

A practical benchmark of open-source MLOps platforms: Comparing MLflow, Metaflow and ZenML across model type

Thursday 18 September 2025 15:00 (15 minutes)

This study offers a rigorous and reproducible comparison of three widely adopted open-source MLOps frameworks - MLflow, Metaflow, and ZenML. These models have been chosen for this study due to their complementary roles within the open-source MLOps landscape.

MLflow excels in experiment tracking, model packaging, and registry, while Metaflow offers seamless data and code versioning with built-in lineage, and ZenML transforms standard Python into portable, production-ready pipelines with automatic artifact tracking. Together, they cover the core pillars of MLOps-tracking, versioning and orchestration, with complementary strengths, making them ideal for a focused, local benchmarking study.

Each framework is evaluated using three representative machine learning tasks: a Random Forest on tabular data, a ResNet-based convolutional neural network on medical imaging, and a BERT-style text classifier for extractive summarization.

Our analysis evaluates installation overhead, developer effort, training duration, pipeline orchestration, and reproducibility, ensuring consistent outputs across identical runs. It further compares performance tracking, model and data versioning, and registry mechanisms through quantitative and visual metrics such as runtime, accuracy, code complexity, and overall operational efficiency, highlighting trade-offs in developer experience.

Empirical results are captured through a unified benchmark that logs execution time, model metrics and a static comparison between the original code and the three versions obtained through integration with each of the studied frameworks. By presenting measurable insights into integration complexity, usability, performance overhead, and reproducibility, this work advances the understanding of local-scale MLOps tool selection.

This study's results are as follows: MLflow proved effortless to integrate in just a few lines of code and imposed negligible runtime overhead (<2%). Metaflow required slightly more setup (~25 extra lines) but delivered robust versioning of both code and data with a modest runtime cost (~8–10%). ZenML involved the most upfront work (~40 lines of boilerplate), yet rewarded that investment with full pipeline orchestration, transparent artifact lineage, and exceptionally stable results (variation within $\pm 0.1\%$ under fixed seeds), while still maintaining moderate runtime overhead (~5%).

Authors: Mr BADEA, Dan Gabriel (National University of Sciences and Technology Politehnica Bucharest); Mr MONEA, Damian (Crowdstrike)

Presenters: Mr BADEA, Dan Gabriel (National University of Sciences and Technology Politehnica Bucharest); Mr MONEA, Damian (Crowdstrike)

Session Classification: Doctoral Symposium

Track Classification: Open Source and GNU in Education and Research