

Weibang Yin

www.mrbang.space

Tokyo, Japan

Email: mrweibang@outlook.com

Mobile: +81-080-7808-5883

SUMMARY

Software Engineer with 3+ years of experience building and operating production systems in **Python** (also Java, C++). Backend lead delivering scalable API and data services, and founder of a live **LLM/GenAI** product (RAG, agent orchestration, real-time inference) serving ~1,000 users. M.Sc. in Computer Vision & Robotics with deep-learning research in PyTorch. Passionate about the open-source AI/ML ecosystem and the ML lifecycle: training, evaluation, deployment, and monitoring.

EXPERIENCE

- Safie Inc.** Tokyo, Japan
Software Engineer Sep. 2025 – Present
 - **Backend Lead – Cloud Camera × POS Integration:** Led backend design for a new retail loss-prevention product integrating Safie cloud cameras with the Smaregi cloud POS platform, delivering theft and staff-misconduct detection on top of existing POS features for hundreds of retail, restaurant, and apparel stores.
 - **Design Ownership & Team Delegation:** Owned and designed 7 REST APIs, 1 database schema, and 2 batch jobs for the camera-POS linkage screens; broke work into tickets, delegated implementation across team members, and reviewed their code and technical designs.
 - **Business Impact:** Drove ~several million JPY of incremental quarterly revenue for the camera product line through the new integration.
 - **API Development & Reliability:** Developed new APIs and improved existing services (FastAPI, gRPC, REST, Celery, SQLAlchemy/Alembic, MySQL); handled production incidents end-to-end, performing root-cause analysis with Datadog and Grafana and shipping fixes plus preventive measures.
 - **Technologies:** Python, FastAPI, gRPC, REST, Celery, SQLAlchemy, Alembic, MySQL, AWS (ECS), Terraform, Docker, GitHub Actions, Datadog, Grafana
- Rakuten Group** Tokyo, Japan
Software Engineer – Rakuten Travel Apr. 2023 – Aug. 2025
 - **Monitoring & Automation:** Built and maintained internal monitoring dashboards visualizing traffic and system metrics across multiple domains, and designed a reporting-automation pipeline that extracted key metrics (QPS, CPU, memory) from Grafana/Jennifer and auto-generated PDF reliability reports.
 - **Load Testing & Performance Engineering:** Led load-testing efforts to ensure stability during high-traffic events (Super Sales); designed test scenarios, ran distributed load generation on GCP, and analyzed logs and metrics to identify and remove bottlenecks.
 - **CI/CD & Infrastructure Automation:** Designed CI/CD pipelines (Jenkins, Groovy), managed Kubernetes clusters for high availability, and automated provisioning with Ansible playbooks.
 - **Technologies:** Python, Java, Kubernetes, Jenkins, Ansible, Terraform, Prometheus, Grafana, Kafka, Elasticsearch, JMeter, GCP, Docker

PROJECTS & OPEN-SOURCE AI/ML WORK

- FastReply – Real-time LLM Interview-Assistant SaaS** Founder & Full-Stack Engineer
Solo-built product: ~1,000 users, 128 paying subscribers (Stripe); 1,300+ commits Jul. 2025 – Present
 - **Real-time GenAI Pipeline:** Designed and built a low-latency audio pipeline: browser mic → WebSocket → Google Cloud Speech (streaming STT) → OpenAI streaming LLM → TTS, delivering answer suggestions live during interviews.
 - **RAG & Agent Orchestration:** Implemented a Retrieval-Augmented Generation layer with **LangChain**, a Redis vector store, and sentence-transformers embeddings to ground each session's responses in the user's resume, and orchestrated prompt generation, evaluation, and grounding across microservices.
 - **Model Evaluation & Monitoring:** Added LLM-as-judge evaluation with a stronger model, session/latency warm-up, and a usage/credit-monitoring service to keep responses accurate and cost-controlled in production.
 - **Platform & Ops:** Architected 8+ Python/FastAPI microservices (gateway, auth, payments, email, STT/TTS, prompt-gen) with a Next.js/React frontend and Stripe billing; containerized with Docker Compose, wired CI/CD (GitHub Actions + Azure Pipelines) with security scanning, Alembic migrations, and automated DB backup/restore; deployed on GCP.

- **Technologies:** Python, FastAPI, LangChain, OpenAI API, RAG, Redis vector store, sentence-transformers, WebSocket, Google Cloud Speech, PostgreSQL, Next.js, React, TypeScript, Stripe, Docker, GCP
- **Reflection Removal Network (M.Sc. Research)** PyTorch, Deep Learning, Computer Vision
Apr. 2020 – Sep. 2022
 - Designed and trained a deep neural network (U-Net + ResNet) in **PyTorch** to remove reflections from images (e.g. glass glare, fences), including data pipeline, model training, and evaluation.

EDUCATION

- **The University of Tokyo** Tokyo, Japan
Apr. 2020 – Sep. 2022
 - M.Sc. in Computer Vision and Robotics*
- **Nanjing University** Nanjing, China
Sep. 2015 – Jun. 2019
 - B.Eng. in Telecommunication Engineering*

TECHNICAL SKILLS

Programming Languages: Python (primary), Java, C++, Go, TypeScript/JavaScript, SQL, Groovy

AI/ML & GenAI: PyTorch, Deep Learning, Computer Vision, LLM/GenAI, LangChain, OpenAI API, RAG, Embeddings & Vector Search, LLM Evaluation, Model Deployment & Monitoring

Backend & Frameworks: FastAPI, gRPC, REST, Celery, SQLAlchemy, Alembic, Spring Boot, Next.js, React

Data & Storage: PostgreSQL, MySQL, MongoDB, Redis (Vector Store), Elasticsearch, Kafka

Cloud & DevOps: AWS (ECS), GCP (GCE/GCS/Speech), Docker, Kubernetes, Terraform, GitHub Actions, Jenkins, CI/CD

Observability: Datadog, Grafana, Prometheus, ELK, Alertmanager

Languages: Chinese (Native), English (Business), Japanese (JLPT N1)