



UNIVERSITÄT
ZU KÖLN



Foto: Thomas Jösel

Platform Engineer*in KI-Infrastruktur (w/m/d)

Abt. Research Infrastructure Services (RIS) | IT Center University of Cologne (ITCC)

Wir sind eine der größten und ältesten Universitäten Europas und gehören zu den größten Arbeitgeber*innen in unserer Region. Durch unser breites Fächerspektrum, die dynamische Entwicklung unserer Forschungsschwerpunkte und unseren Standort mitten in Köln sind wir attraktiv für Studierende und Forschende weltweit. Wir bieten vielfältige Karrierechancen in Wissenschaft, Technik und Verwaltung.

Das IT Center University of Cologne (ITCC) ist eine zentrale Einrichtung der Universität zu Köln, die als Dienstleistungs- und Kompetenzzentrum technische und wissenschaftliche Zusammenarbeit auf Landes-, Bundes- sowie internationaler Ebene anstrebt. Die Abteilung Research Infrastructure Services (RIS) betreibt die zentrale Forschungsinfrastruktur der Universität und baut im Verbundprojekt COSAI gemeinsam mit sechs weiteren Partnern eine souveräne, länderübergreifende KI-Infrastruktur auf.

IHRE AUFGABEN

- » Konzeption, vergleichende Evaluierung und Aufbau einer souveränen, mandantenfähigen Observability-Plattform auf Open-Source-Basis
- » Integration angepasster Sprachmodelle (Bring-Your-Own-Model) in eine föderierte Kubernetes-Inferenzinfrastruktur
- » Forschung zu KI-gestützter Betriebs- und Sicherheitsanalyse mit Sprachmodellen und MCP-Schnittstellen bis hin zu agentenbasierten Ansätzen
- » Untersuchung ressourceneffizienter GPU-Nutzung, z. B. durch Scale-to-Zero, GPU-Virtualisierung und Caching von Modellgewichten
- » Veröffentlichung der Ergebnisse in Fachpublikationen, auf Tagungen und als Referenzarchitekturen und Best Practices für die Hochschulgemeinschaft

IHR PROFIL

- » abgeschlossenes wissenschaftliches Hochschulstudium (Master, Diplom oder vergleichbar) in Informatik, Wirtschaftsinformatik oder einem verwandten Fach
- » fundierte Kenntnisse in Container-Technologien und Kubernetes
- » Erfahrung mit Monitoring- und Observability-Werkzeugen sowie Grundverständnis von IT-Sicherheit in Cloud-Umgebungen
- » wünschenswert: Erfahrung mit dem Betrieb von Sprachmodellen (z. B. vLLM), mit MCP oder agentenbasierten KI-Systemen sowie mit HPC-Umgebungen
- » Interesse an wissenschaftlicher Arbeit und an der Veröffentlichung von Ergebnissen
- » selbstständige, strukturierte Arbeitsweise

- » Freude an der Zusammenarbeit in einem verteilten Verbundteam und gute Kommunikationsfähigkeit
- » gute Deutsch- und Englischkenntnisse

WIR BIETEN IHNEN

- » abwechslungsreiche und spannende Projektarbeiten
- » ein vielfältiges und chancengerechtes Arbeitsumfeld
- » Unterstützung bei der Vereinbarkeit von Beruf und Familie
- » flexible Arbeitszeitmodelle
- » umfangreiches Weiterbildungsangebot
- » Angebote im Rahmen des Betrieblichen Gesundheitsmanagements
- » Möglichkeit zur mobilen Arbeit

Die Universität zu Köln fördert Chancengerechtigkeit und Vielfalt. Bewerbungen von Frauen werden nach Maßgabe des LGG NRW bevorzugt berücksichtigt. Wir begrüßen ausdrücklich alle Bewerbungen – unabhängig von Geschlecht, Nationalität, ethnischer und sozialer Herkunft, Religion, Behinderung, Alter sowie sexueller Orientierung und Identität.

Die Stelle ist ab 01.01.2027 in Vollzeit (39,83 Wochenstunden) zu besetzen. Sie ist bis 30.09.2029 befristet. Sofern die entsprechenden tariflichen und persönlichen Voraussetzungen vorliegen, richtet sich die Vergütung nach der Entgeltgruppe I3 TV-L.

Bitte bewerben Sie sich mit Ihren aussagekräftigen Bewerbungsunterlagen inkl. beigefügten Nachweisen für die gesuchten Qualifikationen ohne Bewerbungsfoto online unter: <https://jobportal.uni-koeln.de>. Die Kennziffer ist TUV2610-02. Die Bewerbungsfrist endet am 08.11.2026.

Bei Fragen wenden Sie sich bitte an Herrn Daniel Wickerroth (wickerroth@uni-koeln.de) und schauen Sie in unsere [FAQs](#).



HR EXCELLENCE IN RESEARCH