YUKI NAKAMURA

yuki.nakamura@gmail.com (+1) 650 555 0123 Palo Alto, USA Research

WORK EXPERIENCE

Principal ML Infrastructure Engineer Apr 2024 – Present

OpenAI (San Francisco, USA)

- Design training infrastructure for large language models with billions of parameters (Python, CUDA, C++)
- Optimize distributed training pipelines reducing training time by 40% and costs by $2M annually

Staff ML Engineer Jul 2022 - Mar 2024

Meta AI (Menlo Park, USA)

- Built PyTorch-based training framework used by 500+ ML researchers (Python, PyTorch, C++)
- Implemented mixed-precision training and model parallelism for LLaMA models

Senior ML Infrastructure Engineer Feb 2020 - Jun 2022

NVIDIA (Santa Clara, USA)

- Developed CUDA kernels and optimizations for deep learning frameworks (CUDA, C++, Python)
- Created profiling tools for GPU performance analysis

ML Engineer Sep 2017 - Jan 2020

Cruise Automation (San Francisco, USA)

- Built ML training pipelines for autonomous vehicle perception systems (TensorFlow, Python)
- Optimized inference engines for real-time object detection on embedded systems

Research Engineer Jan 2015 - Aug 2017

Stanford AI Lab (Stanford, USA)

- Conducted research on computer vision and neural architecture search (Python, Caffe)

TECHNICAL SKILLS

Languages: Python, C++, CUDA, Rust, Assembly

ML Frameworks: PyTorch, TensorFlow, JAX, Triton, ONNX

Infrastructure: Kubernetes, Ray, Horovod, DeepSpeed, Slurm

Core Strengths: ML Systems, GPU Programming, Distributed Training

EDUCATION

Stanford University (Stanford, USA) Dec 2014

Ph.D., Computer Science - Artificial Intelligence

Thesis: Efficient Neural Network Training on Heterogeneous Systems

University of Tokyo (Tokyo, JPN) Mar 2010

Bachelor of Engineering (B.Eng.), Computer Science

Valedictorian

ADDITIONAL SKILLS & INTERESTS

- **Skills:** Performance Optimization, Compiler Design, HPC
- **Interests:** Go (Board Game), Mountain Biking, AI Safety, Cyberpunk 2077, Westworld