Rishi Chhabra

LinkedIn | GitHub | Portfolio
Jersey City, NJ | +1 (201) 423-8684 | Rishi.Chhabra@outlook.com

Summary

Machine Learning Engineer specializing in GenAI, RAG Architectures, and MLOps. Experience building scalable production AI systems using Weaviate/Pinecone, deploying models on AWS/Kubernetes, and optimizing inference latency. Proven track record migrating legacy infrastructure to serverless architectures with high availability and cost efficiency.

Technical Skills

Machine Learning & GenAI: PyTorch, TensorFlow, Transformers, LangChain, RAG, LLMs (Llama 3, GPT-4), vLLM, LoRA/QLoRA, Quantization, Prompt Engineering (Chain-of-Thought, ReAct), CV, NLP
Vector Databases: Weaviate, Pinecone, ChromaDB, AWS Vector Search
Cloud & DevOps: AWS (SageMaker, Lambda, EC2, S3, EMR), Docker, Kubernetes, CI/CD, Terraform, Nginx, gRPC
ML Tools & Tracking: MLflow, Weights & Biases (W&B), TensorBoard
Programming: Python, C++, SQL, JavaScript, Dart

Experience

Ariesview — Jersey City, NJ
AI Engineer Intern — *Sep 2025 – Present*
- Architecting RAG system for internal document processing, integrating Weaviate for hybrid search
- Implementing semantic chunking and re-ranking algorithms, improving accuracy by 40% and reducing hallucinations
- Optimizing CI/CD pipelines and backend caching using Redis/Valkey, cutting deployment time by 30%

Incuwise — Delhi, India
Software Development Engineer I — *Feb 2024 – Aug 2024*
- Migrated legacy infrastructure to serverless AWS architecture, improving scalability by 60% and ensuring 99.9% uptime
- Engineered high-throughput backend services using Node.js/Flutter, reducing API response time by 15% for 10,000+ users
- Launched 4 production applications with optimized high-concurrency performance, improving engagement by 25%

Projects

End-to-End MLOps Pipeline
ML Engineering | *AWS SageMaker, Kubernetes, MLflow*
- Automated ML lifecycle deployment using containerized workflows and Kubernetes orchestration
- Integrated MLflow for experiment tracking, reducing manual intervention by 80% and time-to-production by 60%

RAG-Based Q&A System
GenAI & RAG | *Python, LangChain, Pinecone, OpenAI*
- Engineered RAG system processing 10,000+ documents, achieving 85% accuracy and ¡2s response latency
- Implemented hybrid search algorithms using Pinecone for balanced keyword/semantic search performance

Semantic Search Engine
Information Retrieval | *Weaviate, BERT, FastAPI*
- Developed semantic search platform for 50,000+ docs with Weaviate, improving relevance scores by 70%
- Provided high-performance REST APIs for client dashboards with 99.5% uptime

Multimodal AI Assistant
Computer Vision & NLP | *PyTorch, Transformers (CLIP), AWS*
- Deployed Visual Q&A chatbot using CLIP models, achieving 95% accuracy on custom validation dataset
- Scaled to 1,000+ daily interactions by containerizing with Docker and deploying on AWS EC2

Education

Stevens Institute of Technology — Hoboken, NJ
M.S. in Machine Learning — *Class of 2026*

Central University of Haryana — India
B.Tech. in Computer Science — *Class of 2024*

Certifications

AWS Certified Machine Learning Engineer - Associate

Databricks Certified Developer for Apache Spark - Associate

AWS Certified Developer - Associate