Ruizhang Zhou

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

KI/ML-Ingenieur & Full-stack-Entwickler - Produktive LLM-Systeme, Deep-Learning-Forschung, Datenpipelines, Mandantenfähige Plattformen

Erfahrung

Software Engineer (AI Platform) - Vollzeit | KIconnect - RWTH IT Center | 09/2024 - heute

- Entwicklung einer produktiven Multi-Tenant-LLM-Plattform mit Streaming-Chat, Echtzeit-Kostentracking und Enterprise-SSO
- Implementierung einheitlicher Tokenisierung für OpenAI, Llama, Gemma, Mistral, Qwen, DeepSeek; Entwicklung eines Vue 3 Admin-Dashboards
- **Tech:** C#/.NET 9, ASP.NET Core, Semantic Kernel, Vue 3, MongoDB, SignalR, Azure AI Inference

---

Ausbildung

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

- Masterarbeit: AI-Based Generation of Testing Scenarios for Motion Planners on Connected and Automated Vehicles ↗
  - Implementierung von TimeGAN- und Diffusion-TS-Modellen zur Generierung realistischer Fahrzeugtrajektorien für autonome Fahrzeugtests
  - Automatisierte Generierung von 1000+ diversen Testszenarien aus realen Fahrdatensätzen (inD, rounD, exiD)
- Schwerpunkt KI/ML: Deep Learning, NLP, Large Language Models, Wissensrepräsentation

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 der Knowledge-Graph-Integration mit GNNs (GCN, GAT, GraphSAGE) zur Verbesserung des medizinischen Domänenverständnisses
- 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 - Echtzeit-Meeting-Untertitel & Übersetzung ↗ | Persönliches Projekt | Laufend

- Desktop-Overlay-App für unsichtbare Echtzeit-Meeting-Untertitel mit Übersetzung; erfasst Systemaudio und Mikrofon
- **Tech:** TypeScript, Rust, Electron, Speech-to-Text API, LLM-Integration

Self-hosted Automation Server | Persönliche Infrastruktur | Laufend

- Automatisierte Bots für den täglichen Gebrauch (Aachener Terminüberwachung, Quant-Daten-Alerts) mit Matrix-Benachrichtigungen
- **Tech:** Python, Linux Server, Matrix API, Automation Scripts