Shriyansh Singh

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

SUMMARY

Software Engineer specializing in **AI/ML systems** with expertise developing **scalable GenAI solutions**, optimizing **ML infrastructure**, and implementing **data pipelines** that drive innovation and business impact.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer *April 2024 - Dec 2024*
Hyphenova AI *Los Angeles, CA*

- **Designed** and **implemented** production-ready **GenAI solutions** utilizing LLMs that processed 5TB of unstructured data with 99.7% reliability, enhancing customer engagement by 42%
- **Orchestrated** end-to-end **ML pipelines** using **Airflow** and **Kubernetes** that automated model training, evaluation, and deployment, reducing time-to-production from weeks to days
- **Engineered** modular **Python** libraries for **model optimization** and **feature engineering**, improving inference latency by 68% while maintaining prediction accuracy
- **Collaborated** with cross-functional teams through comprehensive code reviews and documentation, ensuring compliance with best practices and system architecture standards

AI/ML Systems Developer *May 2022 - Oct 2022*
Enterprise Business Technologies Pvt Ltd *Mumbai, India*

- **Developed** distributed **ML training infrastructure** using **TensorFlow** and **PyTorch** that processed multi-modal data, accelerating experimentation cycles by 45%
- **Implemented** CI/CD pipelines with **Git** and **CircleCI** for **ML model versioning** and automatic testing, ensuring reproducibility and code quality
- **Constructed** robust data preprocessing components using **Python** and **SQL** that handled data cleaning, feature extraction, and transformation for ML model training
- **Formulated** technical specifications from ambiguous business requirements into actionable ML system architecture designs with clear acceptance criteria

PROJECTS

Large Language Model Fine-tuning System | *PyTorch, CUDA, Hugging Face, Docker, AWS* *Jan 2024 – Apr 2024*

- **Architected** an end-to-end system for fine-tuning foundation models on domain-specific data using **PyTorch** and **Hugging Face** transformers, achieving 87% task accuracy
- **Implemented** distributed training infrastructure with **CUDA** optimization techniques that reduced training time by 65%, enabling rapid experimentation
- **Designed** model compression and quantization workflows that decreased model size by 75% while maintaining 92% of original performance for efficient deployment

Real-time ML Inference API | *TensorFlow Serving, FastAPI, Kubernetes, Prometheus, Google Cloud* *Sep 2023 – Dec 2023*

- **Engineered** scalable ML inference service using **TensorFlow Serving** and **FastAPI** that efficiently processed 10,000+ prediction requests per second with ¡100ms latency
- **Containerized** the application with **Docker** and orchestrated deployment on **Kubernetes** clusters, ensuring high availability and automatic scaling
- **Integrated** comprehensive monitoring with **Prometheus** that tracked model drift, prediction latency, and system health metrics, enabling proactive issue detection

SKILLS & CERTIFICATIONS

Programming: Python, **C++**, **SQL**, **Go**, Shell Scripting, YAML, JSON
ML/AI Technologies: TensorFlow, **PyTorch**, **Hugging Face**, **scikit-learn**, LangChain, Weights & Biases
Cloud & Infrastructure: Google Cloud, **Kubernetes**, **Docker**, **Terraform**, AWS (Lambda, S3, SageMaker)
Data Processing: Airflow, **Spark**, Kafka, Beam, **TensorFlow Data Validation**, NumPy, Pandas
Software Engineering: System Design, Microservices, **RESTful APIs**, gRPC, **CI/CD**, **Git**
Monitoring: Prometheus, Grafana, **DataDog**, Google Cloud Monitoring

EDUCATION

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

- Relevant Coursework: Machine Learning Systems, Deep Learning, Natural Language Processing, Cloud Computing, Distributed Systems
- GPA: 3.8/4.0