Rishi Chhabra

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

PROFESSIONAL SUMMARY

AI/ML-focused software engineer specializing in designing and deploying intelligent systems (LLMs, semantic search, vector databases) with cloud-native MLOps. Proven ability to build scalable, privacy-preserving solutions for web, mobile, and enterprise applications.

EDUCATION

Stevens Institute of Technology Hoboken, NJ **Sep 2024 - Present**
Master of Science, Machine Learning

Central University of Haryana Haryana, India **Sep 2020 - May 2024**
Bachelor of Technology, Computer Science

Certifications

- **AWS Certified Developer – Associate**: Validates proficiency in developing, deploying, and debugging cloud-based applications using AWS. Demonstrates knowledge of AWS services, SDKs, CI/CD, and security best practices.

TECHNICAL SKILLS

- **Programming Languages**: Python, Java, C++, JavaScript/TypeScript, Dart, SQL, R
- **Machine Learning & AI**: TensorFlow, PyTorch, scikit-learn, Keras, OpenCV, NLTK, Hugging Face Transformers (LLMS)
- **Web & Mobile Development**: Flutter, React.js, Node.js (Express), GraphQL, Flask, Django, RESTful APIs
- **Cloud & DevOps**: AWS (Lambda, EC2, S3, DynamoDB, Rekognition, Bedrock), Microsoft Azure (AI services), Docker, Kubernetes, Terraform, CI/CD (GitHub Actions, Jenkins), MLflow, Apache Airflow
- **Databases & Tools**: PostgreSQL, MongoDB, MySQL, Firebase, Vector DBs (Pinecone, Milvus, FAISS)

PROFESSIONAL EXPERIENCE

Incuwise **Feb 2024 - Aug 2024**
Senior Flutter Developer *Noida, India*

- Developed and launched 4 cross-platform mobile applications using Flutter and Node.js backends; led migration of server infrastructure from Linode to AWS, significantly improving scalability and reliability.
- Enhanced user engagement and transaction security by integrating Firebase Cloud Messaging, Google Maps APIs, Stripe payment processing, and OAuth authentication into mobile apps.
- Collaborated with QA and design teams in an Agile environment to deliver high-quality features on schedule, applying test-driven development and CI/CD best practices.

KEY PROJECTS

LLM Fine-Tuning Pipeline on AWS **2025**

- Implemented an AWS-based pipeline to fine-tune large language models on domain-specific data using SageMaker for distributed training on GPU instances.
- Optimized the workflow using spot instances and MLflow tracking to achieve enhanced model accuracy while controlling costs.

End-to-End MLOps Pipeline **2025**

- Architected a fully automated ML workflow on AWS using Docker, Kubernetes, and Apache Airflow for CI/CD and orchestration.
- Implemented MLflow for model versioning, automated drift detection, and retraining triggers; achieved a 60% reduction in deployment time and improved model reliability.

RAG Q&A System for Internal Docs **2025**

- Built an enterprise-grade Retrieval-Augmented Generation system leveraging Azure AI Search and AWS Bedrock LLMs.
- Created data ingestion and semantic chunking pipelines to enable high-accuracy vector retrieval over internal documentation; exposed secure APIs to power web and mobile Q&A interfaces.

LLM-Powered Semantic Search Engine **2025**

- Designed a semantic search platform using LLM embeddings and vector databases (FAISS, Milvus, Pinecone) for contextual search over unstructured data.
- Automated document indexing and semantic matching, significantly improving search relevance and speed over traditional keyword methods; provided REST APIs for integration with dashboards and apps.

Local Kubernetes LLM Inference **2025**

- Deployed a local Kubernetes cluster on a MacBook (via Docker Desktop/KIND) to host LLM inference services at the edge.
- Configured pods with optimized resource settings to run transformer models efficiently on-device, enabling low-latency, offline NLP inference without cloud dependency.

Multimodal AI Assistant **2025**

- Built a chatbot that processes both text and images using multimodal models for scene descriptions and visual question answering.
- Utilized Hugging Face Transformers (CLIP, Vision-LLM models), PyTorch, FastAPI for API development, and deployed on AWS EC2 with Docker and Kubernetes.