Rishabh Shah

rishabh13prof@gmail.com | (984) 259-5605 | linkedin.com/in/rishabhshah133 | github.com/rishabhshah13

I design machine learning systems that think fast, work quietly, and amplify human potential

Work Experience

Nokia Bell Labs, Murray Hills, NJ *Jun 2024 – Dec 2024*
Machine Learning Engineer Co-op

- Built and deployed a RAG chatbot for financial analysts, enabling complex, free flow QA queries over 10M+ records.
- Engineered data pipelines for document ingestion, embedding generation, and vector indexing for real-time retrieval.
- Developed a query understanding engine to decompose user questions, route sub-queries to vector & document DBs
- Fine-tuned Llama 3.1 using QLoRA to enhance query understanding and free flow QA for complex financial analysis in RAG workflows.

Duke University, Durham, NC *Oct 2023 – Sep 2024*
Research Assistant

- Architected and implemented a real-time, voice-enabled healthcare chatbot for VR nurse training, integrating Llama 2 7B for contextual dialogue and Whisper API for low-latency speech-to-text within a multi-service pipeline.
- Optimized LLM inference pipeline by implementing continuous batching and key-value (KV) caching, significantly reducing time-to-first-token and achieving sub-500ms response latency in voice-driven queries.
- Fine-tuned Llama 2 7B using QLoRA on character-driven dialog datasets, enabling live persona switching and adaptive conversational styles in AI agents.

IBM *Sep 2020 – Aug 2023*
Data Scientist

- Boosted document processing throughput by 30% by automating extraction workflows for 100K+ documents using Computer Vision and BERT-based semantic similarity microservices.
- Reduced API response times by 35% by engineering FastAPI-based RESTful microservices with async data flows.
- Improved deployment efficiency by 15% by containerizing microservices with Docker and automating releases on Azure AKS using Terraform and Jenkins CI/CD.
- Cut auditing costs by 50% ($200K/year) by delivering B2B SaaS automation solutions for document intelligence, collaborate cross-functionally in Agile sprints to refine and deliver features.

IBM *Jan 2020 – Jun 2020*
Data Scientist Intern

- Built a contract analysis system with LSTM and rule-based models for automated clause extraction and classification.
- Engineered Random Forest classifiers and layout-aware pipelines to automate clause segregation in contracts.
- Streamlined legal workflows by automating clause identification and extraction, reducing manual review.

Education

Duke University, Durham, NC *Aug 2023 – Dec 2024*
Master of Engineering in Artificial Intelligence

NIIT University, India *Aug 2016 – Jul 2020*
Bachelor of Engineering in Computer Science and Engineering, Minor in Artificial Intelligence

Projects

Multi-Modal Local File Search Engine

- Developed a multi-modal file search and recommendation system using Weaviate as a vector database.
- Integrated ImageBind for multi-modal embeddings to support text, image, audio, and video searches.
- Fine-tuned TinyLlama-1.1B for query expansion and converting search queries into JSON.

Video Moment Retrieval Engine

- Built a video search engine using CLIP for text-video embeddings and LLM for advanced query parsing.
- Leveraged temporal reasoning to enhance scene discovery (e.g., "Dramatic cliffhanger").
- Achieved a 30% improvement in retrieval precision over tag-based traditional methods.

Technical Skills

Languages: Python | **ML/AI:** PyTorch, TensorFlow, Keras, Scikit-learn, LangChain, LlamaIndex, CrewAI | **Databases:** MongoDB, MySQL, Redis, Pinecone | **LLMs/RAG/Inference:** RAG, Vector Search, vLLM, TensorRT, Triton Inference Server | **Cloud/DevOps:** Azure, AWS, Databricks, Docker, Kubernetes, Terraform, Jenkins | **Frameworks:** FastAPI, Flask, LitServe | **Storage:** MinIO, Azure Blob Storage | **Search:** Elasticsearch | **Web Scraping:** Beautiful Soup