Mickey Szabo

AI Full-Stack Engineer

szabomickey182@gmail.com +49 17643620452 Isny im Allgäu, Germany

linkedin.com/in/mickey-szabo-290a8b3aa

Summary

AI Full-Stack Engineer with experience building production-grade web applications, AI-powered systems, and machine learning solutions across the full software development lifecycle. Strong background in Python, TypeScript, React, Next.js, FastAPI, cloud infrastructure, and modern AI/LLM technologies.

Experienced in designing scalable architectures, developing APIs and distributed systems, integrating large language models, building RAG and agent-based systems, and delivering reliable end-to-end applications. Comfortable working across frontend, backend, AI/ML, databases, infrastructure, testing, and deployment.

Strong focus on clean architecture, engineering quality, performance, scalability, and practical AI integration.

Experience

AI Full-Stack Engineer, *VLinde* 2024 – 2026 | Bavaria, Germany

- Designed and delivered AI-powered solutions that transformed complex business requirements into practical workflows using LLMs, retrieval-augmented generation, intelligent automation, and tool-based AI interactions.
- Tackled challenging AI problems involving unstructured information, context retrieval, inconsistent model outputs, and integration with existing business processes, focusing on making AI functionality reliable enough for real-world use.
- Built AI pipelines that connected models with application data, external services, and business logic, enabling AI systems to move beyond simple chat interfaces and perform useful actions within production workflows.
- Evaluated and improved AI behavior through prompt optimization, structured outputs, retrieval strategies, testing, and iterative model evaluation.
- Designed the supporting full-stack architecture using Python, FastAPI, TypeScript, React, and Next.js, connecting AI services with production-ready APIs, databases, and user interfaces.
- Worked directly with stakeholders to understand where AI could provide genuine business value, demonstrated working solutions, incorporated feedback, and refined the system toward the client's expectations.

AI & Full-Stack Engineer, *SERVOKAT GMBH* 2022 – 2024 | Isny im Allgäu, Germany

- Developed AI-driven application features focused on automating repetitive processes, extracting useful information, and improving how users interact with business data.
- Integrated LLM-based capabilities into existing software workflows, combining model reasoning with application context, structured data, retrieval, and deterministic business logic.
- Addressed reliability challenges in AI workflows by controlling model inputs and outputs, improving prompts and retrieval strategies, and introducing validation around AI-generated results.
- Built intelligent automation workflows that connected AI services with backend systems and external APIs, turning AI capabilities into actionable business processes rather than isolated model responses.

- Developed the full-stack layer supporting these AI systems using Python, FastAPI, React, Next.js, TypeScript, and SQL-based databases.
- Investigated failures across AI, backend, database, and integration layers, identifying root causes and implementing fixes that improved the stability of the overall solution.

Software Developer, *KORNHAAS* 2021 – 2022 | Isny im Allgäu, Germany

- Identified opportunities to replace manual and repetitive workflows with AI-assisted solutions and designed practical automation around real business requirements.
- Built AI-powered features using LLMs, intelligent information retrieval, structured generation, and application-specific context to solve problems that traditional rule-based software could not handle efficiently.
- Designed AI workflows with a strong focus on accuracy, controllability, and usability, balancing the flexibility of generative models with predictable application behavior.
- Connected AI capabilities to business systems through APIs, databases, and custom application logic, allowing users to interact with intelligent functionality directly within the existing product workflow.
- Built the supporting frontend and backend using React, Next.js, TypeScript, Python, and FastAPI, providing the full-stack foundation required to turn AI concepts into usable products.
- Debugged complex issues spanning model behavior, data pipelines, backend services, integrations, and frontend workflows, taking ownership of problems across the complete system.

Skills

Programming Languages
- Python
- TypeScript
- JavaScript
- Java
- Go
- Rust
- SQL
- Bash

Backend Engineering
- FastAPI
- Django
- Flask
- Node.js
- Express.js
- WebSockets
- Authentication & Authorization

Machine Learning & AI
- PyTorch
- TensorFlow
- scikit-learn
- NumPy
- Pandas
- Hugging Face Transformers
- Deep Learning, NLP
- Computer Vision
- Model Training
- Fine-Tuning
- Model Evaluation

Frontend Engineering
- React
- Next.js
- TypeScript
- HTML5
- CSS3
- Tailwind CSS

Generative AI & LLM Engineering
- Large Language Models (LLMs)
- RAG
- Embeddings
- Vector Search
- Prompt Engineering
- Function Calling
- Tool Calling
- AI Agents
- Multi-Agent Systems
- LLM Evaluation
- Model Benchmarking

Cloud & Infrastructure
- AWS
- Docker
- Kubernetes
- Linux
- Terraform
- GitHub Actions
- CI/CD
- Infrastructure as Code

Education

Bachelor of Science (B.Sc.) Computer Science 2019 – 2023 | Mannheim, Germany
University of Mannheim

Projects & Achievements

AI Technical Support Copilot

- Built an AI assistant that helps technical teams find solutions by searching product documentation, previous support cases, and internal knowledge.
- Implemented **RAG, semantic search, embeddings, and LLMs** to generate context-aware answers with relevant source references.
- Added conversation history, response validation, and feedback collection to improve answer quality over time.

Tech: Python, LLMs, RAG, LangChain, Hugging Face, PostgreSQL, pgvector, FastAPI

Intelligent Visual Inspection System

- Developed a computer-vision system for detecting product defects from inspection images and identifying cases that require manual review.
- Trained and evaluated deep-learning models and introduced confidence thresholds to separate reliable predictions from uncertain results.
- Built an inference API and web interface for image submission, prediction review, and inspection history.

Tech: Python, PyTorch, OpenCV, Computer Vision, CNNs, FastAPI, PostgreSQL, Docker

Multi-Tenant SaaS Platform

- Built a SaaS application supporting multiple organizations, user accounts, role-based permissions, dashboards, and business workflows.
- Designed the frontend, backend APIs, database structure, authentication, and authorization as a complete end-to-end system.
- Added caching, validation, error handling, and CI/CD to improve application reliability and deployment consistency.

Tech: React, Next.js, TypeScript, Python, FastAPI, PostgreSQL, Redis, Docker, AWS

Real-Time Collaboration Platform

- Developed a real-time web application where multiple users can work with shared data and receive updates instantly without refreshing the page.
- Implemented **WebSockets, event-driven communication, state synchronization, and concurrent update handling** across connected clients.
- Built the backend APIs, database layer, responsive frontend, authentication, and activity tracking.

Tech: React, Next.js, TypeScript, Node.js, WebSockets, PostgreSQL, Redis, Docker