Friedrich-Alexander-Universität
Erlangen-Nürnberg

AI Inference Platform Engineer (m/f/d)

Regionales Rechenzentrum Erlangen, Erlangen, TV-L E 13, Full time, Temporary employment: until 31.07.2029, Bewerbungsschluss: 07.06.2026

Your Workplace

The Erlangen Regional Computing Center (RRZE) is FAU's internal IT service provider and provides the IT infrastructure and basic IT services. As part of the regional concept, it also supports other colleges and universities in the surrounding area.

The High Performance Computing (HPC@FAU) group works closely with the National Supercomputing Center (NHR@FAU) and the Leibniz Supercomputing Center in Garching near Munich.

Job Benefits

- Regular promotion to the next level and increase in salary pursuant to the collective bargaining agreement for the public service of the German Länder (TV-L) or remuneration pursuant to the Bavarian Public Servants Remuneration Act (BayBesG) plus an additional annual bonus
- 30 days annual leave at five working days per week with additional free days on December 24 and 31
- Occupational pension scheme and asset accumulation savings scheme

Description

Your Role and Responsibilities:

- Designing, implementing and maintaining an AI inference platform based on predominantly open-source components including a web-based user interface and API, all within a friendly and open work environment in a highly motivated, international team
- Conceptualizing and implementing infrastructure components to create a RAG-capable inference environment
- Advising and supporting pilot project partners of select universities using this AI service infrastructure in data quality, data preparation and workflow design to contribute to the transfer of prototypes into production, relying on your friendly personality and communication skills
- Designing and implementing tenant separation concepts for access, data and compute, integrating with federated single sign-on (SSO) institutional identity management systems
- Implementing resource management mechanisms to ensure fair and efficient resource allocation and to allow for usage accounting and cost attribution

Qualifications

Required/Minimum Qualifications

PhD or Master's degree in computer or data science, or other areas of scientific computing,

Other Requirements

- Proficiency working in data center environments (incl. Linux, CLI, Git, Gitlab)
- Extensive knowledge and experience in developing and maintaining platform environments in the context of AI inference workflows, that is utilizing e.g.
    - web server / load balancer (e.g. Nginx), data bases (e.g. MariaDB, SQLite, Redis)
    - containers/OS-level virtualization (e.g. Docker) and container orchestration (e.g. Kubernetes), as well as HPC-based scheduler (e.g. Slurm)
    - monitoring tools for metric collection (e.g. Prometheus) and visualization (e.g. Graphana)
    - Python and JavaScript programming languages for development of frontend components (e.g. Open WebUI)
    - model gateway (e.g. LiteLLM) and inference engines (e.g. vLLM, Triton, SGLang) as well as underlying GPU-based technologies (e.g. torch, ray)
- Knowledge of various types of AI models (e.g. LLMs, vision-language models, …), model guardrails and retrieval-augmented generation (RAG)
- Willingness to keep up with current developments and to learn new technologies in the field of AI
- Knowledge and experience in software deployment and software lifecycle management (ideally based on principles of continuous integration/continuous deployment, CI/CD)
- Basic knowledge and practical skills in software design and engineering
- Basic knowledge and practical skills in IT and cyber security for software and software platform development
- English and German presentation and writing skills

Interessiert?

Die vollständige Stellenausschreibung sowie alle Infos zum Bewerbungsverfahren finden Sie hier: