Projektausschreibung: Themenfeld: "MLOps / AI Engineering, durchgängige Werkzeug- und Prozesskette für die Entwicklung und den produktiven Betrieb eigener KI-Modelle"

1. Projekt

1.1 Titel

MLOps-Pipelines für die Eigenentwicklung von KI: Tooling, Experiment-Tracking, Deployment und Monitoring als durchgängiger Forschungs- und Engineering-Workflow

1.2 Laufzeit, Mittel (Höhe, Mittelgeber), Einbindung in größeres Projekt

3 Semester, Eigenfinanzierung.

1.3 Kurzbeschreibung der Ziele / Aufgaben

Unternehmen und Forschungseinrichtungen entwickeln zunehmend eigene KI-Modelle, scheitern in der Praxis aber häufig nicht am Modell selbst, sondern an der begleitenden Werkzeug- und Prozesskette: Daten werden inkonsistent versioniert, Experimente sind nicht reproduzierbar, Modelle gelangen nur schwer in den Betrieb, und nach dem Deployment fehlt ein verlässliches Monitoring. MLOps adressiert genau diese Lücke zwischen ML-Forschung und produktivem Einsatz.

Im Forschungsprojekt soll eine durchgängige MLOps-Referenzarchitektur für die Eigenentwicklung von KI-Modellen aufgebaut, evaluiert und kritisch weiterentwickelt werden. Die Betrachtung umfasst alle Stationen des ML-Lifecycles:

- **Daten:** Versionierung, Qualitätssicherung und reproduzierbare Preprocessing-Pipelines.
- **ML-Entwicklung:** Experiment-Tracking, Hyperparameter-Management, konfigurierbare Trainings-Workflows.
- **Modelle:** Versionierung, Registry, Lineage zwischen Daten, Code und Modell.
- **Deployment:** Verpackung und Auslieferung der Modelle in unterschiedliche Zielumgebungen (On-Premises, Container, Edge).
- **Betrieb und Monitoring:** Beobachtbarkeit im Betrieb (Drift, Qualität, Ressourcen), Feedback-Schleifen und kontrolliertes Re-Training.

Ziel ist es, eine forschungstaugliche und zugleich praxistaugliche Pipeline zu etablieren und daraus wiederverwendbare Architektur- und Prozessmuster abzuleiten.

1.4 Wissenschaftlicher Anteil für Forschungsmaster

Das Forschungsprojekt kann je nach Interessenslage unterschiedliche Schwerpunkte setzen. Aufbauend auf bestehenden Vorarbeiten im Forschungscluster sind folgende Forschungsthemen möglich:

1. **Daten- und Experiment-Management:** Vergleich und Integration von Tools für Datenversionierung und Experiment-Tracking (z. B. DVC, MLflow, Weights & Biases) unter reproduzierbaren Bedingungen.
2. **Reproduzierbare Trainings-Workflows:** Nutzung konfigurationsgetriebener Pipelines (Hydra + OmegaConf, `uv`, Transformers Trainer) zur Durchführung vergleichbarer Experimente über Daten, Modelle und Hyperparameter hinweg.
3. **Modell-Lifecycle und Registry:** Entwurf einer Modell-Registry mit nachvollziehbarer Kopplung an Daten- und Code-Stand; Untersuchung von Rollback- und Promotion-Strategien.
4. **Deployment-Strategien:** Systematischer Vergleich von Deployment-Pfaden (Container, On-Premises-Serving, Edge) hinsichtlich Aufwand, Performance und Wartbarkeit.
5. **Monitoring und Qualitätssicherung im Betrieb:** Erkennung von Daten- und Modell-Drift, Aufbau geeigneter Metriken und Alarmierung, Rückkopplung in das Re-Training.
6. **Architekturforschung für MLOps:** Ableitung wiederverwendbarer Architektur- und Prozessmuster für MLOps-Pipelines.

Das Projekt ist explizit als Forschungsbeitrag an der Schnittstelle von Software-Architektur, ML-Engineering und Betrieb angelegt.

---

2. Durchführende Stelle

2.1 Institut / Labor / Arbeitsplatz

Hochschule Ansbach, CCS – Center for Signal Analysis of Complex Systems

2.2 Betreuerin / Co-Betreuer / Betreuende Laboringenieurinnen

Prof. Nicolas Weeger

2.3 Notwendige Ausstattung vorhanden / wird in Projektlaufzeit beschafft

Die notwendige Ausstattung ist überwiegend vorhanden: GPU-Workstation mit NVIDIA RTX 5090 sowie Zugang zu einer DGX-Spark-Plattform (GB10 Blackwell, Unified Memory) für Trainings- und Serving-Experimente. Für Vergleichsmessungen kann zusätzlich auf performantere Hardware zurückgegriffen werden. Standard-Software (Python, `uv`, gängige MLOps-Tools) ist etabliert.

---

3. Reporting

3.1 Rahmen für Projekt- / Masterseminar vorhanden

Projektfortschritt und Ergebnisse sind im Rahmen der APR-Seminare und Tagungen darzustellen.

3.2 Veröffentlichung geplant auf Konferenz / in Zeitschrift / als Patentanmeldung

Veröffentlichungen sind entsprechend dem Projektfortschritt vorgesehen, u. a. auf Konferenzen im Bereich Software-Architektur und Software Engineering (ICSA, ECSA, ICSE) sowie in peer-reviewed Zeitschriften (z. B. TSE, TOSEM, IEEE Software, Software: Practice and Experience). Zusätzlich sind Beiträge auf anwendungsorientierten Formaten möglich.

---

4. Anforderungen an Bewerber

4.1 Gewünschte/vorausgesetzte Fachrichtung eines Hochschulabschlusses

Bachelor of Science oder Bachelor of Engineering in Informatik, Software Engineering, KI, Data Science oder verwandten Disziplinen. Interesse am Zusammenspiel von Software-Architektur, MLOps, AI-Engineering und Betrieb sowie die Bereitschaft, sich in neue Themenfelder einzuarbeiten.

4.2 Vorteilhaft sind folgende Vertiefungen / praktische Erfahrungen

- Solide Python-Kenntnisse
- Erfahrung mit Machine Learning / Deep Learning (z. B. PyTorch, Tensorflow, CUDA, Transformers)
- Grundkenntnisse in Software Architektur und DevOps (Docker, CI/CD, Container-Orchestrierung) sind willkommen, aber nicht zwingend
- Erste Erfahrung mit MLOps-Werkzeugen (z. B. MLflow, DVC, Weights & Biases) ist willkommen, aber nicht zwingend
- Sicheres Englisch in Wort und Schrift (wissenschaftliche Publikationen)

Bei Interesse oder Fragen wenden Sie sich bitte an Prof. Nicolas Weeger (nicolas.weeger@hs-ansbach.de)