

YANG JIZHOU

Applied AI & Product Engineer — Curriculum Vitae

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

PROFILE

Applied AI engineer and project lead specializing in taking systems from research prototype to customer-operated production. At Corpy, I lead an end-to-end autonomous-driving system for a Japanese automaker, with core ownership of architecture, model/data integration, Python-to-C++ transfer, inference formatting, and delivery; I also supported early cross-team R&D for fisheye-camera person Re-ID. Earlier, I built medical-AI and robotic-navigation systems; outside work, I operate production SaaS, document-AI, RAG, and voice systems. I am 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 demo production — SOTA research → working prototype → 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	LLMOps (multi-provider routing with config-driven fallback chains, adaptive concurrency and error-class-aware retries, schema-enforced output gated by code-side validators, semantically-keyed response caching, hybrid dense+BM25 retrieval with reranking, model/prompt A-B evaluation against hand-built golden sets), MLOps for inference and data (multi-model serving, human-in-the-loop annotation pipelines, production quality auditing)
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

EMPLOYMENT

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 and took the system from research prototype to real-time operation on a vehicle.
- Participated in part of the model training and owned the data preprocessing and post-processing pipeline.
- Solved cross-language data transfer from the Python model pipeline to the C++ runtime; introduced FlatBuffers and optimized the transfer path for lower overhead.
- Optimized final inference-format conversion so outputs reached the C++ side in the most suitable structure for downstream processing and presentation; contributed to inference optimization that cut latency ~70% at metric parity.
- Progressed from engineer to Project Leader (team 4 → 7), owning requirements, technical documentation, module coordination, and weekly delivery — module ownership raised delivery efficiency ~50%.

C++ · Python · PyTorch · FlatBuffers · cross-language data transfer · preprocessing/post-processing · Docker · GitLab CI

Fisheye-Camera Person Re-Identification — Cross-Team R&D 2026 | cross-project

- Supported early-stage, cross-team/cross-project R&D by comparing person segmentation and B-box detection as front-end approaches for fisheye-camera person Re-ID.

Python · PyTorch · person segmentation · B-box detection · person Re-ID

AMIT Co., Ltd. — Medical AI / Robotics Engineer Apr 2022 – Apr 2024
Surgical navigation / robot-assisted liver ablation: pre-op 3D CT fused with intra-op 2D ultrasound — joint research with Waseda University

- Owned the arm-kinematics core: multi-space coordinate transforms between ultrasound-image space, patient physical space, and arm base frame; ROS 2 auto-tracking control with force feedback, doctor-in-the-loop.
- Built a doctor-in-the-loop semi-automated annotation workflow (~7–8k images) and trained a medical-image segmentation model to ~0.89 IoU, sharply cutting manual annotation.
- Converted and accelerated model inference with ONNX / TensorRT — ~70% faster on an RTX 2060 while maintaining accuracy.
- Validated on a liver phantom and in human (clinical) trials.

PyTorch · medical-image segmentation · ONNX · TensorRT · ROS 2 · CT / ultrasound · image registration

SELECTED PERSONAL PROJECTS (PRODUCTION & OPEN SOURCE)

EXA Restaurant — multi-tenant restaurant-operations SaaS 2026 | solo, in production

- Used daily by a real Kawasaki restaurant: 11 staff clock in, check shifts, run approvals and receive payslips. Backend, web console, mobile app and deployment all built and operated solo.

- Keeping one restaurant's data invisible to another is enforced by the database itself — 65 tables and 256 row-level-security policies — rather than by remembering to filter in every query, where a single miss leaks data across tenants. CI rebuilds every policy from code and re-checks it against a real Postgres on each run. (The single-tenant pilot runs SQLite; switching is a deployment step, not a rewrite.)
- 877 automated tests across API, web and mobile. Before anything ships, the release gate rebuilds a fresh instance from scratch and makes it pass its own checks.
- Front and back end are generated from one shared API contract, so if the two ever disagree the build fails — a whole class of “the API changed and nobody told the frontend” bugs simply cannot happen.
- Encrypted NFC clock-in that cannot be cloned or replayed, so staff cannot punch in for each other (NTAG 424 DNA; built and tested, QR is the rolled-out path today).

FastAPI · PostgreSQL RLS · SQLAlchemy / Alembic · React · TypeScript · Flutter · NFC · AWS

LinguaHaru — open-source AI translator

2025 | 270+ GitHub stars

- One-click translation across 16+ document formats with layout preserved; desktop app + web console on one engine.
- OCR plugin (PP-OCRV6 / RapidOCR ONNX, ~60× GPU speedup in a child process), manga re-lettering mode, video/audio transcription, glossary enforcement, SHA-256-verified atomic cache.

Python · LLM · PP-OCRV6 · SenseVoice / Whisper · Docker · github.com/YANG-Haruka/LinguaHaru

HaruYukaFlow — OCR-driven inventory & customs suite

2025 | live, operating a real business

- First end-to-end client project: requirements → build → delivery → ongoing after-sales, for a cross-border resale business.
- Vision-LLM OCR with a three-model fallback chain (~55 brands' receipts); "the LLM only reads, code does all the reconciliation" — every number survives arithmetic cross-checks before entering inventory; Playwright crawlers auto-operate 11 EC platforms.

FastAPI · vision-LLM chain · Playwright · PyMuPDF · Fly.io · live: haruflow-yuka.fly.dev

HaruQQ — hybrid-retrieval RAG and persona memory

2025–2026 | personal use

- Dense+BM25 retrieval fused by RRF then cross-encoder reranked, over a 28k-document store. Built a labelled retrieval set and measured it: one threshold constant served two incomparable score scales and on the live reranked path discarded 77% of context (2.2 → 7.8 items kept after the fix); the RRF constant sat in a saturated region where k=60 and k=200 tied, and k=10 beat both at every cutoff; the cross-encoder cost 24 points of top-1 accuracy.

Personal site (harukayang.com) + HaruAccompany

2025–2026 | live

- My portfolio site and the AI chat that runs on it. Deployed on AWS with no always-on server: pages served from a CDN, requests handled per-invocation by Lambda, data in DynamoDB, and the whole infrastructure described in code (CDK) so it can be rebuilt with one command.
- The backend uses no third-party libraries at all — about 2,300 lines on Node.js built-ins, including hand-written request signing to reach the database.
- The chat looks up my background and project notes before answering, and replies stream in word by word as they are generated; a separate intent step handles contact, booking and resume requests.
- Budget alerts and cost-anomaly detection cap the damage from any unexpected spike.

Node.js (zero-dependency) · AWS Lambda · DynamoDB · CloudFront + S3 · CDK · RAG

HACKATHONS

Gemini Gym — Google Cloud / Gemini AI Agent Hackathon, Tokyo

Jun 2026 | team

- Natural-language health logging growing a living 3D avatar. Role: backend & showcase — ported the deterministic avatar-state engine JS → Python with byte-for-byte parity (0 diff); wired Gemini 2.5 Flash via Google ADK on Vertex AI; offline-proof demo; AI-produced showcase video.

FASHINI — AI virtual try-on mirror (team StyleAI)

2026 | judging in progress

- Smart mirror: gesture + voice drive an AI stylist agent (Gemini via Google ADK, Search grounding) proposing outfits from store inventory with real-time AI try-on (MediaPipe Pose + Decart Lucy). Role: 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, University of Electronic Science and Technology of China (UESTC)

Chengdu | Sep 2015 – Jun 2019

LANGUAGES & STATUS

- Chinese (native) · English (professional working) · Japanese (conversational, in active study toward business level).
- Tokyo-based; Highly-Skilled Professional visa; permanent residency in application.