Shriyansh Singh

+1-930-333-5141 | shriyansh.singh24@gmail.com | LinkedIn

SUMMARY

ML Systems Engineer specializing in **distributed LLM training** and **inference optimization**. Expert in developing scalable **ML infrastructure** that enables advanced research in **RLHF**, instruction tuning, and multi-modal models.

PROFESSIONAL EXPERIENCE

ML Systems Engineer *April 2024 - Dec 2024*

Hyphenova AI *Los Angeles, CA*

- **Architected** a **distributed training framework** using **PyTorch FSDP** and **DeepSpeed** that scaled to 70B parameter models across 128 A100 GPUs, reducing training time by 63%
- **Implemented** custom **CUDA kernels** and integrated **FlashAttention-2** that improved inference throughput by 2.8x while reducing memory footprint by 42%
- **Designed** an end-to-end **RLHF pipeline** with distributed reward modeling and **PPO training** that improved alignment scores by 37% while maintaining training stability
- **Collaborated** with ML researchers to optimize model architectures and training recipes, accelerating experimentation cycles by 4.2x through parallel evaluation frameworks

Deep Learning Infrastructure Engineer *May 2022 - Oct 2022*

Enterprise Business Technologies AI Lab *Mumbai, India*

- **Developed multi-node training infrastructure** using **PyTorch** and **Horovod** that enabled efficient fine-tuning of transformer models on distributed hardware
- **Engineered** model optimization techniques including quantization, knowledge distillation, and **gradient checkpointing** that reduced memory requirements by 56%
- **Built** a comprehensive **inference serving platform** with dynamic batching and tensor parallelism that achieved sub-100ms latency at scale
- **Created** detailed performance profiling tools that identified and resolved bottlenecks in data loading, gradient computation, and communication patterns

ML SYSTEMS PROJECTS

Distributed RLHF Training Platform | *PyTorch, CUDA, JAX, Ray, Transformers, FlashAttention* *May 2024 – Dec 2024*

- **Designed** a scalable system for **Reinforcement Learning from Human Feedback** that enabled efficient preference learning and policy optimization for LLMs
- **Implemented** custom **distributed data loading** and **sharding** techniques that improved GPU utilization by 83% during reward model training
- **Optimized** the **PPO training loop** with mixed precision, gradient accumulation, and efficient KL divergence computation that maintained training stability at scale

High-Performance LLM Inference Engine | *C++, CUDA, TensorRT, Triton, PyTorch, FasterTransformer*

- **Engineered** a highly optimized inference engine with **continuous batching** and **kernel fusion** that achieved 7.2x higher throughput than standard implementations
- **Implemented tensor parallelism** and **key-value caching** strategies that enabled serving 70B+ parameter models with minimal latency on commodity hardware
- **Developed** advanced **quantization techniques** including AWQ and GPTQ that reduced model size by 75% while preserving 96% of full-precision performance

TECHNICAL EXPERTISE

ML Systems: Distributed Training, **FSDP**, **DeepSpeed**, **Tensor Parallelism**, **RLHF**, **Model Sharding**, Data Parallelism
ML Frameworks: PyTorch, **Transformers**, JAX, **FlashAttention**, FasterTransformer, TensorRT, TorchDynamo
Programming: Python, **C++**, **CUDA**, OpenCL, Shell Scripting
Infrastructure: Kubernetes, **Ray**, Slurm, Docker, **MLflow**, Weights & Biases, NCCL
LLM Techniques: Instruction Tuning, **RLHF**, **Tool Use**, Multimodal Training, LoRA, QLoRA, Model Merging
Computer Science: Distributed Systems, High-Performance Computing, Memory Optimization, Network Communication

EDUCATION

Indiana University Bloomington Aug 2023 – May 2025

Master of Science in Data Science *Indiana, United States*

- Research Focus: Distributed Systems for Large Language Models, High-Performance ML Computing