Technische Universität Berlin

Bei der Technischen Universität Berlin ist/sind folgende Stelle/n zu besetzen:

Wiss. Mitarbeiter*in (d/m/w) - Entgeltgruppe 13 TV-L Berliner Hochschulen

unter dem Vorbehalt der Mittelbewilligung; Teilzeitbeschäftigung ist ggf. möglich

Fakultät IV - Institut für Telekommunikationssysteme / FG Verteilte Systeme und Betriebssysteme (DOS)
Kennziffer: IV-195/25 (besetzbar ab 01.07.2025 / befristet bis 30.06.2027 / Bewerbungsfristende 30.05.2025)

Aufgabenbeschreibung:

Mitarbeit in Forschung und Lehre im Fachgebiet Verteilte Systeme und Betriebssysteme; Veröffentlichung von Forschungsergebnissen.

Große Sprachmodelle (LLMs) liegen im Trend. Die zunehmende Modellgröße erfordert jedoch die Entwicklung und Bereitstellung komplexerer IT-Infrastrukturen. Aufgrund von Speicherbeschränkungen wird zunehmend ein verteiltes LLM-Training durchgeführt, welches jedoch große und komplexe IT-Infrastrukturen erfordert. Dadurch steigt auch die Ausfallwahrscheinlichkeit einzelner Komponenten, die wiederum zu höheren Betriebskosten und Ressourcenverschwendung führt. Eine effektive Fehlerüberwachung erfordert daher ein umfassendes Verständnis der IT-Infrastruktur unter Berücksichtigung des Zusammenspiels von Metriken aus Inter-/Intra-Host-Netzwerken CPUs, NPUs, GPUs, Kommunikationsmuster sowie den Besonderheiten eines LLM-Trainings. Ziel dieses Projekts ist die Entwicklung eines Frameworks zur Erkennung und Vorhersage von Fehler in großen Sprachmodellen, insbesondere in Mixture of Experts-Architekturen. Dies basiert auf einer umfassenden Analyse und dem Verständnis von Fehlermechanismen in Kommunikations-, Rechen- und Speicherkomponenten während des Trainings und der Inferenz.

Wir konzentrieren uns auf folgende Themen: Verständnis und Analyse der während des LLM-Trainings generierten Signale, Simulation von Szenarien durch Hinzufügen von synthetisch generierten Fehlern, Verständnis von Wechselwirkungen zwischen Komponenten in großen KI-Infrastrukturen, Überwachung und Interpretation von Daten aus der physikalischen Schicht (Hardware), der Datenschicht (Speicherung und Übertragung), der Rechenschicht und der Anwendungsschicht (Modelle). Unser Ziel ist es, gemeinsame Repräsentationen aus den verschiedenen Systemdatenquellen zu erlernen, um Anomalien und deren Ursachen zu erkennen. Dies beinhaltet die Entwicklung einer allgemeinen Methode, die Implementierung eines Prototyps im Kontext bestehender Open-Source-Systeme sowie die experimentelle Evaluierung des Prototyps mit Testdaten aus experimentellen und Produktionsdaten.

Die Möglichkeit zur Promotion ist gegeben.

Erwartete Qualifikationen:

- Erfolgreich abgeschlossenes wissenschaftliches Hochschulstudium (Master, Diplom oder Äquivalent) in Informatik mit Spezialisierung auf Betrieb komplexer IT-Infrastrukturen und Maschinelles Lernen
- Erfahrung mit Statistiksoftware, Monitoring-Tools und Betriebssystemen
- Erfahrung mit ML-Methoden für Erkennungs- und Klassifizierungsaufgaben
- Erfahrung im Umgang mit großen Clustersystemen
- Aufbau und Betrieb von Containern (z. B. Singularity, Docker)
- Erfahrung mit TensorFlow/PyTorch/Keras
- Gute Deutsch- und/oder Englischkenntnisse sind erforderlich; Bereitschaft, die jeweils fehlenden Sprachkenntnisse zu erwerben

Wünschenswert:

- Interesse an der Systementwicklung und dem Betrieb groß angelegter Softwarearchitekturen sowie die Bereitschaft, aktuelle Forschungsergebnisse in die Praxis umzusetzen
- Erfahrung im Verfassen und Publizieren wissenschaftlicher Arbeiten
- Vertraut mit Methoden und Methodiken aus dem Bereich der Zeitreihenanalyse
- Erfahrung und Interesse an den Themen KI und KI-Infrastrukturen
- Erfahrung im Umgang mit erklärbaren Methoden des maschinellen Lernens und Daten aus heterogenen Quellen
- Erfahrung in der Entwicklung zugänglicher Technologien
- Interesse an Projektmanagement und agilen Entwicklungsmethoden

Ihre **schriftliche** Bewerbung richten Sie bitte unter **Angabe der Kennziffer** mit den üblichen Unterlagen (Lebenslauf, Notenliste, ggf. Nachweise von Sprachkenntnissen) an die Technische Universität Berlin, Herrn Prof. Odej Kao: **odej.kao@tu-berlin.de**.

Mit der Abgabe einer Onlinebewerbung geben Sie als Bewerber*in Ihr Einverständnis, dass Ihre Daten elektronisch verarbeitet und gespeichert werden. Wir weisen darauf hin, dass bei ungeschützter Übersendung Ihrer Bewerbung auf elektronischem Wege keine Gewähr für die Sicherheit übermittelter persönlicher Daten übernommen werden kann. Datenschutzrechtliche Hinweise zur Verarbeitung Ihrer Daten gem. DSGVO finden Sie auf der Webseite der Personalabteilung: https://www.abt2-t.tu-berlin.de/menue/themen_a_z/datenschutzerklaerung/ oder Direktzugang: 214041.

Zur Wahrung der Chancengleichheit zwischen Frauen und Männern sind Bewerbungen von Frauen mit der jeweiligen Qualifikation ausdrücklich erwünscht. Schwerbehinderte werden bei gleicher Eignung bevorzugt berücksichtigt. Die TU Berlin schätzt die Vielfalt ihrer Mitglieder und verfolgt die Ziele der Chancengleichheit. Bewerbungen von Menschen aller Nationalitäten und mit Migrationshintergrund sind herzlich willkommen.

Technische Universität Berlin - Die Präsidentin - Institut für Telekommunikationssysteme, FG Verteilte Systeme und Betriebssysteme, Prof. Dr. Odej Kao, Sekr. EN 22, Einsteinufer 17, 10587 Berlin

Die Stellenausschreibung ist auch im Internet abrufbar unter:
https://www.personalabteilung.tu-berlin.de/menue/jobs/

Zertifikat seit 2008
audit familiengerechte hochschule