Hiro Tanaka

AI Engineer (Senior)

Contact

Email: hiro.tanaka@example.com
Phone: +81 70-1234-5678
Address: Tokyo, Japan
https://linkedin.com/in/hirotanaka | https://github.com/htanaka | https://hiroai.dev
Date of Birth: 1988-02-04
Nationality: Japanese

Profile

Senior AI Engineer specializing in LLM systems, evaluation, and optimization for latency and cost at scale.

Experience

Senior AI Engineer — NeuraWave (Tokyo, Japan)

Sep 2021 – Present

- Designed multi-tenant inference with KV-cache sharing.
- Built eval harness with rubric and human-in-the-loop review.

Tech: PyTorch, vLLM, Ray, CUDA, Weights & Biases

Machine Learning Engineer — CyberAgent (Tokyo, Japan)

Apr 2018 – Aug 2021

- Developed recommendation systems for a large-scale advertising platform.
- Deployed models using Docker and Kubernetes.

Tech: TensorFlow, scikit-learn, Docker, Kubernetes

Projects

EvalLab — Python, Ray, Weights & Biases

Automated LLM eval pipeline with bias and safety checks.

https://github.com/htanaka/evallab

Feb 2022 – Present

Education

Master — University of Tokyo

Faculty: Computer Science • Major: Artificial Intelligence • CGPA: 3.8

2012 – 2014

Bachelor — Kyoto University

Faculty: Information Science • Major: Information Science • CGPA: 3.7

2008 – 2012

Skills

- PyTorch
- Distributed Training
- Optimization
- RAG
- Prompt Engineering
- Evaluation
- MLOps

Awards

Innovation Award — NeuraWave, 2023

Latency reduction project

Certificates

- AWS Machine Learning – Specialty

Languages

- Japanese — Native
- English — Fluent