

YANG JIZHOU

Applied AI & Product Engineer

Tokyo, Japan | yjz.haruka@gmail.com | www.harukayang.com | github.com/YANG-Haruka | linkedin.com/in/yang-haruka

SUMMARY

Applied AI engineer and project lead who turns research prototypes into reliable, operated systems. At Corpy, I lead an end-to-end autonomous-driving system for a Japanese automaker, with ownership across architecture, model/data integration, Python-to-C++ transfer, inference formatting, and delivery. Outside work, I have built SaaS, OCR, RAG, and voice-AI products used by real businesses. Seeking Applied AI / AI Product / Forward-Deployed roles. Highly-Skilled Professional visa; permanent residency in application.

SKILLS

AI-Native Dev	Agentic coding (Claude Code, Codex), rapid technical research, image/video generation, AI slide-to-video demos — SOTA research → client-ready demo in days
Languages	Python, TypeScript, C++, SQL, Dart
AI / ML / Robotics	PyTorch, end-to-end autonomous driving, computer vision (detection, segmentation, tracking, person Re-ID), model deployment (ONNX / TensorRT), RAG, LLM agents, model evaluation, OCR / speech, ROS 2, OpenGL, CUDA
LLMOps / MLOps	Multi-provider model routing with config-driven fallback chains, adaptive rate limiting, code-validated structured output, hybrid dense+BM25 retrieval with reranking, model/prompt A-B evaluation on hand-built golden sets
Backend	FastAPI, PostgreSQL (multi-tenant row-level security), SQLAlchemy / Alembic, SQLite, Node.js, OpenAPI contract-driven APIs
Frontend / Mobile	React, TypeScript, Vite, Flutter
Cloud / DevOps	AWS (Lambda, DynamoDB, S3, CloudFront, CDK), GCP (Vertex AI, Cloud Run), Docker, Fly.io, CI/CD (GitHub Actions / GitLab CI), Linux

WORK EXPERIENCE

Corpy & Co., Inc. — Applied AI Engineer → Project Leader

Tokyo | May 2024 – Present

End-to-end autonomous-driving system for a Japanese automaker — research prototype to on-vehicle operation

- Led the overall system architecture, taking the system from research prototype to real-time operation on a vehicle.
- Participated in model training and owned data preprocessing/post-processing plus final inference-format conversion; contributed to inference optimization that cut latency ~70% at metric parity.
- Solved Python-to-C++ cross-language data transfer and introduced FlatBuffers to improve transfer efficiency and runtime performance.
- Promoted to Project Leader (team 4 → 7), owning requirements, weekly delivery, and module coordination — module ownership raised delivery efficiency ~50%; also supported early cross-team R&D for fisheye person Re-ID (segmentation vs B-box detection).

AMIT Co., Ltd. — Medical AI / Robotics Engineer

Apr 2022 – Apr 2024

Surgical navigation and robot-assisted ablation fusing pre-op CT with live ultrasound — joint research with Waseda University

- Owned arm kinematics and multi-space coordinate transforms (image ↔ patient ↔ arm base); ROS 2 auto-tracking control; validated on phantom and in clinical trials.
- Built a doctor-in-the-loop annotation workflow (~8k images) and trained a medical-image segmentation model to ~0.89 IoU, sharply reducing manual annotation work.
- Converted and accelerated model inference with ONNX / TensorRT — ~70% faster on constrained hardware (RTX 2060) while maintaining accuracy.

SELECTED PROJECTS

- EXA Restaurant** — multi-tenant restaurant-ops SaaS, built solo and used daily by a real Kawasaki restaurant (11 staff), on AWS. Keeping one restaurant's data invisible to another is enforced by the database itself rather than by remembering to filter in every query — 65 tables and 256 row-level-security policies, rebuilt from code and re-checked against a real Postgres on every CI run. 877 automated tests across API, web and mobile, and the release gate rebuilds a fresh instance from scratch and makes it pass its own checks before anything ships. Also built encrypted-NFC clock-in that cannot be cloned or replayed (NTAG 424 DNA).
- LinguaHaru / HaruYukaFlow** — production document AI: 16+ format translation with OCR/manga/audio modes (270+ GitHub stars), plus a three-model receipt-OCR and customs workflow operating a real cross-border business.
- Personal site (harukayang.com) + HaruAccompany** — my portfolio site and the AI chat on it, running on AWS with no always-on server: pages from a CDN, requests handled per-invocation by Lambda, data in DynamoDB, infrastructure described in code (CDK). The backend uses no third-party libraries — Node.js built-ins only. The chat looks up my background and projects before answering, and replies stream in as they are generated.

HACKATHONS

- Gemini Gym** — Google Cloud Gemini AI Agent Hackathon Tokyo (Jun 2026, team): backend & showcase; JS → Python engine port at 0-diff parity, Gemini via Google ADK on Vertex AI.
- FASHINI** — AI virtual try-on mirror, DevOps × AI Agent Hackathon 2026 (team StyleAI, ongoing): backend development; live on Google Cloud Run.

EDUCATION

M.Eng. in Mechanical and Intelligent Systems Engineering, The University of Electro-Communications	Tokyo Apr 2021 – Mar 2023
B.Eng. in Mechanical Design, Manufacturing & Automation, UESTC	Chengdu Sep 2015 – Jun 2019

ADDITIONAL

- Languages: Chinese (native), English (professional working), Japanese (conversational, advancing toward business) · Permanent residency in application.