

杨际舟

应用 AI & AI 产品工程师 — 完整履历

日本东京 | yjz.haruka@gmail.com | www.harukayang.com | github.com/YANG-Haruka | linkedin.com/in/yang-haruka

概要

应用 AI 工程师兼项目负责人，专注把研究原型推进为客户长期使用的生产系统。在 Corpy 主导某日本车企的端到端自动驾驶系统，核心负责整体架构、模型与数据集成、Python 到 C++ 的跨语言传输、推理格式转换和持续交付；同时参与 Cross-team、Cross-project 的鱼眼人物 Re-ID 前期 R&D。此前研发医疗 AI 与机器人导航，业余独立运营真实用户使用的 SaaS、文档 AI、RAG 与语音系统。求职方向为应用 AI、AI 产品工程与 FDE。在留资格：高度専門職1号，永住申请中。

专业技能

AI 原生开发 Agent 编码 (Claude Code、Codex)、快速技术调研、生图/生视频、AI 绘制幻灯片并渲染成演示视频——短时间调研 SOTA、落地原型并产出面向客户的完整 demo

语言 Python、TypeScript、C++、SQL、Dart

AI / ML / 机器人 PyTorch、端到端自动驾驶、计算机视觉 (检测/分割/跟踪/人物 Re-ID)、模型部署 (ONNX / TensorRT)、RAG、LLM Agent、模型评测、OCR / 语音、ROS 2、OpenGL、CUDA

LLMOps / MLOps LLMOps (多供应商路由与配置化兜底链、自适应并发与按错误类型重试、结构化输出经代码侧校验才放行、语义化响应缓存、稠密+BM25 混合检索与重排、对人工基准集做模型/提示词 A-B 评测)、面向推理与数据的 MLOps (多模型服务、人在环标注流程、生产质量审计)

后端 FastAPI、PostgreSQL (多租户行级安全)、SQLAlchemy / Alembic、SQLite、Node.js、OpenAPI 契约驱动

前端/移动 React、TypeScript、Vite、Flutter

云 / 运维 AWS (Lambda、DynamoDB、S3、CloudFront、CDK)、GCP (Vertex AI / Cloud Run)、Docker、Fly.io、CI/CD (GitHub Actions / GitLab CI)、Linux

工作经历

Corpy & Co., Inc. — 应用 AI 工程师 → 项目负责人

东京 | 2024.05 – 至今

某日本车企端到端自动驾驶系统——研究原型到实车运行

- 主导整体系统架构，将系统从研究原型推进到实车实时运行。
- 参与部分模型训练，负责完整的数据前处理与后处理流程。
- 解决 Python 模型侧到 C++ 运行端的跨语言数据传输问题，引入 FlatBuffers 并优化传输路径，降低数据交换开销。
- 优化最终推理阶段的格式转换，使输出以最适合 C++ 下游处理和展示的结构进行传递；参与推理优化，关键指标持前提下推理提速约 70%。
- 从工程师成长为项目负责人 (团队 4 → 7 人)，负责需求、技术文档、模块协调与每周交付——模块负责制使交付效率提升约 50%。

C++ · Python · PyTorch · FlatBuffers · 跨语言数据传输 · 数据前后处理 · Docker · GitLab CI

车载鱼眼相机人物 Re-ID — Cross-team 前期 R&D

2026 | Cross-project

- 作为 Cross-team、Cross-project 项目参与前期研究，比较人物分割与 B-box 检测两种前端方案在鱼眼人物 Re-ID 中的适用性。

Python · PyTorch · 人物分割 · B-box 检测 · 人物 Re-ID

AMIT株式会社 — 医疗 AI / 机器人工程师 (正式员工)

2022.04 – 2024.04

术前 3D CT 与术中 2D 超声融合的手术导航 / 机械臂辅助肝脏消融 —— 与早稻田大学联合研究项目

- 独立负责机械臂运动学核心：超声图像空间↔患者物理空间↔臂基座的多空间坐标变换；ROS 2 自动跟踪控制、力反馈、医生在环。
- 搭建医生在环的半自动标注流程 (约 7–8 千张)，训练医学图像分割模型至 IoU 约 0.89，大幅降低医生手工标注成本。
- 使用 ONNX / TensorRT 完成模型转换与推理加速——RTX 2060 上提速约 70%，同时保持识别效果。
- 肝脏模体与人体 (临床) 试验双重验证。

PyTorch · 医学图像分割 · ONNX · TensorRT · ROS 2 · CT/超声 · 图像配准

代表个人项目 (生产运营与开源)

EXA Restaurant — 多租户餐厅运营 SaaS

2026 | 独立开发，生产运营中

- 川崎一家真实餐厅每天在用：11 名员工打卡、查排班、走审批、领工资单。后端、Web 后台、手机 App 与上线运维全部由我一人完成。

- 「A 店看不到 B 店的数据」这条规则交给数据库自己执行——65 张表、256 条行级安全策略——而不是靠每次写查询都记得加过滤条件，因为漏一次就是跨店数据泄漏。CI 每次运行都会从代码重建全部策略，并在真实 Postgres 上重新核对。（单店试点目前跑 SQLite，切换是部署动作而非重写。）
- 877 个自动化测试覆盖后端 / Web / 手机端。发布前会从零重建一个全新实例，让它跑通自己的全部检查才放行。
- 前后端代码由同一份接口契约生成，两边一旦对不上就直接构建失败——「接口改了前端不知道」这类问题从根上不会发生。
- 加密 NFC 打卡：标签无法复制、刷卡记录无法重放，员工没法互相代打卡（NTAG 424 DNA；功能已完成并测试，店里目前铺开的仍是扫码）。

FastAPI · PostgreSQL 行级安全 · SQLAlchemy / Alembic · React · TypeScript · Flutter · NFC · AWS

LinguaHaru — 开源 AI 翻译器 2025 | GitHub 270+ stars

- 16+ 文档格式一键翻译且保留排版；桌面端 + Web 控制台共用一套引擎。
- OCR 插件（PP-OCRv6 / RapidOCR ONNX，GPU 提速约 60 倍、子进程隔离）、漫画重排模式、音视频转写、术语表强制、SHA-256 校验的原子缓存。

Python · LLM · PP-OCRv6 · SenseVoice/Whisper · Docker · github.com/YANG-Haruka/LinguaHaru

HaruYukaFlow — OCR 库存与报关自动化 2025 | 在线，服务真实业务

- 第一个端到端客户项目：需求收集 → 开发 → 交付 → 至今持续售后，服务跨境转售业务。
- 视觉 LLM OCR 三模型兜底链（约 55 个品牌小票）；「LLM 只读、代码核对」——每个数字过算术交叉验证才入库；Playwright 自动登录操作 11 个电商平台。

FastAPI · 视觉 LLM 链 · Playwright · PyMuPDF · Fly.io · 在线演示：haruflow-yuka.fly.dev

HaruQQ — 混合检索 RAG 与人格记忆 2025–2026 | 自用

- 2.8 万条库上做向量 + BM25 混合检索（RRF）接 cross-encoder 重排。建标注查询集实测后修掉三处：一个阈值常数被同时用于余弦与重排两种量纲，在实际启用的重排路径砍掉了 77% 上下文（修复后每查询留存 2.2 → 7.8 条）；RRF 常数落在饱和区（k=60 与 200 同分），改 k=10 后各截断点全面提升；并量出 cross-encoder 使 top-1 下降 24 个百分点。

个人网站 (harukayang.com) + HaruAccompany 2025–2026 | 在线

- 自己做的作品集网站，以及跑在上面的 AI 聊天助手。整站部署在 AWS，不需要一台一直开着的服务器：静态页面走 CDN 分发，请求由 Lambda 按次处理，数据存在 DynamoDB；全部基础设施用代码（CDK）描述，一条命令即可重建。
- 后端不依赖任何第三方库，只用 Node.js 自带能力写成（约 2300 行），连访问数据库的请求签名也是自己实现的。
- 聊天助手会先检索我的经历与项目资料再回答，回复边生成边逐字显示；另有意图识别，可处理联系、预约与简历索取等请求。
- 配置了预算告警与费用异常检测，给意外账单设了上限。

Node.js（零依赖）· AWS Lambda · DynamoDB · CloudFront + S3 · CDK · RAG · 向量检索

黑客松

Gemini Gym — Google Cloud / Gemini AI Agent 黑客松 (东京) 2026.06 | 团队

- 自然语言健康记录养成 3D 虚拟形象。负责后端与展示：确定性状态引擎 JS → Python 逐字节一致移植 (0 diff)；Google ADK 接入 Gemini 2.5 Flash (Vertex AI)；断网也能演示的兜底设计；AI 全流程制作展示视频。

FASHINI — AI 虚拟试衣智能镜 (StyleAI 团队) 2026 | 评审中

- 手势+语音驱动 AI 造型师 Agent (Gemini via Google ADK + Search grounding)，从店铺库存推荐搭配并实时 AI 试衣 (MediaPipe Pose + Decart Lucy)。负责后端开发；已上线 Google Cloud Run。

教育背景

电气通信大学 (东京) · 机械智能系统学专攻 · 工学硕士 2021.04 – 2023.03

电子科技大学 (成都) · 机械设计制造及其自动化 · 工学学士 2015.09 – 2019.06

语言与在留资格

- 中文 (母语)、英语 (工作语言)、日语 (日常会话，向商务水平进阶中)。
- 常驻东京；在留资格：高度専門職1号；永住申请中。