Ruizhang Zhou

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

AI/ML Engineer & Full-stack Developer - Production LLM systems, Deep Learning research, Data pipelines

Experience

Software Engineer (AI Platform) - Full-time | KIconnect - RWTH IT Center | 09/2024 - Present

- Built production multi-tenant LLM platform with streaming chat, real-time cost tracking, and enterprise SSO
- Implemented unified tokenization for OpenAI, Llama, Gemma, Mistral, Qwen, DeepSeek models; developed Vue 3 admin dashboard
- **Tech:** C#/.NET 9, ASP.NET Core, Semantic Kernel, Vue 3, MongoDB, SignalR, Azure AI Inference

---

Education

M.Sc. Computer Science | RWTH Aachen University | Oct. 2022 - Jun. 2024

- Master's Thesis: AI-Based Generation of Testing Scenarios for Motion Planners on Connected and Automated Vehicles ↗
  - Implemented TimeGAN and Diffusion-TS models to generate realistic vehicle trajectory data for autonomous vehicle testing
  - Automated generation of 1000+ diverse test scenarios from real-world driving datasets (inD, rounD, exiD)
- Focus on AI/ML: Deep Learning, NLP, Large Language Models, Knowledge Representation

B.Sc. Computer Science | RWTH Aachen University | Oct. 2019 - Sep. 2022

B.A. German Language and Literature | Tongji University | Sep. 2014 - Aug. 2018

Selected previous roles

Research Assistant (Part-time) | RWTH Chair of Embedded Software | Aug. 2023 - Mar. 2024

- Maintained CPM Remote Web Application for autonomous vehicle algorithm testing with real-time visualizations
- **Tech:** TypeScript, Angular, Docker, GitLab CI/CD

Research Assistant (Part-time) | RWTH Chair of DBIS | Jul. 2023 - Mar. 2024

- Research on medical imaging AI: combined BLIP vision-language model with LLMs for chest X-ray report generation
- Investigated Knowledge Graph integration with GNNs (GCN, GAT, GraphSAGE) to improve medical domain understanding
- Set up LLM serving with Ollama; built FastAPI REST service and end-to-end LLM chat platform
- **Tech:** Python, PyTorch, Transformers, PyG, BLIP, CUDA, Ollama, FastAPI, vLLM

Projects

TalkEcho - Real-time Meeting Subtitle & Translation ↗ | Personal Project | Ongoing

- Desktop overlay app for invisible real-time meeting subtitles with translation; captures system audio and microphone
- **Tech:** TypeScript, Rust, Electron, Speech-to-Text API, LLM Integration

Self-hosted Automation Server | Personal Infrastructure | Ongoing

- Run automated bots for daily use (Aachen appointment monitoring, quant data alerts) with Matrix notifications
- **Tech:** Python, Linux Server, Matrix API, Automation Scripts