
MLflow


AI

Gateway

(Experimental)


Stack


LLMOps

Building


an

28

MLflow AI Gateway

localhost:5001/docs#/

MLflow AI Gateway 2.17.2 OAS 3.1

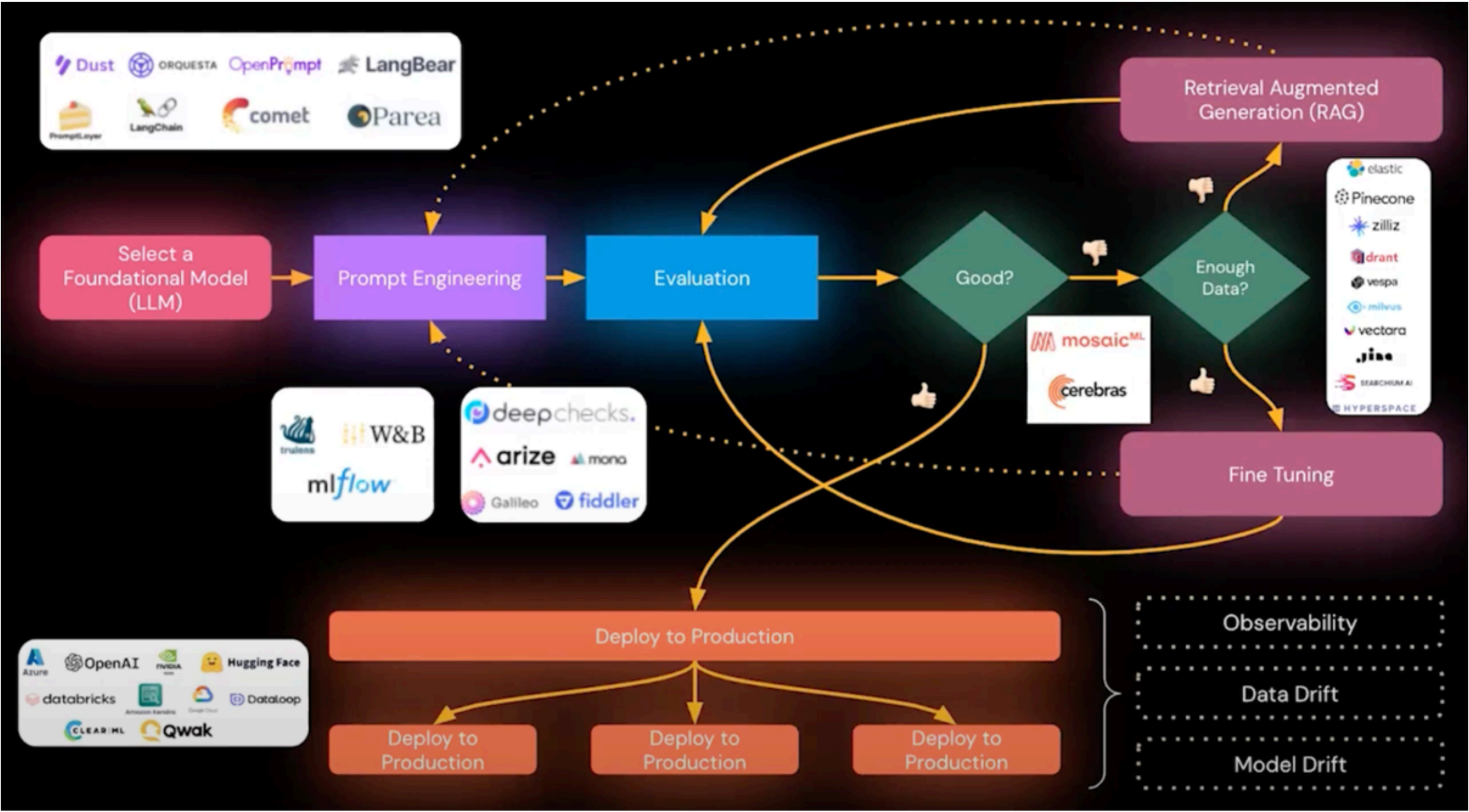
/openapi.json

The core deployments API for reverse proxy interface using remote inference endpoints within MLflow

default

- POST `/endpoints/completions/invocations` Completions Completions Deployments
- POST `/endpoints/chat/invocations` Chat
- GET `/health` Health
- GET `/api/2.0/endpoints/{endpoint_name}` Get Endpoint
- GET `/api/2.0/endpoints/` List Endpoints
- GET `/api/2.0/endpoints/limits/{endpoint}` Get Limits
- POST `/api/2.0/endpoints/limits/` Set Limits
- POST `/v1/chat/completions` Openai Chat Handler
- POST `/v1/completions` Openai Completions Handler
- POST `/v1/embeddings` Openai Embeddings Handler

Schemas


Dust
ORQUESTA
OpenPrompt
LangBear
PromptLayer
LangChain
comet
Parea
Retrieval Augmented Generation (RAG)
elastic
Pinecone
zilliz
drant
vespa
milvus
vectara
jina
SEARCHIUM AI
HYPERSPACE
Select a Foundational Model (LLM)
Prompt Engineering
Evaluation
Good?
Enough Data?
mosaicML
cerebras
truleens
W&B
mlflow
deepchecks.
arize
mona
Galileo
fiddler
Fine Tuning
Azure
OpenAI
NVIDIA
Hugging Face
databricks
Amazon Kendra
Google Cloud
Dataloop
CLEARML
Qwak
Deploy to Production
Deploy to Production
Deploy to Production
Deploy to Production
Observability
Data Drift
Model Drift

29