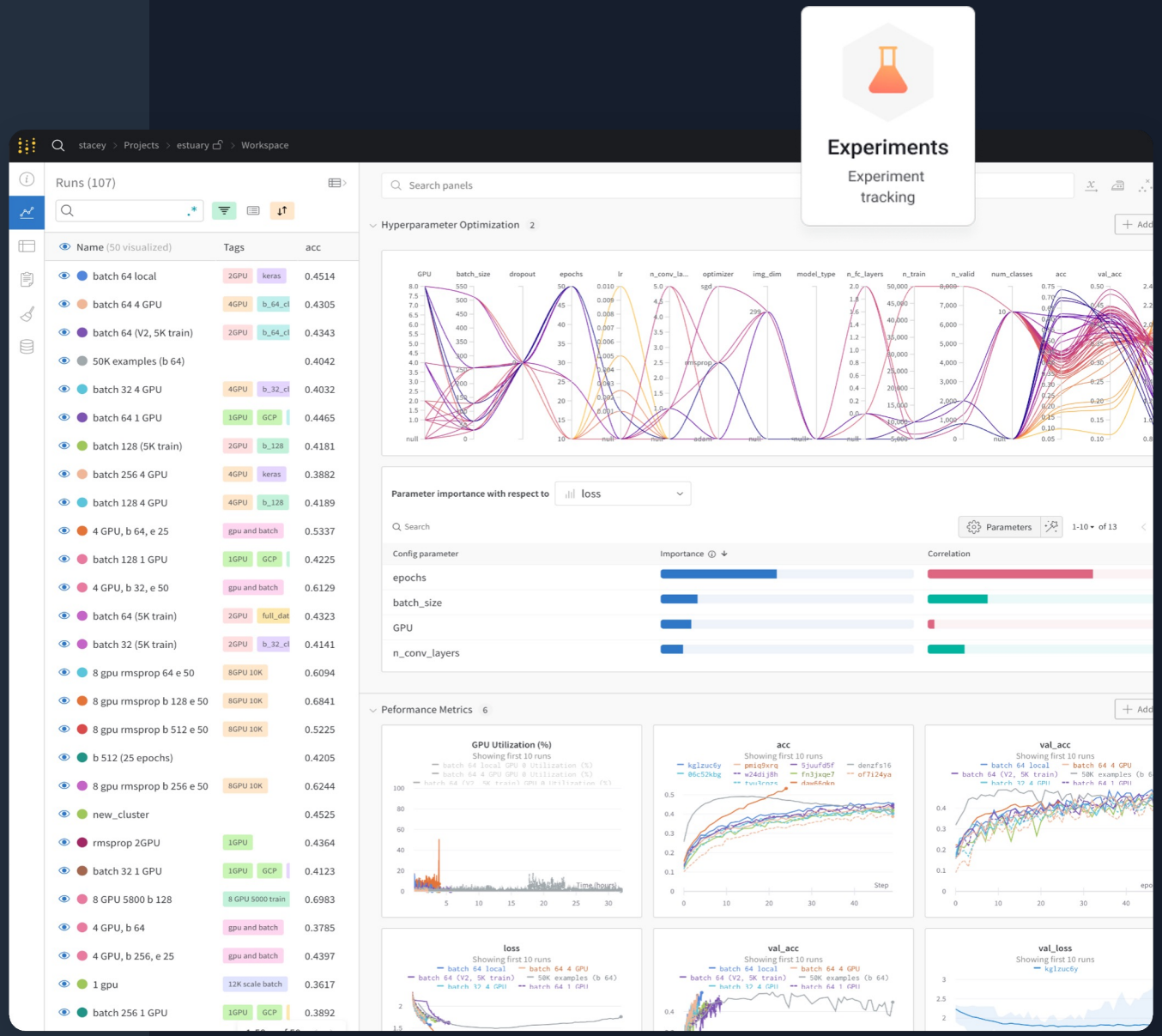
SOFTWARE 2.0 - TRAIN MODELS

Version data and pipelines	Prep and visualize data	Experiment tracking	Optimize models	Model registry	CI/CD
Artifacts	Tables	Experiments	Sweeps	Models	Automations & Monitoring

© 2023, Amazon Web Services, Inc. or its affiliates.

W&B models for building and fine-tuning models

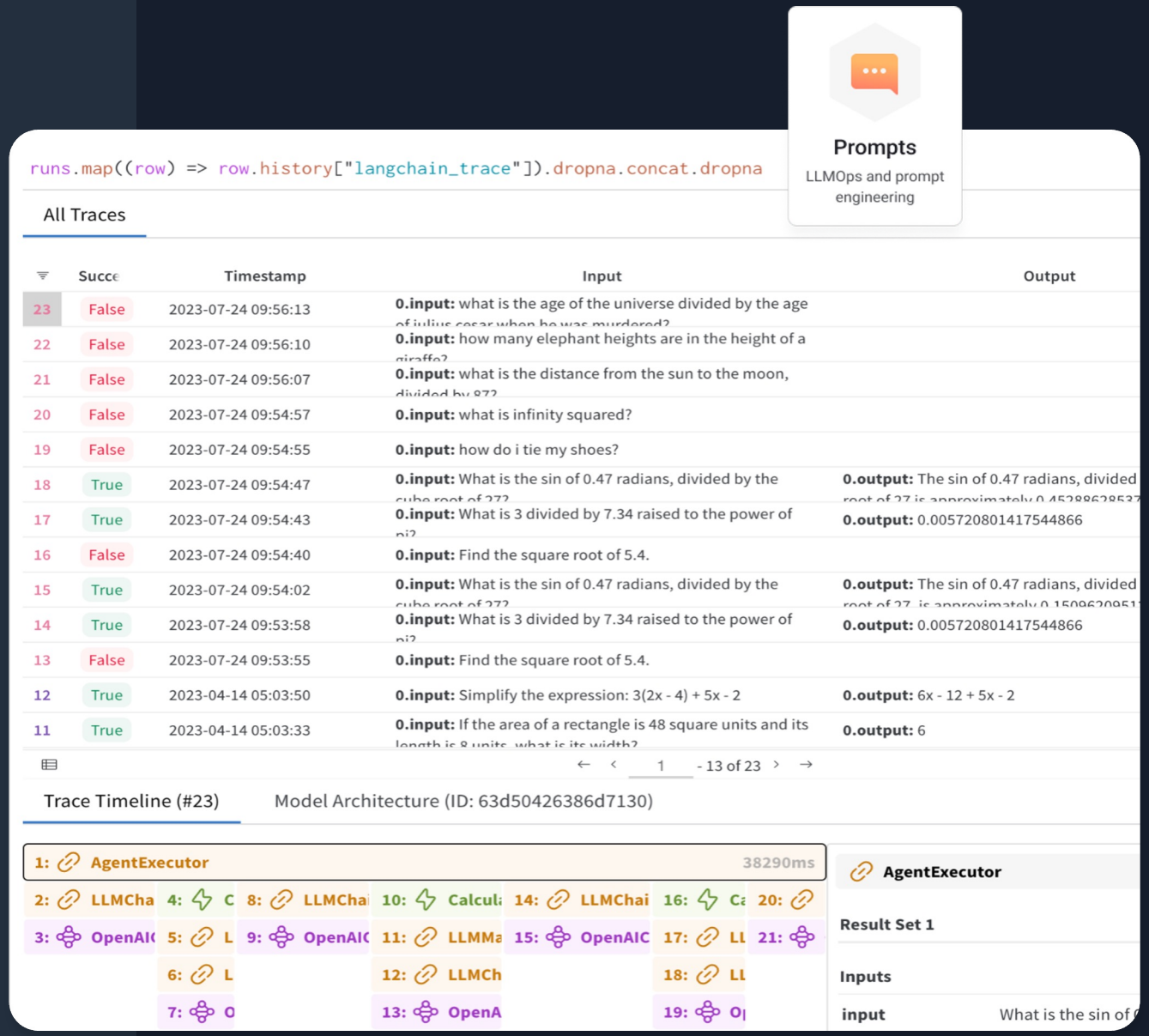
- Integrate into your training or fine-tuning workflows with **five lines of code**
- Version, prepare, analyze, and secure all your **artifacts** with access controls
- Track, measure, reproduce, audit, optimize, and improve every **experiment**
- Collaborate on **insights** and best practices to get better models to production faster
- Reduce cycle time to **launch** approved models to production inference environments



© 2023, Amazon Web Services, Inc. or its affiliates.

W&B prompts for building and monitoring LLM-based apps

- Track all inputs, outputs, and costs across any LLM using **API Gateway**
- Observability of LLM latency, token usage, and costs. Simplify **model selection** and trade-off analysis
- Debug LLM pipelines with detailed **evaluation traces** to engineer reliable and consistent outputs
- Live tracking of all evaluations across training and production with **drill downs** and customizable visualizations



aws marketplace

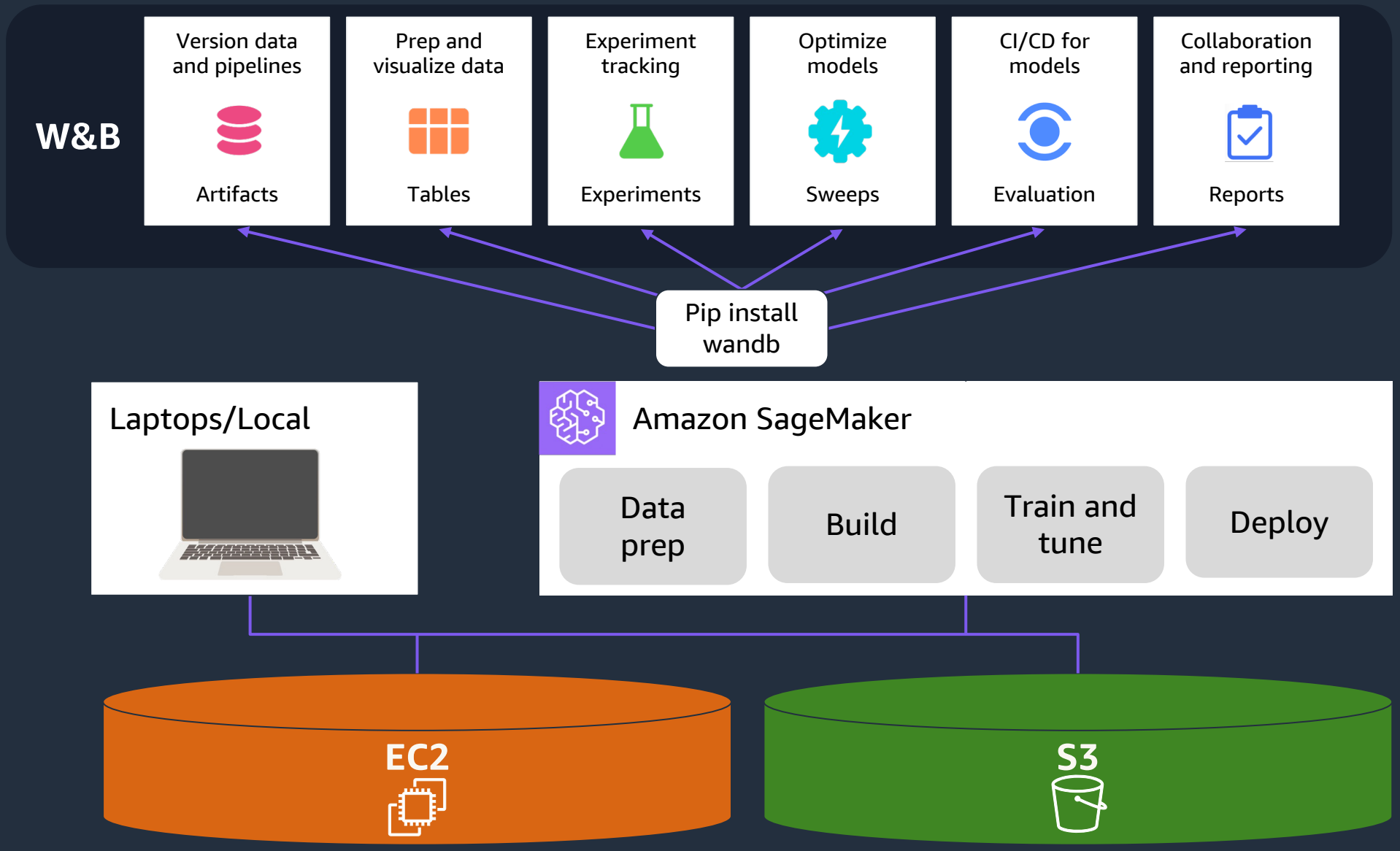
© 2023, Amazon Web Services, Inc. or its affiliates.

Weights & Biases and Amazon SageMaker

A system of record for ML development

W&B plugs into Amazon SageMaker

Infrastructure



aws marketplace

© 2023, Amazon Web Services, Inc. or its affiliates.

W&B for LLMs | ML Practitioner Types

1. **Model creators: train LLMs from scratch**
   - **Workflow:** Take OSS model e.g. GPT-NeoX → make architectural changes → pre-train → instruction tune → fine-tune / prompting and chaining → ongoing user/data feedback
   - **Example customer:** Cohere, Eleuther

2. **Model fine-tuners: take OSS LLMs and adapt to specific tasks**
   - **Workflow:** Take OSS model e.g. GPT-NeoX → instruction tune → fine-tune / prompting and chaining → ongoing user/data feedback
   - **Example customer:** Hedge fund / technical startups, in regulation / with sensitive data

3. **Model prompt engineers: use commercial LLM API services**
   - **Workflow:** Take commercial LLM API → prompting and chaining → optionally fine-tune the LLM using LLM API services once user data/feedback available
   - **Example customer:** Jasper, less technical teams and startups

© 2023, Amazon Web Services, Inc. or its affiliates.

We have *wandbot*: a LLM powered app answering questions on our Discord server!

Weights & Biases

8 Events

Browse Channels

INFORMATION

announcements

wandb-tips

rules

support

getting-started

GET INVOLVED

general

introduce-yourself

weave

wandbot

jobs-hiring

jobs-looking

日本語

EDUCATION

courses

VOICE CHANNELS

tcapelle Online

100+ new messages since 4:35 PM on August ... Mark As Read

Corey_S 08/26/2023 3:47 PM
@wandbot (beta) How can I use prompts for my LLM?

wandbot (beta) started a thread: Thread. See all threads. 08/26/2023 3:47 PM

Thread 4 Messages ›
Corey_S Document for reference https://docs.wandb.ai/gui

Corey_S 08/26/2023 3:50 PM
@wandbot (beta) what is WandbTracer?

wandbot (beta) started a thread: Thread. See all threads. 08/26/2023 3:50 PM

Thread 4 Messages ›
Corey_S Document for reference: https://docs.wandb.ai/re

@Zain Raza (he/him) Hello, sorry if this is not the cha

Corey_S 08/26/2023 3:58 PM
As another update here is the updated documentation on Weave and Production Monitoring @Zain Raza (he/him)

Weave: https://docs.wandb.ai/guides/weave

Weave, Production Monitoring: https://docs.wandb.ai/guides/weave/prod-mon

Message #wandbot

Thread

wandbot (beta) BOT 08/26/2023 3:47 PM
Hi @Corey_S: Please note that **wandbot is currently in alpha testing** and will experience frequent updates.

Please do not share any private or sensitive information in your query at this time.

Generating response...
You can use Weights & Biases Prompts for the development of LLM-powered applications. Here's a step-by-step guide on how to use W&B Prompts with LangChain:

1. Import WandbTracer from wandb.integration.langchain and optionally define a dictionary with arguments for wandb.init() that will later be passed to WandbTracer.

```
from wandb.integration.langchain import WandbTracer

wandb_config = {"project": "wandb_prompts_quickstart"}
```

2. Set up your LangChain Agent. Import an OpenAI model and create a math tool with load_tools. Next, create a math agent with the initialize_agent method and pass the tool and model objects to initialize_agent:

```
from langchain.llms import OpenAI
from langchain.agents import load_tools, initialize_agent,
AgentType

llm = OpenAI(temperature=0)
```

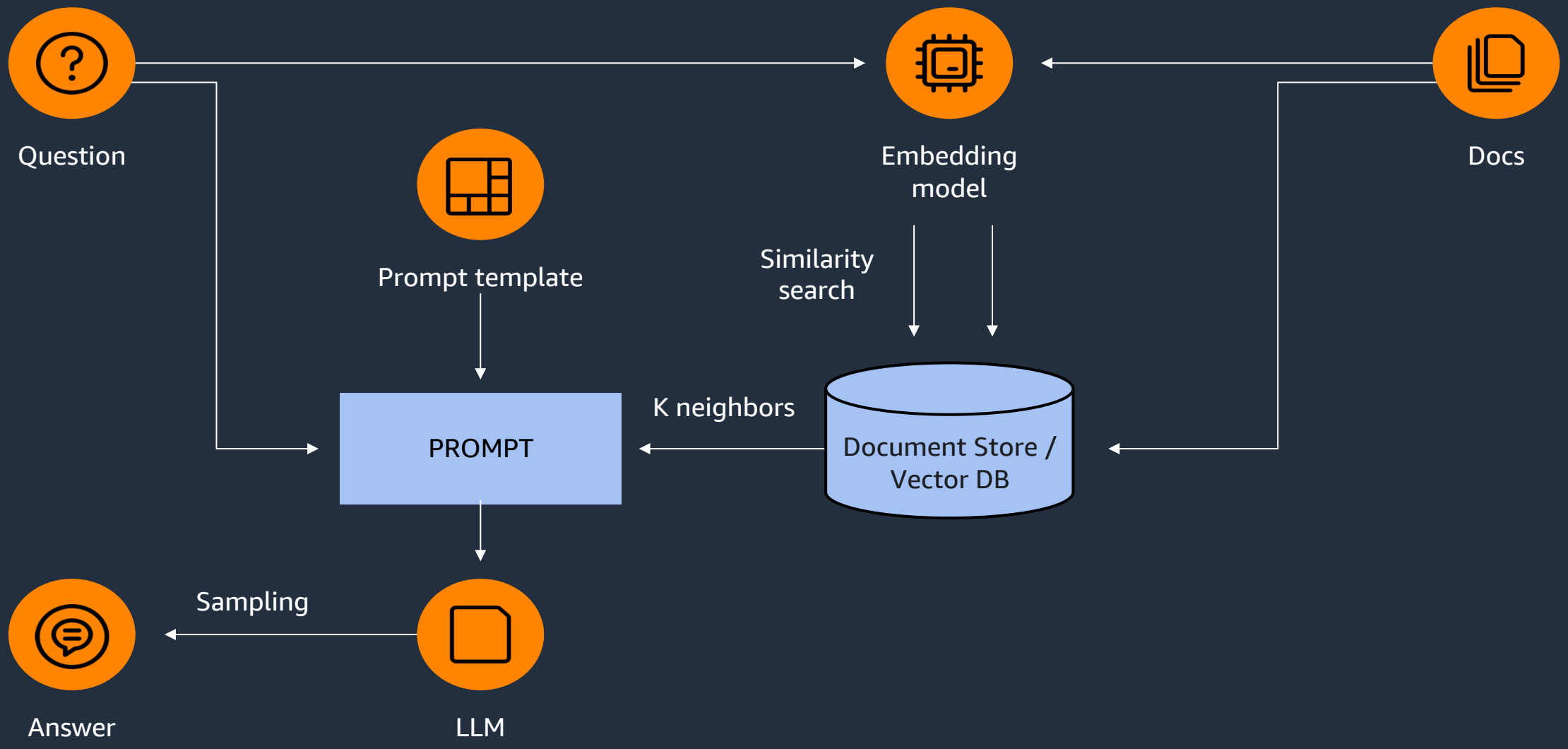
Message "Thread"

wandb.me/wandbot

aws marketplace

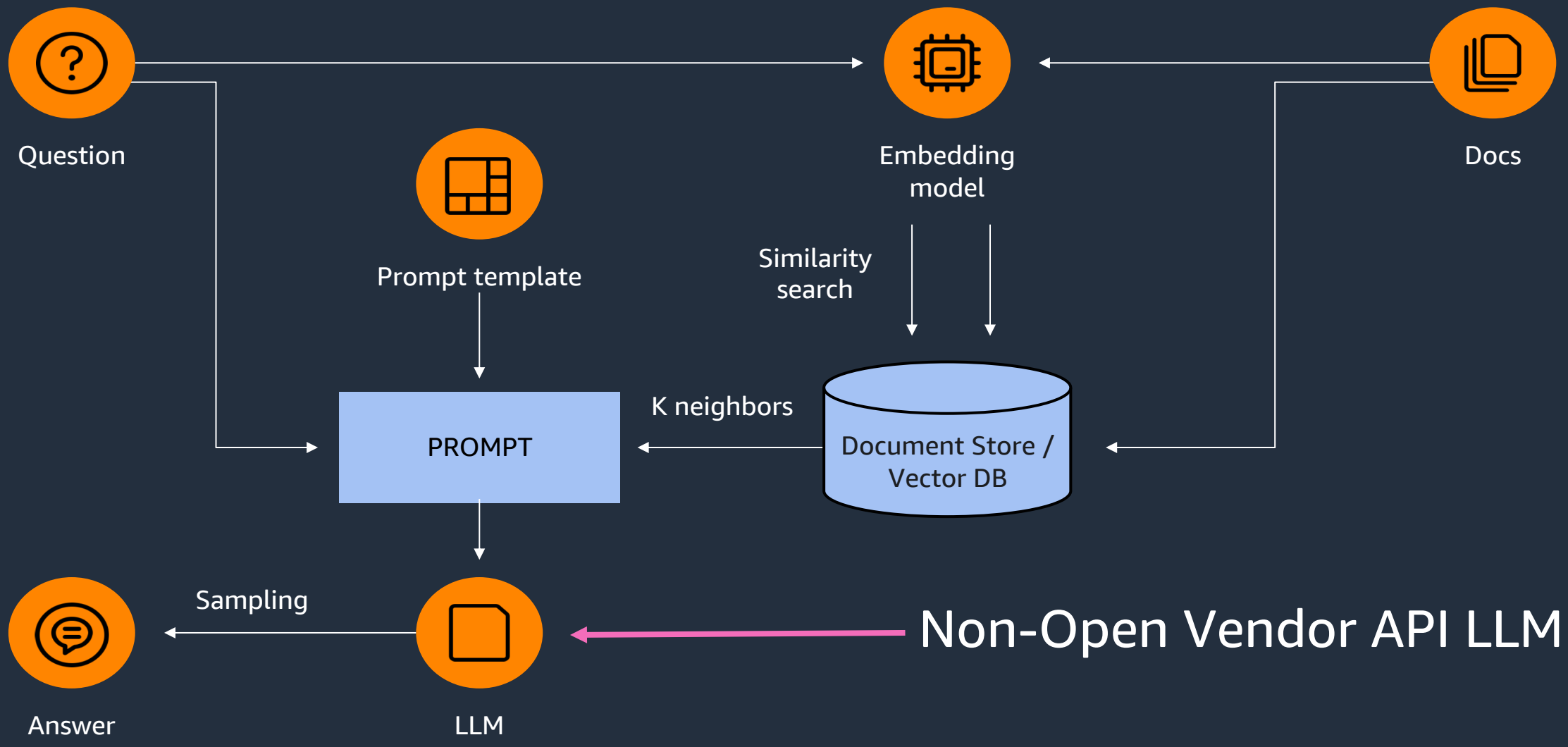
© 2023, Amazon Web Services, Inc. or its affiliates.

How does it work?



© 2023, Amazon Web Services, Inc. or its affiliates.

How does it work?



© 2023, Amazon Web Services, Inc. or its affiliates.

We have been keeping track of the generated questions

```
runs.summary["chat"]
```

	question	answer	model	source_documents	time_taken	start_time	end_time
8	I want to run a multi-gpu sweep with Pytorch lightning. What is the best way to run this	To run a multi-GPU sweep with PyTorch Lightning and Weights & Biases, you can follow	gpt-4	[{"source": "https://docs.wandb.ai/gu agents", "score": 0.4109834730625152	63.96	2023-07-14 13:08:55	2023-07-14 13:09:59
9	how do i update an alias of a model version inside the model registry using the wandb public api?	To update an alias of a model version inside the model registry using the wandb public api, you	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.3561517596244812	25.496	2023-07-14 13:03:13	2023-07-14 13:03:38
10	how do i update an alias of a model version inside the model registry programmatically?	You can update an alias of a model version inside the model registry programmatically by	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.4024070501327514	9.451	2023-07-14 12:54:17	2023-07-14 12:54:27
11	how do i update an alias of a model version inside the model registry programmatically?	You can update an alias of a model version inside the model registry programmatically by	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.4205527901649475	15.394	2023-07-14 12:40:09	2023-07-14 12:40:24
12	how do i update an alias of a model version inside the model registry programmatically?	Hmm, The question is not seem to be related to wandb. As a documentation bot for	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.4096565246582031	7.4	2023-07-14 12:35:49	2023-07-14 12:35:56
13	how do i update an alias of a model version inside the model registry programmatically?	To update an alias of a model version inside the model registry	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.4096565246582031	9.881	2023-07-14 12:27:41	2023-07-14 12:27:51
14	how do i update an alias of a model version inside the model registry programmatically?	You can update an alias of a model version inside the model registry	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.4204981923103325	10.512	2023-07-14 12:25:11	2023-07-14 12:25:21

8 - 14 of 29

Export as CSV Columns... Reset Table

© 2023, Amazon Web Services, Inc. or its affiliates.

We have been keeping track of the generated questions

```
runs.summary["chat"]
```

	question	answer	model	source_documents	time_taken	start_time	end_time
8	I want to run a multi-gpu sweep with Pytorch lightning. What is the best way to run this	To run a multi-GPU sweep with PyTorch Lightning and Weights & Biases, you can follow	gpt-4	[{"source": "https://docs.wandb.ai/gu agents", "score": 0.4109834730625152	63.96	2023-07-14 13:08:55	2023-07-14 13:09:59
9	how do i update an alias of a model version inside the model registry using the wandb public api?	To update an alias of a model version inside the model registry using the wandb public api, you	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.3561517596244812	25.496	2023-07-14 13:03:13	2023-07-14 13:03:38
10	how do i update an alias of a model version inside the model registry programmatically?	You can update an alias of a model version inside the model registry programmatically by	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.4024070501327514	9.451	2023-07-14 12:54:17	2023-07-14 12:54:27
11	how do i update an alias of a model version inside the model registry programmatically?	You can update an alias of a model version inside the model registry programmatically by	gpt-4	[{"source": "https://docs.wandb.ai/gu "score": 0.4205527901649475	15.394	2023-07-14 12:40:09	2023-07-14 12:40:24
12	how do i update an alias of a model version inside the model registry programmatically?	Hmm, The question is not seem to be related to wandb. As a documentation bot for	gpt-4				
13	how do i update an alias of a model version inside the model registry programmatically?	To update an alias of a model version inside the model registry	gpt-4				
14	how do i update an alias of a model version inside the model registry programmatically?	You can update an alias of a model version inside the model registry	gpt-4				

For more details, you can refer to the W&B Prompts documentation.

🤖 If you still need help please try re-phrase your question, or alternatively reach out to the Weights & Biases Support Team at support@wandb.com

Was this response helpful? Please react below to let us know

👍 2 👎 1

and the quality

© 2023, Amazon Web Services, Inc. or its affiliates.

We can use this data to train our own LLM now

- Using this dataset, we will fine-tune a Llama2 model in SageMaker
- We will evaluate the performance of the fine-tuned model
- This new model replaces the vendor's model API
- We can serve this model using SageMaker Inference

© 2023, Amazon Web Services, Inc. or its affiliates.

Hands-on workshop

James Yi
Senior Partner Solutions Architect in AI/ML, AWS

Thomas Capelle
Machine Learning Engineer, Weights & Biases

aws marketplace

© 2023, Amazon Web Services, Inc. or its affiliates.

Questions and answers

© 2023, Amazon Web Services, Inc. or its affiliates.

Conclusion/next steps

- Thank you for attending.
- Please fill out the AWS event survey and let us know what you thought.
- If you have any further questions or wish to schedule a call with AWS and Weights & Biases experts, please contact: wb-aws-contact-us@wandb.com

© 2023, Amazon Web Services, Inc. or its affiliates.

aws marketplace

Thank you!

In collaboration with

Weights & Biases

© 2023, Amazon Web Services, Inc. or its affiliates.