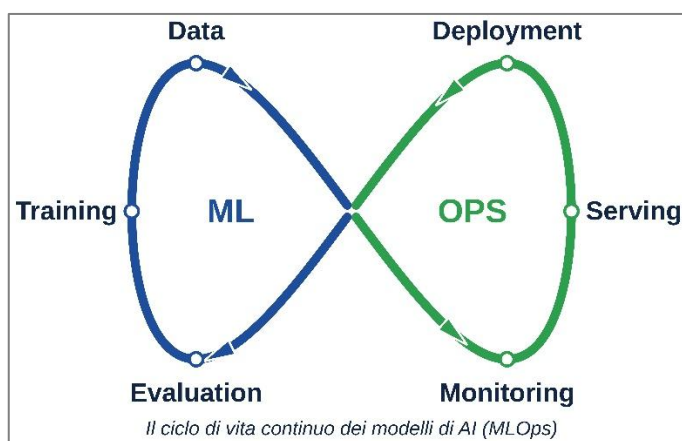




RISE

Infrastruttura Strategica della Ricerca
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Servizio integrato di automazione della costruzione e gestione dei modelli di AI



Per informazioni rivolgersi al responsabile scientifico:

Prof. Filippo Lanubile
filippo.lanubile@uniba.it

Polo Scientifico Tecnologico

Dipartimento di Informatica

Via Orabona 4, 70125 (BA)

ENG: Within the framework of the PNC (National Plan for Complementary Investments to the PNRR), the University of Bari has developed an integrated service for automating the construction and management of Artificial Intelligence models, covering the entire life cycle of AI-enabled components: from requirements definition to the monitoring of models in production. The service builds on the MLOps (Machine Learning Operations) methodologies investigated by the collaborative software engineering research group (COLLAB) of the Department of Computer Science, which extend DevOps principles to the development of AI systems (Calefato et al., 2022).

The service ensures the reproducibility and traceability of machine learning experiments through version control of code and data, experiment tracking, and a model registry for archiving and registering trained models. Development workflows are fully automated by means of training pipelines and continuous integration and deployment (CI/CD) mechanisms, which allow AI models to be built, updated, and released in a controlled and repeatable way.

Quality assurance components verify the quality of code, data, and models, and specific checks can be integrated to guarantee compliance with current regulations. Models are exposed through Web APIs for interoperability, while dedicated components support Large Language Models, including fine-tuning, Retrieval Augmented Generation (RAG), and conversational interfaces. Models in production

are continuously monitored, with an alerting system that signals component unavailability or suboptimal performance (Mallardi et al., 2024; Mallardi et al., 2025).

The service is hosted on the new hardware infrastructure of the Department of Computer Science of the University of Bari, based on a Dell PowerEdge XE7745 server equipped with 4 NVIDIA H200 GPUs (141 GB each), 2 AMD EPYC 9555 processors (128 cores in total), 1.5 TB of RAM, complemented by a Dell PowerScale NAS storage cluster (4 A300L nodes, 120 TB of raw capacity).

The objective is to offer public and private stakeholders an integrated automation service for building, releasing, and operating reliable and compliant AI models.

ITA: Nell'ambito del PNC (Piano Nazionale Complementare al PNRR), l'Università di Bari ha sviluppato un servizio integrato per l'automazione della costruzione e della gestione di modelli di Intelligenza Artificiale, che copre l'intero ciclo di vita delle componenti basate sull'IA: dalla definizione dei requisiti al monitoraggio dei modelli in produzione. Il servizio si fonda sulle metodologie MLOps (Machine Learning Operations) studiate dal gruppo di ricerca di ingegneria del software collaborativa (COLLAB) del Dipartimento di Informatica, che estendono i principi DevOps allo sviluppo di sistemi di IA (Calefato et al., 2022).

Il servizio garantisce la riproducibilità e la tracciabilità degli esperimenti di machine learning attraverso il controllo di versione di codice e dati, il monitoraggio degli esperimenti (experiment tracking) e un model registry per l'archiviazione e la registrazione dei modelli addestrati. I workflow di sviluppo sono completamente automatizzati mediante pipeline di addestramento e meccanismi di integrazione e rilascio continui (CI/CD), che consentono di costruire, aggiornare e rilasciare i modelli di IA in modo controllato e ripetibile.

Componenti di quality assurance verificano la qualità di codice, dati e modelli, ed è possibile integrare controlli specifici per garantire la compliance con le regolamentazioni vigenti. I modelli sono esposti tramite Web API per l'interoperabilità, mentre componenti dedicate supportano i Large Language Model, incluse le funzionalità di fine-tuning, Retrieval Augmented Generation (RAG) e interfacce conversazionali. I modelli in produzione sono monitorati in modo continuo, con un sistema di allarme che segnala l'indisponibilità delle componenti o prestazioni subottimali (Mallardi et al., 2024; Mallardi et al., 2025).

Il servizio è ospitato sulla nuova infrastruttura hardware del Dipartimento di Informatica dell'Università di Bari, basata su un server Dell PowerEdge XE7745 dotato di 4 GPU NVIDIA H200 (141 GB ciascuna), 2 processori AMD EPYC 9555 (128 core complessivi), 1,5 TB di RAM, affiancato da un cluster storage NAS Dell PowerScale (4 nodi A300L, 120 TB di capacità raw).

L'obiettivo è offrire a stakeholder pubblici e privati un servizio integrato di automazione per costruire, rilasciare e gestire modelli di IA affidabili e conformi alle regolamentazioni.

References

- 1) F. Calefato, F. Lanubile, L. Quaranta. "A Preliminary Investigation of MLOps Practices in GitHub". In ESEM '22: ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp. 283–288, 2022. <https://doi.org/10.1145/3544902.3546636>
- 2) G. Mallardi, F. Calefato, L. Quaranta, F. Lanubile. "An MLOps Approach for Deploying Machine Learning Models in Healthcare Systems". In IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2024), pp. 6832–6837, 2024.
- 3) G. Mallardi, L. Quaranta, F. Calefato and F. Lanubile, "Towards Ensuring Responsible AI for Medical Device Certification," 2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE), Ottawa, ON, Canada, 2025, pp. 29-32, doi: 10.1109/RAIE66699.2025.00009.