Automated Benchmarking Suite for Hybrid Models

Author: Jakub Polak, Lucas Cosier, Tiago Mateus
Affiliation: DataHow AG, R&D Team
Scope: Internship
Skills: Coding in Python, Machine Learning Operations, Model delployments
Date: March 11, 2026
Contact: h.narayanan@datahow.ch

Executive Summary

This project proposes the development of an internal Benchmarking Suite that automates the evaluation of all hybrid models developed at DataHow. The tool will run benchmarks against curated datasets, track performance and production-focused metrics (including resource utilisation, model weights, stability, and inference behavior), and present results through an intuitive dashboard. The desired outcome is an internal company tool that replaces the current manual workflow, enables continuous evaluation tied to model releases, and provides clear visual reporting for the R&D team.

Background

Context

As the number of models, datasets, and deployment configurations increases, the need for a systematic and continuous evaluation framework becomes critical. It is an absolute necessity to robustly evaluate the performance, ensure consistency across releases, and accelerate the development cycle. Without an automated benchmarking framework, the team risks undetected quality issues, slow feedback loops, and difficulty communicating model readiness to stakeholders.

Current approaches

An internal benchmarking workflow already exists and provides a solid starting point. Models run as Docker containers to simulate deployment, metrics are logged via MLflow, and an internal dashboard visualizes results. While the core architecture is sound, several limitations motivate the improvements proposed here.

Challenges

MLflow's metric organization and querying do not align well with comparing model families across datasets, requiring a separate dashboard layer. Benchmarking runs require manual triggering and are not tied to new model versions or CI/CD pipelines. Metric coverage is limited to predictive performance; production-relevant metrics such as model weight distributions, training/inference stability, latency, and versioning metadata are not systematically tracked.

Approach

The internship is structured around a set of progressive objectives, each building on the previous phase. The intern will work closely with the R&D team throughout, iterating on requirements and design decisions.

Objectives

1. **Investigate existing capabilities:** Review the current workflow end-to-end, document findings and pain points.
2. **Research and tool evaluation:** Assess alternative experiment-tracking, dashboard, and automation tools that could replace or enhance parts of the stack.
3. **Stakeholder iteration:** Collaborate with R&D to define additional metrics, visualizations, and reporting needs
4. **Automation:** Implement CI/CD-triggered benchmarking runs for each new model version, eliminating manual intervention.
5. **Dashboard and reporting:** Build an internal dashboard for at-a-glance comparisons and support on-request report generation for each release cycle.

Methods and Work Plan

Data and Resources

All necessary datasets and computational resources will be provided.

Timeline

Timeline subject to the scope. The expected duration of this project is 4-6 months

Phase	Tasks	Target Dates
Discovery	Familiarize with DataHowLab, hybrid models, and current benchmarking workflow. Document gaps.	Wk 1-3
Research and Analysis	Evaluate alternative tools and frameworks. Present recommendations.	Wk 4-6
Design and Prototyping	Architect the benchmarking suite. Build initial prototypes of pipeline and dashboard.	Wk 7-10
Implementation	Develop automated triggers, expanded metrics collection, storage, and visualization.	Wk 11-18
Integration and Testing	Integrate with CI/CD, run end-to-end tests, iterate on feedback.	Wk 19-22
Wrap-up	Finalize docs, deliver release benchmarking report, handover the benchmarking suite knowledge to the team.	Wk 23-24

Expected Outcome

- Benchmarking Suite: Automated evaluation tool covering all hybrid models with expanded metric coverage (performance, weights, stability, latency).
- Internal Dashboard: Visual interface for comparing models, versions, and datasets at a glance.
- Automated Pipeline: CI/CD-integrated triggers eliminating manual benchmarking runs.
- Release Reports: On-request benchmarking summaries for each model release cycle.
- Technical Documentation and Research Summary: Architecture docs, setup guides, and a written evaluation of tools considered.