MLOps Lab3 Assignment

Task Summary

Add MLFlow tracking for the training service.

1. To do that, add MLFlow backend (you can use your server or Databricks or any other system you want)
2. Save training artefacts to the object storage
3. Make model service scalable
4. Use challenger-champion scheme for deployment

Note: all artefacts for lab results demo are located in the `MLOps/reranker_mlflow_demo` folder
Note (one more): the whole code from the first version `MLOps/reranker_pipeline_demo` was reconsidered, refactored, and simplified for the sake of improvements of interpretability and further cloud deployment

Local Deployment

As mentioned earlier, the final version of this task implementation was simplified for better understanding of the next stage, so for MLFlow and challenger-champion pipeline tests we only lift the necessary resources:

- FastAPI: http://localhost:8000
- MLflow UI: http://localhost:5000
- MinIO: http://localhost:9001

The setup includes four core services:

- `mlflow-server` handles experiment tracking and model registry
- `rag-api` exposes FastAPI-based endpoint for challenger training and evaluation
- `minio` acts as an S3-compatible object storage backend
- `postgres` acts as the backend store for MLflow's tracking metadata

Both `mlflow-server` and `rag-api` are built from custom Dockerfiles, share volumes for model artifacts (`mlruns/`) and communicate via Docker's internal network.

```
docker compose up --build
```

```
[+] Building 4.6s (21/21) FINISHED                                                                                  docker:desktop-linux
 => [mlflow internal] load build definition from Dockerfile.mlflow                                                                    0.0s
 => => transferring dockerfile: 149B                                                                                                  0.0s
 => [mlflow internal] load metadata for ghcr.io/mlflow/mlflow:v2.12.1                                                                 0.0s
 => [mlflow internal] load .dockerignore                                                                                              0.0s
 => => transferring context: 2B                                                                                                       0.0s
 => [mlflow 1/2] FROM ghcr.io/mlflow/mlflow:v2.12.1@sha256:00fd66fe10c93315eb6a5e98f872bc237aed03b5010cdc145b82b01ab1a4cf92            0.8s
 => => resolve ghcr.io/mlflow/mlflow:v2.12.1@sha256:00fd66fe10c93315eb6a5e98f872bc237aed03b5010cdc145b82b01ab1a4cf92                    0.7s
 => [mlflow 2/2] RUN pip install psycopg2-binary                                                                                      1.8s
 => [mlflow] exporting to image                                                                                                       0.4s
 => => exporting layers                                                                                                               0.3s
 => => exporting manifest sha256:1b2bbdffdf00b7c363819ee4e3e3760a1394fadbadc6160ce1e4aa167e48b73b                                     0.0s
 => => exporting config sha256:0f5a881b7a7f65842b2ff43cc7f75d4e90adb97da59b89c8da8bb3474396c346                                      0.0s
 => => exporting attestation manifest sha256:cacf3ae950f509f36b848efc76e4b1351af5e72b012fa8f506651552e77caca1                         0.0s
 => => exporting manifest list sha256:0e605c38ad6e9f7e734e33ff5f188541d7a7bee04d0df6d807f0d5cec4328744                                0.0s
 => => naming to docker.io/library/reranker_mlflow_demo-mlflow:latest                                                                 0.0s
 => => unpacking to docker.io/library/reranker_mlflow_demo-mlflow:latest                                                              0.1s
 => [mlflow] resolving provenance for metadata file                                                                                   0.0s
 => [reranker-api internal] load build definition from Dockerfile                                                                     0.0s
 => => transferring dockerfile: 540B                                                                                                  0.0s
 => [reranker-api internal] load metadata for docker.io/library/python:3.11-slim                                                      1.0s
 => [reranker-api auth] library/python:pull token for registry-1.docker.io                                                            0.0s
 => [reranker-api internal] load .dockerignore                                                                                        0.0s
 => => transferring context: 2B                                                                                                       0.0s
 => [reranker-api 1/7] FROM docker.io/library/python:3.11-slim@sha256:139020233cc412efe4c8135b0efe1c7569dc8b28ddd88bddb109b764f8977e30  0.0s
 => => resolve docker.io/library/python:3.11-slim@sha256:139020233cc412efe4c8135b0efe1c7569dc8b28ddd88bddb109b764f8977e30              0.0s
 => [reranker-api internal] load build context                                                                                        0.0s
 => => transferring context: 2.77kB                                                                                                   0.0s
 => CACHED [reranker-api 2/7] WORKDIR /app                                                                                            0.0s
 => CACHED [reranker-api 3/7] RUN apt-get update && apt-get install -y     gcc     g++     curl     && rm -rf /var/lib/apt/lists/*   0.0s
 => CACHED [reranker-api 4/7] COPY requirements.txt .                                                                                 0.0s
 => CACHED [reranker-api 5/7] RUN pip install --no-cache-dir -r requirements.txt                                                      0.0s
 => [reranker-api 6/7] COPY . .                                                                                                       0.0s
 => [reranker-api 7/7] RUN mkdir -p models/reranker                                                                                   0.1s
 => [reranker-api] exporting to image                                                                                                 0.1s
 => => exporting layers                                                                                                               0.0s
 => => exporting manifest sha256:20f7a372f71c3bab17770e0e78a846f0b75c371f579e7b77b6b61845d68d2fcd                                     0.0s
 => => exporting config sha256:f371a6abae35622f576641accf68fcfb38e1bad014675460b0ba5690e56de02d                                       0.0s
 => => exporting attestation manifest sha256:40c4ec1357290e41441357608583378 01c50dd761ae8dc8941298239fd17d01                         0.0s
 => => exporting manifest list sha256:70da0a118419117a6114124b2afb6bf46ba494f437d5c42f154680d506cd1613                                0.0s
 => => naming to docker.io/library/reranker_mlflow_demo-reranker-api:latest                                                           0.0s
 => => unpacking to docker.io/library/reranker_mlflow_demo-reranker-api:latest                                                        0.0s
 => [reranker-api] resolving provenance for metadata file                                                                             0.0s
[+] Running 5/5
 ✔ Network reranker_mlflow_demo_default           Created                                                                             0.0s
 ✔ Container reranker_mlflow_demo-postgres-1      Created                                                                             0.1s
 ✔ Container reranker_mlflow_demo-minio-1         Created                                                                             0.1s
 ✔ Container reranker_mlflow_demo-mlflow-1        Created                                                                             0.1s
 ✔ Container reranker_mlflow_demo-reranker-api-1  Created                                                                             0.1s
Attaching to minio-1, mlflow-1, postgres-1, reranker-api-1
postgres-1      |
postgres-1      | PostgreSQL Database directory appears to contain a database; Skipping initialization
postgres-1      |
postgres-1      | 2025-07-20 13:53:41.371 UTC [1] LOG:  starting PostgreSQL 13.21 (Debian 13.21-1.pgdg120+1) on aarch64-unknown-linux-gnu, compiled by gcc (D
ebian 12.2.0-14) 12.2.0, 64-bit
postgres-1      | 2025-07-20 13:53:41.371 UTC [1] LOG:  listening on IPv4 address "0.0.0.0", port 5432
postgres-1      | 2025-07-20 13:53:41.371 UTC [1] LOG:  listening on IPv6 address "::", port 5432
postgres-1      | 2025-07-20 13:53:41.375 UTC [1] LOG:  listening on Unix socket "/var/run/postgresql/.s.PGSQL.5432"
postgres-1      | 2025-07-20 13:53:41.392 UTC [27] LOG:  database system was shut down at 2025-07-20 13:49:48 UTC
postgres-1      | 2025-07-20 13:53:41.397 UTC [1] LOG:  database system is ready to accept connections
minio-1         | MinIO Object Storage Server
minio-1         | Copyright: 2015-2025 MinIO, Inc.
minio-1         | License: GNU AGPLv3 - https://www.gnu.org/licenses/agpl-3.0.html
minio-1         | Version: RELEASE.2025-06-13T11-33-47Z (go1.24.4 linux/arm64)
minio-1         |
minio-1         | API: http://172.19.0.3:9000  http://127.0.0.1:9000
minio-1         | WebUI: http://172.19.0.3:9001 http://127.0.0.1:9001
```

reranker_mlflow_demo	-	-	-		0.68%	53 seconds ago	
minio-1	442366c31132	minio/minio:latest	9000:9000 Show all ports (2)		0%	53 seconds ago	
postgres-1	bcc9e5f1dd85	postgres:13			0.13%	53 seconds ago	
mlflow-1	d4fc7aa783d6	reranker_mlflow_demo-mlflow	5000:5000		0.04%	53 seconds ago	
reranker-api-1	5ea190e0c5a0	reranker_mlflow_demo-reranker-api	8000:8000		0.51%	53 seconds ago	

The full logic of training a challenger, comparing two models, and deploying them was moved to the `POST /train_challenger` method of the main model API for convenience.

localhost:8000/docs#/

Document Reranker API 2.0.0 OAS 3.1

/openapi.json

default

GET	/	Root
GET	/health	Health Check
POST	/rerank	Rerank Documents
POST	/score	Score Pairs
POST	/train_challenger/	Train Challenger

Register Base Model

In order to be able to compare the trained model in the future, we first register and promote the basic `CrossEncoder("cross-encoder/ms-marco-MiniLM-L-6-v2")` in a `pyfunc` wrapper to the production version enabling the loading and inference logic to be fully encapsulated:

```
class CrossEncoderWrapper(mlflow.pyfunc.PythonModel):
    def load_context(self, context):
        from sentence_transformers import CrossEncoder
        self.model =
CrossEncoder(os.path.join(context.artifacts["model_path"]))

    def predict(self, context, model_input):
        return self.model.predict(model_input)
```

In the future, we will also continue to use this wrapper to properly log the model and easily retrieve it from the production phase.

```
python pipeline/register_base_model.py
```

The newly logged model was automatically registered under the specified name and transitioned to the "Production" stage.

Register base CrossEncoder

Overview | Model metrics | System metrics | Traces | Artifacts

Description

No description

Details

Created at	07/20/2025, 10:16:36 PM
Created by	anastasiiamazur
Experiment ID	0
Status	Finished
Run ID	bfd0bcabc2cb44c0992eca3ece50945c
Duration	12.9s
Datasets used	—
Tags	Add tags
Source	register.py d8bcdf5
Registered models	CrossEncoderReranker v2
Registered prompts	—

The entire model logging and registration process is traceable within the MLflow UI, including parameters, source run ID, and artifacts directory.

Register base CrossEncoder

Overview | Model metrics | System metrics | Traces | Artifacts

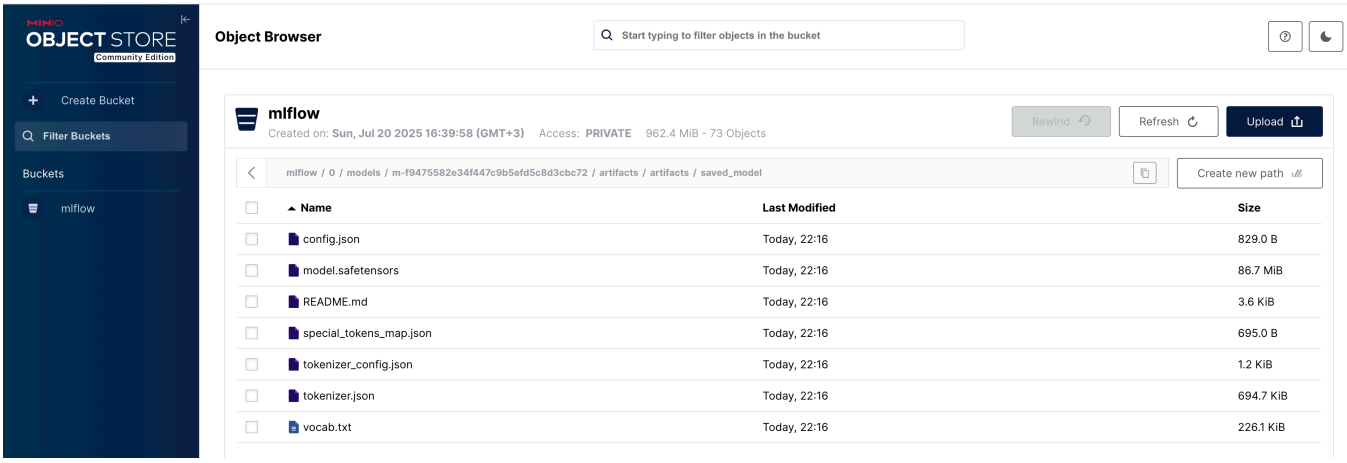
- artifacts
  - saved_model
    - README.md
    - config.json
    - model.safetensors
    - special_tokens_map.json
    - tokenizer.json
    - tokenizer_config.json
    - vocab.txt
  - MLmodel
  - conda.yaml
  - python_env.yaml
  - python_model.pkl
  - requirements.txt

You're viewing artifacts assigned to a logged model associated with this run.

artifacts/saved_model

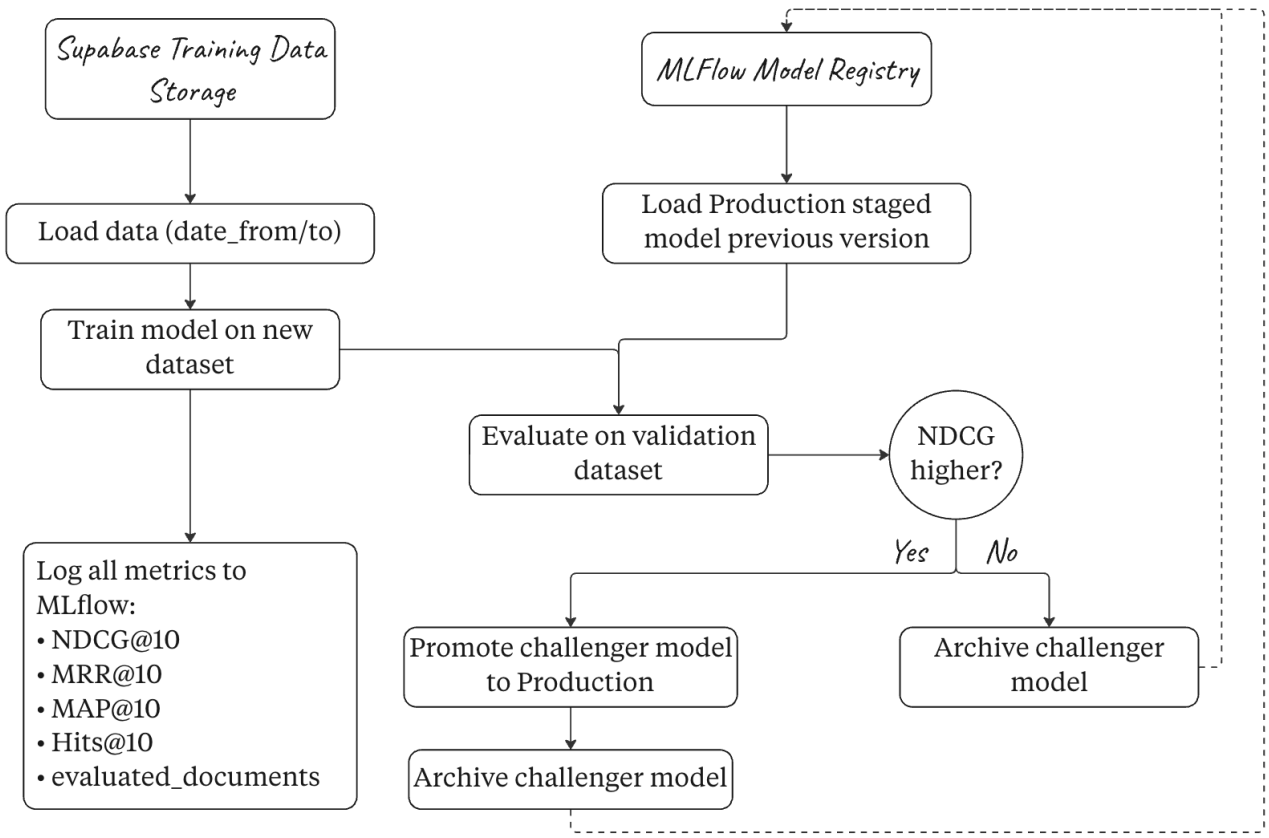
Path: s3://mlflow/0/bfd0bcabc2cb44c0992eca3ece50945c/artifacts/artifacts/saved_model

The same artifacts can be directly observed in the MiniO object storage.



Champion–Challenger Pipeline

If the challenger outperforms the champion, it is promoted to the "Production" stage, and the previous model version is archived.



Every model is being registered under a new Version, so that single consistent model name is maintained. This ensures that any system referencing the production model can always rely on

a stable name.

Registered Models ›

CrossEncoderReranker

Created Time: 07/20/2025, 04:56:31 PM Last Modified: 07/20/2025, 10:46:21 PM

› Description Edit

› Tags

⌄ Versions Compare

New model registry UI

Version	Registered at	Created by	Tags	Aliases	Description
Version 5	07/20/2025, 10:46:21 PM		Add	Add	
Version 4	07/20/2025, 10:39:36 PM		Add	Add	
Version 3	07/20/2025, 10:24:04 PM		Add	Add	
Version 2	07/20/2025, 10:16:49 PM		Add	Add	
Version 1	07/20/2025, 05:57:44 PM		Add	Add	

Thus, we can always refer to any of the model versions, but in the Production phase, only one is found, which is used for queries and uploaded for comparison with the challenger:

Registered Models › CrossEncoderReranker ›

Version 5

Registered At: 07/20/2025, 10:46:21 PM Last Modified: 07/20/2025, 10:46:21 PM Source Run: challenger-ms-marco-MiniLM-L-6-v2-2025-07-01_to_2025-07-14

Stage (deprecated): Production

And all others remain archived:

Registered Models › CrossEncoderReranker ›

Version 2

Registered At: 07/20/2025, 10:16:49 PM Last Modified: 07/20/2025, 10:24:05 PM Source Run: Register base CrossEncoder

Stage (deprecated): Archived

To run the training in this case, we simply make a request to the custom Fast API:

```
url = "http://localhost:8000/train_challenger/"
payload = {
    "date_from": "2025-07-01",
    "date_to": "2025-07-20"
}
response = requests.post(url, json=payload)
```

List of experiments:

Experiments

Search experiments

- Default
- ChallengerTraining

Displaying Runs from 2 Experiments Share

Runs Evaluation Traces

metrics.rmse < 1 and params.model = "tree" Time created State: Active Datasets Sort: Created Columns Group by

Run Name	Created	Dataset	Duration	Source	Models
challenger-ms-marco-...	5 minutes ago	-	3.4min	register....	CrossEncoderReran... +1
challenger-ms-marco-...	8 minutes ago	-	13.0s	register....	CrossEncoderReran... +1
Register base CrossEnc...	24 minutes ago	-	8.0s	register....	model
Register base CrossEnc...	31 minutes ago	-	12.9s	register....	model

Show more columns (16 total)

Logs from the last experiment with a full training session:

mlflow 3.1.1 Experiments Models Prompts

ChallengerTraining ›

challenger-ms-marco-MiniLM-L-6-v2-2025-07-01_to_2025-07-14

Overview | Model metrics | System metrics | Traces | Artifacts

Description

No description

Details

Created at	07/20/2025, 10:42:56 PM
Created by	anastasiiamazur
Experiment ID	1
Status	Finished
Run ID	e916a7bfce2241d0bc74c2424763443e
Duration	3.4min
Datasets used	—
Tags	Add tags
Source	register.py d8bcdf5
Registered models	CrossEncoderReranker v5
Registered prompts	—

ChallengerTraining ›

challenger-ms-marco-MiniLM-L-6-v2-2025-07-01_to_2025-07-14

Overview | Model metrics | System metrics | Traces | Artifacts

Metrics (5)

Search metrics

Metric	Value
ndcg_at_5	0.61
mrr_at_5	0.58
map_at_5	0.53
hits_at_5	0.77
evaluated_documents	2541

Parameters (7)

Search parameters

Parameter	Value
base_model	cross-encoder/ms-marco-MiniLM-L-6-v2
epochs	3
batch_size	16
num_negatives	10
date_from	2025-07-01
date_to	2025-07-14
archived_model_version	4

Logged models (1)

Model attributes							No dataset			
Type	Step	Model name	Status	Created	Registered models	Dataset	ndcg_at_5	mrr_at_5	map_at_5	hits_
Output	0	model_pyfunc	Ready	2 minutes ago	CrossEncoderReranker v5	-	0.61	0.58	0.53	0.77