

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、JavaScript
- AI / ML** PyTorch、端到端自动驾驶、计算机视觉 (检测/分割/跟踪/人物 Re-ID)、ONNX / TensorRT 部署、混合 RAG、LLM Agent、评测 (Evals)、OCR / ASR
- 后端** FastAPI、SQLAlchemy、Alembic、PostgreSQL (行级安全 RLS)、SQLite、Node.js、OpenAPI 契约驱动、事务性 Outbox、幂等设计
- 前端/移动** React、Vite、TanStack Query、Vitest、Flutter、Riverpod
- 基础设施** Docker、CUDA、Fly.io、GCP (Vertex AI / Cloud Run)、GitHub Actions / GitLab CI、Playwright、pytest、ROS 2、OpenGL

工作经历

Corpy & Co., Inc. — 应用 AI 工程师 → 项目负责人 东京 | 2023.04 – 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，关键指标持平前提下推理提速约 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 – 2023.03

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

- 川崎真实餐厅每日在用：11 名员工扫码打卡、排班、审批、工资单分发。
- 模块化单体基座：12 个自注册模块；62 张租户表全量命令级 PostgreSQL 行级安全 (RLS)；不透明可撤销会话；幂等写入 + 事务性 Outbox；真实 Postgres 上屏障同步并发测试。
- 契约驱动：OpenAPI 快照 → 生成 TypeScript client，两侧字节级拒绝漂移；API/Web（三入口）/Flutter 共 698 个测试。

FastAPI · SQLAlchemy · Alembic · PostgreSQL RLS · React · TypeScript · Flutter · Fly.io

LinguaHaru — 开源 AI 翻译器

2025 | GitHub 1000+ 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 / HaruAccompany — 对话式 RAG 与记忆系统

2025–2026 | 在线

- 构建向量检索 + 关键词检索的混合 RAG，并加入结果融合与重排、查询扩展、元数据过滤、增量摄取以及用户/人格隔离。
- 将检索与长期记忆、OneBot 消息、工具/插件和主动消息连接；网站人格助手另含并行意图 Agent（联系/预约）、双语料风格检索、签名访客记忆与 SSE 流式响应。

Python · 向量数据库 · Embedding · 关键词检索 · 结果重排 · Node.js · SSE

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 – 2022.03
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。