

Rui Jin

Surigao del Norte, Philippines | +63 931 207 1477 | ruijin.developer@gmail.com | [Portfolio](#) | [GitHub](#) | [X](#) | [FILLI.ME](#)

Professional Summary

Senior AI/ML Engineer with 7+ years of experience building computer vision, on-device machine learning, and large language model (LLM) systems, along with the data pipelines and full-stack software around them. Built real-time pose and gesture tracking that runs on-device on iOS and Android phones at a 30 fps target, YOLO-based object detection for phones and drones, and 20+ full-stack and AI integration projects for clients on AWS and Google Cloud. Currently train and evaluate LLMs through prompt engineering, rubric-based evaluation, and hallucination review. Experienced in taking models from data pipelines and training through optimization, deployment, and real-world accuracy testing.

Skills

- **Languages:** Python, TypeScript, JavaScript, Java, C#, SQL, PHP
- **Machine learning & computer vision:** PyTorch, TensorFlow, scikit-learn, OpenCV, YOLO, Deep learning, Computer vision, Object detection, Pose estimation, Model evaluation
- **On-device & real-time ML:** On-device ML, Model optimization, INT8 quantization, ONNX, Real-time inference, Sensor fusion, Video encoding, iOS, Android
- **LLMs & NLP:** LLM evaluation, Prompt engineering, Rubric-based evaluation, Hallucination review, LLM integration, Natural language processing (NLP)
- **Data & MLOps:** ML pipelines, MLOps, ETL pipelines, NoSQL, Data analytics, Data visualization
- **Cloud & DevOps:** AWS, Google Cloud (GCP), Docker, Kubernetes, CI/CD, Git
- **Backend & APIs:** Node.js, Flask, Laravel, REST APIs
- **Frontend:** React, Next.js, Tailwind CSS

Work Experience

Outlier & AfterQuery

Mar 2025 - Present

AI Training & Evaluation Specialist

Help AI companies train and evaluate their large language models (LLMs) through the Outlier and AfterQuery platforms, improving the data those models learn from and are measured against.

- Write and refine prompts across classification, summarization, code generation, and multi-step reasoning tasks, including hard prompts built to expose where models fail.
- Score and compare model responses side by side against detailed rubrics for accuracy, reasoning, and instruction following, and write a clear justification for each rating.
- Flag hallucinations and factual errors, and rewrite weak answers into clear, correct reference responses.
- Review generated code for correctness and readability.

Nex

Dec 2022 - Mar 2025

AI/ML Engineer

Worked on the real-time body tracking behind Nex's motion-controlled games, where players control the game by moving in front of their phone's camera.

- Built and optimized lightweight computer vision models for real-time pose and gesture tracking on iOS and Android phones, targeting 30 fps with all inference running on-device.
- Worked on two of the hardest parts of the pipeline: encoding high-quality video on the phone and keeping the network connection stable.
- Wrote low-latency code for camera input, sensor fusion, and model inference, and connected it to the SDK our game designers and outside Unity developers used to build motion-controlled games.
- Set up data pipelines and MLOps tooling to train, evaluate, and deploy models, plus debugging tools to check tracking accuracy across different lighting and room setups.

Upwork

Jun 2021 - Nov 2022

Freelance Software Developer

- Completed 20+ full-stack and AI integration projects while keeping a 5-star client rating.
- Built and shipped web apps with React, Next.js, Flask, and Python on AWS and Google Cloud for SaaS, e-commerce, and analytics clients.
- Added LLM features and third-party API integrations to client products, and automated manual workflows along the way.

Customized Limited

Sep 2019 - Nov 2022

Full-Stack Developer

- Built, tested, and deployed front-end and back-end features for the company's platform using Java, Python, and Node.js.
- Used Python to connect the platform to the company's computer vision and machine learning models.
- Worked with full-time engineers and ML interns who processed and labeled the video data behind those models.

Cloudbreakr

Jun 2018 - Aug 2019

Software Engineer Intern

Hong Kong

- Built features for an influencer marketing platform, using PHP and Laravel on the back end and JavaScript and React on the front end.
- Added tracking charts to influencer profiles and built dashboards that showed brand clients the ROI of their campaigns.

Projects

On-Device Rep Counter

github.com/ruijindev0801/pose-rep-counter

PyTorch, MediaPipe, ONNX, INT8 quantization, TypeScript

Counts exercise reps live from a camera, right in the browser: a small PyTorch model, trained from scratch, reads MediaPipe body landmarks and runs on INT8 weights. On the RepCount-A benchmark it counts 62% of test videos within one rep; the best published pose-based method counts 56%.

Small LLM Judges vs. Human Rubric Grades

github.com/ruijindev0801/llm-judge-agreement

Python, llama.cpp, LLM-as-a-judge, Rubric evaluation

Tested whether a small open model on a 4 GB GPU can grade answers against detailed rubrics like a person. On 320 human-scored BiGGen-Bench answers, Qwen3.5 4B (4-bit) matched the human grades as closely as GPT-4 and Claude 3 Opus did. It graded harsher than people; a cross-validated calibration brought it within one point on 86% of answers.

EmbeddingGemma 2 Code Search Benchmark

github.com/ruijindev0801/embeddinggemma2-benchmark

Python, PyTorch, sentence-transformers, Hugging Face

An open benchmark of Google's EmbeddingGemma 2 against the first EmbeddingGemma on code search, at 768, 512, 256, and 128 dimensions, plus CPU speed and memory. At 256 dimensions, EmbeddingGemma 2 matched the original model at 768.

Marine Object Detection

Client work · NDA

Python, YOLO26, TensorFlow, OpenCV, Drone SDK

Real-time detection of marine objects for a marine company: kelp, plastic bottles, surfers, whales, boats, and containers. Built on YOLO26 as a stable model that runs on phones and drones.

Instagram Avatar Protection

Client work · NDA

Python, TensorFlow, OpenCV, Instagram API

Uses image recognition and watermarking to catch and prevent unauthorized reuse of users' Instagram profile photos.

SaaS Landing Page

trashlion.com

Next.js, React, Tailwind CSS, Resend

A fast, responsive landing page for a SaaS product, with email integration through Resend.

Education

City University of Hong Kong

May 2016 - Jan 2020

Bachelor of Computer Science

Hong Kong

- Graduated with honors, focusing on software development and AI
- Thesis: AI-driven web applications

Certifications

Certified AI/ML Engineer, micro1

May 2026