Ruizhang Zhou

Full-stack AI Platform Engineer - Streaming chat / Token & cost governance / Multi-tenant security / SSO

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

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

09/2024 - Present

- Chat pipeline (ASP.NET Core 9 + Semantic Kernel + Azure AI Inference/OpenAI): streaming responses, title generation, strict role alternation, content-filter handling (reasoning / think tags).
- Tokenization and costing: unified token counting (OpenAI, Llama 3, Gemma 3, Mistral, Qwen, DeepSeek); Python.NET bridge; image tokenization for VLMs; detailed cost events including cached tokens; reporting with heatmaps and metrics.
- Admin UI (Vue 3 + Inertia): model and deployment management, parameters (temperature / top-p / penalties), system prompts, quotas; soft-delete and restore; auto-deactivation on credential removal.
- Documents and context: upload and extraction via Microsoft KernelMemory; per-deployment and per-model token guardrails; on-demand inclusion of document text and images; daily cleanup with Hangfire.
- Security and compliance: Shibboleth SSO; MongoDB CSFLE (local master key) for API keys; EU data-processing region; tenant and role checks.
- Realtime and resilience: SignalR broadcasting (live UI), resilient HTTP via Polly, NLog; DevOps context with Traefik and Nomad; embedded Python runtime setup.

Education

- RWTH Aachen University - M.Sc. Computer Science (GenAI, LLM), Grade 2.5 (2022.10 - 2024.06).
- RWTH Aachen University - B.Sc. Computer Science, Grade 2.2 (2019.10 - 2022.09).
- Tongji University - B.A. German Language and Literature (2014.09 - 2018.08).

Selected previous roles

- Research Assistant - RWTH Chair of Embedded Software (2023.08 - 2024.03): CPM Remote Web App maintenance; real-time visualizations; CI/CD.
- Research Assistant - RWTH Chair of DBIS (2023.07 - 2024.03): LLM / GNN / Knowledge Graph research; GPU cluster experiments; LLaMA / Ollama chat platform.

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.