Karlsruhe Institute of Technology

Scientific Computing Centre (SCC)
D3A - Data Analytics, Access and Applications

MLflow Pilot Service for Helmholtz Researchers

Lisana Berberi*, Christophe Laures, Khadijeh Alibabaei, Valentin Kozlov, Achim Streit

Karlsruhe Institute of Technology (KIT), Germany
**lisana.berberi@kit.edu*

Problem Statement & Motivation

Key challenges for researchers:

- ❓ How do I **reproduce** an AI/ML experiment I ran weeks ago with the exact same parameters, data, and code?
- ❓ How do I **compare** dozens of AI/ML model runs across different hyperparameters, datasets, and metrics?
- ❓ How do I **track which version** of a dataset was used to train which version of an AI/ML model?
- ❓ How do I **manage the lifecycle** of AI/ML models from development to staging to production?
- ❓ How do I **evaluate** and **monitor** GenAI/LLM applications for quality and hallucination?

MLflow is an **open-source** platform [2] that assists AI/ML **practitioners** and teams in handling the complete AI/ML lifecycle, ensuring that each stage is *manageable*, *traceable*, and *reproducible*.

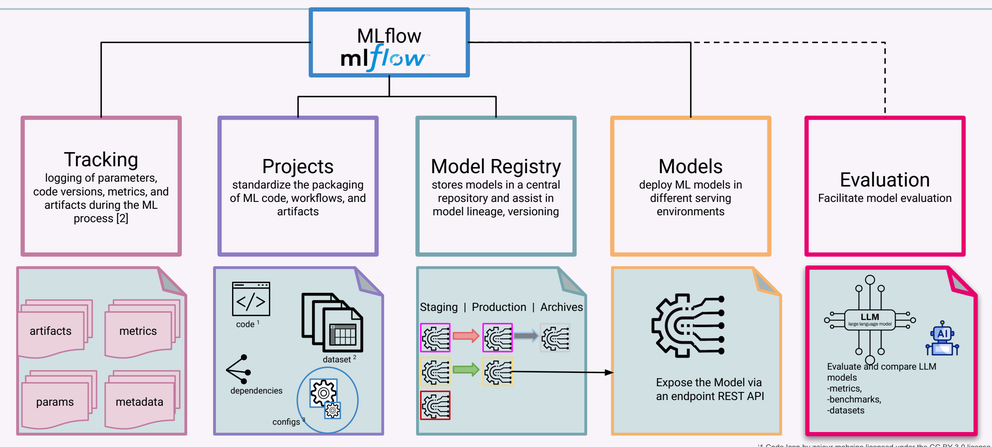


Fig 1: MLflow Components

[1] Code Icon by zaiour mohcine licensed under the CC BY 3.0 license
[2] Data set icon by H Alberto Gongora under the license CC BY 3.0 license
[3] Config icon by Madalin Jefferson under the license CC BY 3.0 license
MLFlow logo by MLflow Project, a Series of LF Projects, LLC licensed under the CC BY 4.0 license

MLflow Pilot Instances

MLflow instances are actively provided within multiple EU-funded projects, supporting diverse scientific domains:

- **AI4EOSC** – AI for the European Open Science Cloud
  Usage Report: **Users**: 22 | Active **Experiments/Runs**: 61/1992
- **iMagine** – Image-based AI for marine and aquatic research
  Usage Report: **Users**: 17 | Active **Experiments/Runs**: 78/2006
- **FLUID-AI** – Upcoming project
- **EOSC-ARENA** – Upcoming project
  (See **poster #219**)

mlflow.scc.kit.edu

The pilot instance for **Helmholtz** researchers [3], available at: mlflow.scc.kit.edu | mlops.data.kit.edu

Key Features:

- Coupled with Helmholtz ID/AAI for authentication
- Integrated with LSDF large storage for artifacts
- Experiment and model sharing among users
- Collaborative development workflows
- Group-sharing capabilities for teams
- Quota management

Usage Report: **Users**: 15 | Active **Experiments/Runs**: 32/339

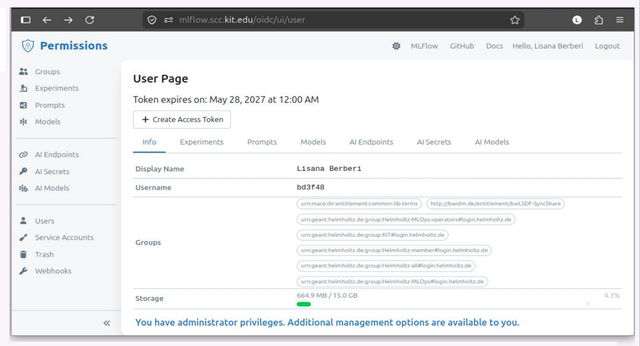


Fig 2: Permissions user page and storage quota dashboard

Use Case 1: Thermal Urban Feature Segmentation

- Semantic segmentation of thermal anomalies in urban environments using UAV-captured thermal
- U-Net model with ResNet-152 backbone trained on 793 images across 9 classes (buildings, cars etc.)
- Comparison of Federated Learning algorithms
- Experiment tracking with MLflow, energy consumption monitoring with Perun[5], and FL orchestration via NVFlare [6]

a) b)

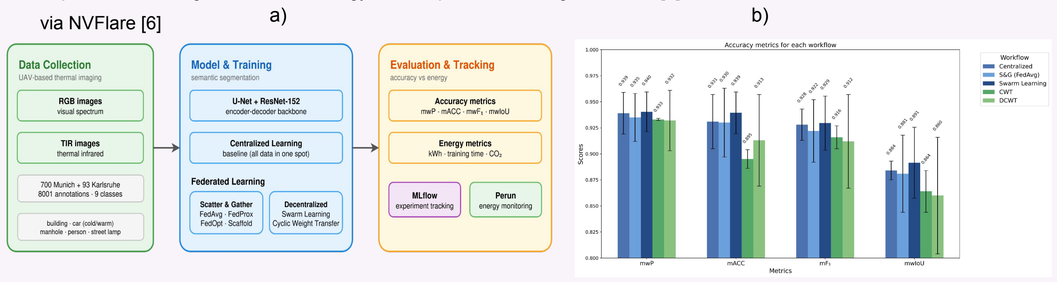


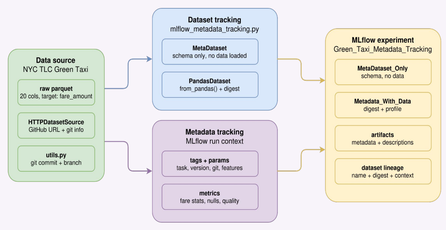
Fig 3: a) UAV-based thermal image segmentation comparing centralized vs federated learning workflows; b) Accuracy metrics per workflow [7]

https://github.com/ai4os-hub/thermal-urban-feature-segmenter

Use Case 2: Taxi ride data lineage pipeline

- Track NYC Green Taxi dataset [8] metadata and lineage using MLflow's `mlflow.data` API
- Log dataset provenance via `HTTPDatasetSource`
- Enrich MLflow runs with tags (prediction task, data version), parameters, and metrics

a)



b)

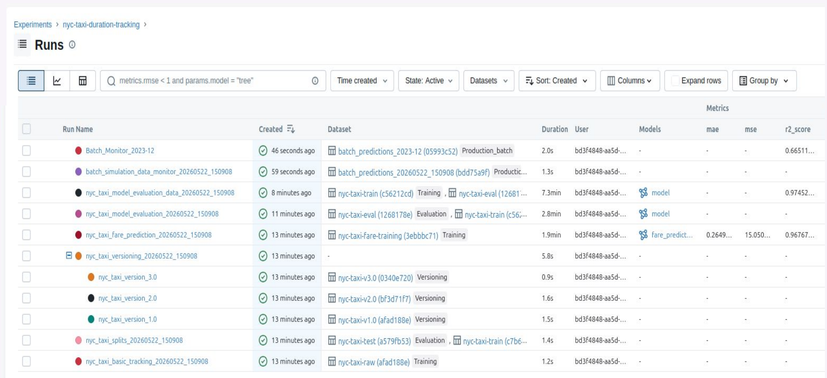


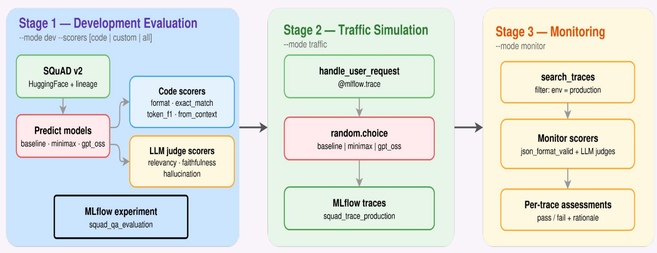
Fig 4: a) Pipeline architecture showing (meta)dataset tracking; b) Exp-runs showing the full experiment hierarchy

https://codebase.helmholtz.cloud/m-team/ai/nyc-taxi-dataset

Use Case 3: GenAI QA evaluation pipeline

- Evaluate GenAI question-answering models using MLflow's `mlflow.genai` module with HuggingFace SQuAD v2 dataset[4]
- Compare multiple LLM endpoints
- Score outputs with code-based scorers (exact match, token F1 etc.)
- Score outputs with custom LLM-as-judge scorers
- Generate production traces via `@mlflow.trace` with user, model tags

a) b)



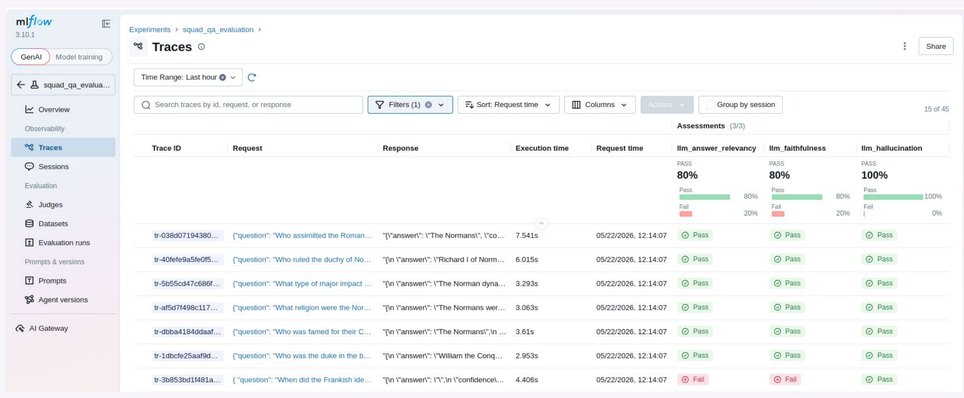


Fig 5: a) Pipeline architecture diagram showing the three stages from development to monitoring; b) MLflow Traces UI showing production traces

https://codebase.helmholtz.cloud/m-team/ai/mlflow-genai-evaluate

References

[1] Berberi, L., Kozlov, V., Nguyen, G. et al. Machine learning operations landscape: platforms and tools. Artif Intell Rev 58, 167 (2025). doi.org/10.1007/s10462-025-11164-3
[2] mlflow.org | mlflow.org/docs
[3] mlflow.scc.kit.edu | mlops.data.kit.edu
[4] Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In I. Gurevych & Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the ACL: https://doi.org/10.18653/v1/P18-2124
[5] https://github.com/Helmholtz-AI-Energy/perun
[6] https://github.com/nvidia/nvflare
[7] Duda, L., Alibabaei, K., Vollmer, E., et al. (2025). Federated learning for thermal urban feature segmentation. In *ICCSA 2025*, 285–302. Springer.
[8] Berberi, L., Gavoçi, E., Bushati, S., Kroni, F. (2026). Evaluating Machine Learning Models for Trip Duration Prediction in Taxi Data. Springer, Cham.

Contact

MLflow support | Scientific Computing Centre (SCC) |
Karlsruhe Institute of Technology | mlops-support@lists.kit.edu

MLflow: mlflow.scc.kit.edu **MLOps Info:** mlops.data.kit.edu