Ruizhang Zhou

ruizhang.zhou@mail.com | GitHub | LinkedIn | Website

Applied AI / AI Platform Engineer - Produktive LLM-Systeme · Backend & Infrastruktur · ML Engineering

Erfahrung

Software Engineer (AI Platform) | KI:connect - RWTH IT Center | 09/2024 - heute

- Mitarbeit an OpenAI-kompatiblen Chat-Completions- und Responses-APIs, einschließlich Raw-Request-Weitergabe und Streaming
- Arbeit an modellspezifischer Token-Schätzung, Nutzungs- und Kostenberechnung, Cache-Write-Preisen und Quoten
- Bewertung von Websuche für Agent-Workflows und RAG-Integration: Provider-Grenzen, Mandantenfreigaben, Ingestion und Service-Authentifizierung
- Entwicklung von ASP.NET-Core-Diensten und Vue-3-Admin-Oberflächen; Analyse von Releases, SSO und Fehlern über CI/CD, Nomad, IIS, Grafana/OpenSearch und Provider hinweg
- **Tech:** C#/.NET, ASP.NET Core, Vue 3, MongoDB, SignalR, OpenAI-kompatible APIs

---

Ausbildung

M.Sc. Informatik | RWTH Aachen University | Okt. 2022 - Jun. 2024

- Masterarbeit: KI-basierte Testszenarien für autonome Fahrzeuge ↗
  - Implementierung von TimeGAN- und Diffusion-TS-Modellen zur Generierung realistischer Fahrzeugtrajektorien für autonome Fahrzeugtests
  - Verarbeitung realer Fahrdatensätze (inD, rounD, exiD) und Evaluation generierter Trajektorien

B.Sc. Informatik | RWTH Aachen University | Okt. 2019 - Sep. 2022

B.A. Germanistik | Tongji Universität | Sep. 2014 - Aug. 2018

Ausgewählte frühere Tätigkeiten

Wissenschaftliche Hilfskraft (Teilzeit) | RWTH Lehrstuhl Embedded Software | Aug. 2023 - Mär. 2024

- Wartung der CPM Remote Web-Anwendung für autonome Fahrzeug-Algorithmentests mit Echtzeit-Visualisierungen
- **Tech:** TypeScript, Angular, Docker, GitLab CI/CD

Wissenschaftliche Hilfskraft (Teilzeit) | RWTH Lehrstuhl DBIS | Jul. 2023 - Mär. 2024

- Forschung zu medizinischer Bildverarbeitung: Kombination von BLIP Vision-Language-Modell mit LLMs für Röntgenbild-Berichterstellung
- Untersuchung von Knowledge Graphs und GNNs für medizinische Daten
- Einrichtung von LLM-Serving mit Ollama; Entwicklung eines FastAPI-REST-Services und End-to-End-LLM-Chat-Plattform
- **Tech:** Python, PyTorch, Transformers, PyG, BLIP, CUDA, Ollama, FastAPI, vLLM

Projekte

TalkEcho - Meeting-Untertitel und Diktierfunktion ↗ | Aktiv weiterentwickelter persönlicher Fork

- Windows-Diktier- und Tray-Funktionen, stabilere Verarbeitung langer Meetings sowie sichere Speicherung von Provider-Schlüsseln

LLM Gateway - Aufgabenabhängiges Modell-Routing ↗ | Eigenständiges Open-Source-Projekt

- Verteilt Jobs auf verfügbare Modelle, mit Provider-Fallback und Wechsel beim Streaming nur vor der ersten Ausgabe