Shriyansh Singh

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

SUMMARY

GenAI Research Scientist with proven track record advancing **large language models** through novel **fine-tuning techniques**, **RLHF optimizations**, and architectures that **democratize AI access** across industries.

PROFESSIONAL EXPERIENCE

Research Scientist, Generative AI Lab *April 2024 - Dec 2024*
Hyphenova AI Research *Los Angeles, CA*

- **Pioneered** novel **parameter-efficient fine-tuning methods** that reduced computational requirements by 83% while maintaining 97% of model performance, resulting in company-wide adoption
- **Architected** scalable training infrastructure for 70B parameter models using distributed learning techniques that accelerated research iterations by 4.2x, enabling rapid empirical validation
- **Developed** state-of-the-art **RLHF pipeline** incorporating innovative preference modeling approaches that improved benchmark scores by 28% across complex reasoning tasks
- **Published** research on **context window extension techniques** at leading NLP conference, demonstrating efficient methods for extending token context without proportional compute increase

AI Research Engineer *May 2022 - Oct 2022*
Enterprise Business Technologies Research Division *Mumbai, India*

- **Investigated** efficient techniques for knowledge insertion into pre-trained LLMs that improved factual accuracy by 36% while requiring only 2% additional parameters
- **Implemented** novel **retrieval augmentation architecture** that dynamically selected appropriate knowledge sources, resulting in 42% improved accuracy on domain-specific tasks
- **Conducted** systematic empirical studies comparing different **fine-tuning approaches** across model scales ranging from 1B to 20B parameters, identifying optimal techniques for different scales
- **Contributed** to open-source GenAI frameworks by developing efficient implementations of research techniques that have been adopted by over 5,000 practitioners globally

RESEARCH PROJECTS

Multi-modal Retrieval-Augmented Generation System | *PyTorch, Transformers, JAX, Vector Databases*

- **Designed** and **implemented** a novel architecture for integrating multi-modal retrieval with generative models that achieved state-of-the-art performance on three industry benchmarks
- **Pioneered distributed training methodology** that enabled efficient fine-tuning of 100B+ parameter models on diverse data sources with 78% reduced memory footprint
- **Formulated** mathematical framework for analyzing retrieval quality impact on generation performance, providing theoretical foundations for optimal retrieval corpus construction

Reinforcement Learning from Human Feedback (RLHF) at Scale | *PyTorch, TRL, Transformers, Ray, MLflow, Accelerate*

- **Created** an **end-to-end RLHF pipeline** for efficiently aligning pre-trained models with human preferences, incorporating innovations in reward modeling that improved alignment by 43%
- **Invented** a novel approach for preference data synthesis that reduced labeled data requirements by 65% while maintaining alignment quality comparable to fully human-labeled datasets
- **Engineered training infrastructure optimizations** that reduced cost per trained model by 58% through advanced parallelization techniques and optimized hyperparameter search

TECHNICAL SPECIALIZATIONS

Research Areas: Large Language Models, **Fine-tuning**, **RLHF**, **Retrieval-Augmented Generation**, **Multimodal Learning**, Distillation, Scaling Laws
Deep Learning: PyTorch, **JAX**, **Transformers**, **Accelerate**, **DeepSpeed**, FSDP, TRL, HuggingFace, Triton
Distributed Computing: Ray, Spark, **MLflow**, **Weights & Biases**, SLURM, **Kubernetes**, Parameter Server
Languages: Python, C++, CUDA, **SQL**, **Bash**, Julia
Data Processing: Distributed ETL, **Data Filtering**, **Tokenization**, **Vector Databases**, **Embeddings**, Feature Engineering

EDUCATION

Indiana University Bloomington Aug 2023 – May 2025
Master of Science in Data Science *Indiana, United States*

- Research Focus: Advanced Training Methodologies for Large Language Models, Parameter-Efficient Fine-Tuning Techniques