Takumi Shimizu

Senior AI & Full Stack Engineer

tobeiokita35@gmail.com 81-6-6354-2252 Osaka, Japan

SUMMARY

Senior AI & Full Stack Engineer with 8+ years of experience building scalable SaaS platforms, AI applications, and cloud-based systems.
I have strong experience with LLMs, RAG, AI agents, LangChain, OpenAI APIs, FastAPI, React, Kubernetes, and AWS.
I focus on writing clean, reliable, and easy-to-maintain code. I can handle the full development process, from system design and backend development to AI integration, frontend development, deployment, and performance improvements.
My goal is to build practical and reliable software that solves real business problems, works well at scale, and provides real value to clients and users.

PROFESSIONAL EXPERIENCE

AI | Full-stack Engineer 03/2024 – 02/2026
Chinougijutsu Co., Ltd. Osaka, Japan
- Designed and deployed production AI applications using FastAPI, LangChain, OpenAI APIs, and vector databases.
- Built Retrieval-Augmented Generation (RAG) pipelines that improved response accuracy and reduced hallucinations for enterprise knowledge systems.
- Developed AI agents to automate multi-step business workflows, reducing manual effort and improving operational efficiency.
- Optimized cloud infrastructure and inference services, reducing API latency by 40% while lowering infrastructure costs.
- Led a team of eight engineers delivering scalable AI services and collaborating with product managers to ship customer-facing AI features.
- Collaborated with product managers and researchers to ship customer-facing AI features.

AI Developer 06/2022 – 02/2024
Rokken Osaka, Japan
- Built machine learning pipelines for real-time sentiment analysis with high prediction accuracy.
- Implemented privacy-preserving machine learning techniques for healthcare applications.
- Integrated AI-powered features into the company's SaaS platform, improving customer retention and user satisfaction.
- Developed backend services using Python, FastAPI, PostgreSQL, and Docker for scalable AI deployment.

Full Stack Developer 09/2020 – 06/2022
Fenrir Inc Osaka, Japan
- Developed a microservices-based e-commerce platform using React, Node.js, GraphQL, and PostgreSQL.
- Reduced page load time by 30% through frontend optimization and API performance improvements.
- Designed GraphQL APIs supporting high-volume production traffic.
- Implemented CI/CD pipelines using GitHub Actions and Docker, reducing deployment time from 3 hours to 25 minutes.
- Collaborated with cross-functional teams to integrate machine learning recommendations into customer-facing products.

Junior Full stack Developer 05/2018 – 08/2020
KINTO Technologies Osaka, Japan
- Built and maintained RESTful APIs using Django, Python, and PostgreSQL.
- Migrated legacy JavaScript applications to TypeScript, improving maintainability and reducing production bugs.
- Developed automated testing pipelines using Jest and Cypress, significantly increasing code coverage.
- Collaborated with frontend developers to deliver responsive web applications using React.

EDUCATION

Bachelor of Science in Computer Science 04/2014 – 03/2018
International Professional University of Technology Osaka, Japan
- Artificial Intelligence

SKILLS

AI & Machine Learning
GPT-5.6, Claude, Llama, Mistral, LangChain, LangGraph, RAG, AI Agents, OpenAI API, Vector Databases, Prompt Engineering, Fine-tuning (LoRA, PEFT)

Cloud & DevOps
AWS, Docker, Kubernetes, Terraform, GitHub Actions, Nginx, Linux, CI/CD

Frontend
React, Next.js, TypeScript, Tailwind CSS, Redux, React Query

Backend
Python, FastAPI, Django, Node.js, Express, Go, GraphQL, REST APIs

Databases
PostgreSQL, MongoDB, MySQL, Redis, Pinecone, Weaviate, Elasticsearch

LANGUAGES

English - Fluent

Japanese - Fluent