Ruizhang Zhou

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

Applied AI / AI Platform Engineer - Production LLM Systems · Backend & Infrastructure · ML Engineering

Experience

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

- Contribute to OpenAI-compatible Chat Completions and Responses APIs; validate streaming and provider behavior
- Work on token estimation, usage/cost metering, cache-write pricing, quotas, and Vue 3 admin interfaces
- Review agent web-search and RAG integration: tenant access, ingestion, and service authentication
- Diagnose releases, SSO, and request failures across CI/CD, Nomad, IIS, Grafana/OpenSearch, and providers
- **Tech:** C#/.NET, ASP.NET Core, Vue 3, MongoDB, SignalR, OpenAI-compatible APIs

---

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
  - Processed real-world driving datasets (inD, rounD, exiD) and evaluated generated trajectories

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 - Meeting captions and dictation ↗ | Actively developed personal fork

- Built Windows dictation/tray workflows; improved long-meeting queues, retries, and secure provider-key storage

LLM Gateway - Task-aware model routing ↗ | Independent open-source project

- Routes jobs across discovered models with provider fallback and safe pre-output streaming failover