Video Introduction Script Opening

Opening

Hello, I'm Yuki Nakamura, and I'm based in Palo Alto, California. I hold a Ph.D. in Computer Science specializing in Artificial Intelligence from Stanford, and a Bachelor's from the University of Tokyo where I was valedictorian. I've been working on machine learning infrastructure for over 10 years. For me, ML infrastructure isn't just about running models—it's about building the systems that enable researchers to push the boundaries of what's possible with AI.

Recent Experience

Most recently, I'm a Principal ML Infrastructure Engineer at OpenAI, where I design training infrastructure for large language models with billions of parameters. In this role, I optimize distributed training pipelines, reducing training time by 40% and saving $2M annually in compute costs, while enabling our researchers to experiment faster and more efficiently.

Leadership Philosophy

My leadership style is research-enabling and efficiency-focused. I believe infrastructure teams should be force multipliers for researchers—our job is to remove friction and make impossible experiments possible. I've learned that the best infrastructure is invisible—researchers shouldn't need to think about distributed systems, they should focus on their models.

Problem-Solving Approach

What distinguishes me is my combination of deep systems expertise and ML understanding. At Meta AI, I built PyTorch-based training frameworks used by 500+ researchers, implementing mixed-precision training and model parallelism for the LLaMA models. I always keep in mind that infrastructure choices directly impact research velocity—a 10% speedup in training means researchers can run 10% more experiments.

Full-Stack ML Infrastructure Experience

I've worked across the entire ML infrastructure stack: from CUDA kernel optimization and GPU programming, through distributed training frameworks and model serving, to cluster orchestration and resource allocation. My work has spanned everything from low-level performance optimization to high-level API design.

I believe strongly in measuring everything and optimizing systematically. In ML infrastructure, intuition often misleads—you need profiling data to identify real bottlenecks.

Technical Expertise

Over the past five years, I've primarily worked with Python for ML frameworks, C++ and CUDA for performance-critical components, and Rust for systems programming where

safety matters. I'm expert in PyTorch and TensorFlow internals, and I understand distributed training frameworks like DeepSpeed and Horovod deeply.

But my real strength is in performance optimization for ML workloads. I love tackling questions like: How do we minimize communication overhead in distributed training? How do we maximize GPU utilization? How do we make training numerically stable at large scale? These low-level optimizations compound to enable breakthrough research.

Personal Interests

Outside of work, I contribute to PyTorch and other open-source ML frameworks, follow AI safety research closely, and enjoy mountain biking in the Bay Area hills. I also play Go at a competitive level, which I find shares interesting parallels with strategic thinking in system design.

Closing

You can find my research publications and open-source contributions on my personal website, which I've included in my application. I'd be excited to discuss ML systems, GPU optimization, or the infrastructure challenges of training frontier AI models.