TARO SUZUKI

Senior AI/ML Engineer

Tokyo, Japan | suzukitaro008@gmail.com | github.com/omni-front | linkedin.com/in/tarosuzuki

PROFESSIONAL SUMMARY

Senior AI/ML Engineer with 7+ years of experience designing and deploying production machine learning systems and scalable software solutions. Expertise spans predictive modeling, NLP, LLM-based applications (RAG, agents, fine-tuning), and end-to-end ML lifecycle management. Proven ability to deliver high-impact solutions in enterprise environments across retail, manufacturing, and telecom sectors. Strong software engineering foundations with hands-on experience from backend systems to cloud-native AI pipelines.

PROFESSIONAL EXPERIENCE

Independent AI/ML Engineer | Independent

2025 — Present · Tokyo, Japan

- Designing LLM-based applications including RAG pipelines and agent-style orchestration systems for structured task execution
- Exploring parameter-efficient fine-tuning approaches (LoRA/QLoRA) and structured evaluation strategies to improve reliability and cost efficiency
- Building intelligent automation workflows integrating LLMs with production APIs and real-time data sources

Tech: Python, PyTorch, Hugging Face, LangChain, LangGraph, Vector DBs, MLflow, Docker

Senior AI/ML Engineer | BrainPad Inc.

2022 — 2024 · Tokyo, Japan

- Delivered predictive modeling and analytics solutions for enterprise clients in retail and manufacturing industries
- Led end-to-end ML development including feature engineering, model selection, validation, and deployment coordination
- Standardized experiment tracking workflows using MLflow and supported production evaluation and retraining processes
- Mentored junior engineers and led technical discussions with client stakeholders

Tech: LightGBM, XGBoost, TensorFlow, MLflow, Airflow, AWS SageMaker, Docker, SQL

AI/ML Engineer | ALBERT Inc.

2019 — 2021 · Tokyo, Japan

- Built supervised learning models for demand forecasting and customer behavior prediction in retail and telecom sectors
- Developed Japanese-language NLP classification systems using transformer-based architectures
- Designed preprocessing workflows, engineered time-series features, and implemented cross-validation strategies
- Implemented lightweight inference APIs and integrated ML services into enterprise systems

Tech: Python, Scikit-learn, PyTorch, Hugging Face, Pandas, Spark, FastAPI

Software Engineer | TIS Inc.

2017 — 2019 · Tokyo, Japan

- Developed and maintained enterprise financial systems as a backend software engineer
- Implemented server-side modules using Java and Spring Framework, developed REST APIs and transaction processing systems
- Contributed to performance improvements through SQL optimization and restructuring of batch-processing workflows

Tech: Java, Spring Framework, PostgreSQL, Oracle DB, AWS EC2/S3, Git, Linux

KEY PROJECTS

InterviewIQ — Forensic Interview AI

- PEACE-framework forensic interview AI achieving 10/10 compliance at ~750ms per turn (7.2x faster than baseline)
- 4-node LangGraph agentic pipeline with LoRA fine-tuned Phi-4 (3.8B) and Neo4j knowledge graph

AdFlow — GCP Ad Trafficking Automation

- End-to-end ad trafficking pipeline on Google Cloud with Vertex AI agents and human-in-the-loop approval
- Automated email monitoring, OMS validation, trafficking sheet generation, and ad server updates

ConvoStack — Real-Time Voice Agent Platform

- Real-time voice agent with Deepgram STT/TTS, FastAPI backend, and React frontend
- Low-latency bidirectional audio streaming with LLM-driven conversational logic

TECHNICAL SKILLS

AI / ML: PyTorch, TensorFlow, Hugging Face, LangChain, LangGraph, scikit-learn, MLflow, LightGBM, XGBoost

NLP & LLM: LLMs, RAG, Prompt Engineering, Transformers, Japanese NLP, LoRA / QLoRA, Embeddings, Vector DBs

Languages: Python (expert), Java, SQL, TypeScript, Bash

Cloud & Infra: AWS (SageMaker, EC2, S3), GCP (Vertex AI, Cloud Functions, Pub/Sub), Docker, Kubernetes, Airflow

Databases: PostgreSQL, Oracle DB, Neo4j, Vector Databases

Frameworks: FastAPI, Flask, Spring Framework, React, Next.js

EDUCATION

Master of Science (M.S.), Information Science and Technology

The University of Tokyo · 2015 — 2017 · GPA: 3.7 / 4.0

- Specialized in machine learning, statistical modeling, and computational data analysis
- Research: supervised learning under limited and imbalanced data conditions

Bachelor of Science (B.S.), Computer Science

The University of Tokyo · 2011 — 2015 · GPA: 3.6 / 4.0

- Strong foundations in algorithms, systems programming, and software engineering
- Capstone: predictive analytics system using SVM, Random Forest, and Logistic Regression

CERTIFICATIONS

Google Cloud Professional Machine Learning Engineer — Google Cloud, 2023

AWS Certified Solutions Architect — Associate — Amazon Web Services, 2022

JDLA Deep Learning for ENGINEER — Japan Deep Learning Association, 2020

LANGUAGES

Japanese (Native) **English** (Professional)