Ruizhang Zhou

Full-stack KI-Plattform-Ingenieur - Streaming-Chat / Token- und Kosten-Governance / Mandantensicherheit / SSO

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

Erfahrung Software Engineer (AI Platform), KIconnect - RWTH IT Center

09/2024 - heute

- Chat-Pipeline (ASP.NET Core 9 + Semantic Kernel + Azure AI Inference/OpenAI): Streaming-Antworten, Titelgenerierung, strikte Rollenfolge, Handling von Content-Filtern (Reasoning- / Think-Tags).
- Tokenisierung und Kosten: Einheitliche Tokenerfassung (OpenAI, Llama 3, Gemma 3, Mistral, Qwen, DeepSeek); Python.NET-Bridge; Bild-Tokenisierung für VLMs; detaillierte Kostenereignisse inklusive Caching-Tokens; Reporting mit Heatmaps und Kennzahlen.
- Admin-Oberfläche (Vue 3 + Inertia): Modell- und Deploymentverwaltung, Parameter (Temperature / Top-p / Penalties), System-Prompts, Quoten; Soft-Delete und Restore; automatische Deaktivierung bei Credential-Entzug.
- Dokumente und Kontext: Upload und Extraktion via Microsoft KernelMemory; Token-Grenzprüfung je Deployment und Modell; bedarfsgerechte Einbindung von Dokumenttexten und Bildern; tägliche Aufräumjobs mit Hangfire.
- Sicherheit und Compliance: Shibboleth-SSO; MongoDB CSFLE (lokaler Master-Key) für API-Schlüssel; EU-Datenregion; Mandanten- und Rollenprüfungen.
- Echtzeit und Stabilität: SignalR-Broadcasting (Streaming-UI), resiliente HTTP-Pipelines via Polly, NLog; DevOps-Kontext mit Traefik und Nomad; Setup einer eingebetteten Python-Runtime.

Ausbildung

- RWTH Aachen University - M.Sc. Informatik (GenAI, LLM), Note 2,5 (2022.10 - 2024.06).
- RWTH Aachen University - B.Sc. Informatik, Note 2,2 (2019.10 - 2022.09).
- Tongji Universität - B.A. Germanistik (2014.09 - 2018.08).

Ausgewählte frühere Tätigkeiten

- Wissenschaftliche Hilfskraft - RWTH Lehrstuhl Embedded Software (2023.08 - 2024.03): Wartung der CPM Remote Web App; Echtzeit-Visualisierungen; CI/CD.
- Wissenschaftliche Hilfskraft - RWTH Lehrstuhl DBIS (2023.07 - 2024.03): LLM- / GNN- / Knowledge-Graph-Forschung; GPU-Cluster-Experimente; LLaMA- / Ollama-Chatplattform.

Tech-Stack C# / .NET 9, ASP.NET Core, MongoDB (CSFLE), Semantic Kernel, Azure AI Inference, OpenAI .NET, Microsoft.ML.Tokenizers, pythonnet, SkiaSharp, Hangfire, SignalR, Vue 3, Inertia.js, Vite, NLog, Traefik, Nomad, KernelMemory.