Rushir Bhavsar

Tempe, AZ | rushirbhavsar@gmail.com |+1 (480)-875-6417 | linkedin.com/in/rushir-bhavsar/ | github.com/rushirb2001

EDUCATION

Arizona State University *Tempe, AZ*
Master of Science in **Data Science** | **Concentration:** High-Performance Computing ***Graduation Date:** May 2025*
Relevant Courses: Statistical ML, Data Processing at Scale, Convex Optimisation, Information Assurance and Security

Nirma University *Ahmedabad, GJ, India*
Bachelor of Technology (B.Tech) in **Computer Science and Engineering** ***Graduation Date:** June 2023*
Relevant Courses: Data Structures & Algorithm, NLP, Deep Learning, Computer Vision

WORK EXPERIENCE

ML Researcher | Arizona State University, Tempe, AZ ***November 2025 – Present***

- **Architecting** a modular CCP plasma **simulation framework** in PyTorch Lightning and JAX with swappable **architectures, samplers, collocation strategies**, and interpolators for Applied Materials **semiconductor R&D**.
- **Executed 60+ experiment** configurations across architectures and sampling strategies with adaptive loss balancing, **identifying optimal training** configuration through **reproducible** experiment tracking.

ML Engineer Intern (Sci-Dev) | OpenEye, Cadence Design Systems, Santa Fe, NM ***July 2025 – October 2025***

B2B SaaS computational molecular design leader ($18M Revenue, 150+ employees), accelerating drug discovery for top 20 pharma since 1997.

- **Architected** end-to-end protein **thermostability prediction pipeline** from scratch, achieving **7.2ms/seq** throughput across **1M+ sequences** using ESM2 (650M-param pLM), Hugging Face, cuDF, and PyTorch Lightning.
- **Developed contrastive learning** architecture for protein property prediction, **replacing GPR** with cuML and RAFT, **scaling batches 333x** (150 → 50K+ items) on 5120-dim pLM embeddings.
- **Built** a unified **configuration framework** with OmegaConf and Pydantic across training, testing, and evaluation, **parallelising 20+ experiments** for **antibody developability** prediction.

GenAI Engineering Intern |Talin Labs, Los Angeles, CA ***June 2024 – September 2024***

B2B SaaS consulting startup delivering AI and blockchain solutions to 10+ Fortune500 clients across 5+ countries since 2019.

- **Deployed** fine-tuned **Mistral-7B-Q8** on K8S to serve multi-document synthesis across **12+ enterprise on-prem** environments, hitting **<200ms p95 latency** at **10K users** through FastAPI.
- **Developed RAG evaluation** framework to validate factual grounding across **10K human-evaluated queries**, achieving **88% accuracy** by measuring chunk precision, citation accuracy, and cross-document consistency.
- **Engineered** a document **retrieval pipeline** with FAISS and LangChain to automate analysis and summary generation, **cutting manual review from weeks to minutes** across uploaded documents.
- **Architected 6-agent** LangChain **orchestration** with intent-based routing across PDF, XLSX, and DOCX parsing, geospatial rendering, and **conversational AI**.

TECHNICAL SKILLS

- **Gen AI & LLMs:** LangChain, RAG Systems, Prompt Engineering, AI Agents, Tool Calling, LLM Orchestration, Model Context Protocol (MCP), Fine-tuning, Model Quantisation.
- **ML Frameworks:** PyTorch, JAX/Flax, Keras, Scikit-Learn, CUDA, RAPIDS, Hugging Face Transformers
- **Vector & Data Stores:** Qdrant, FAISS, Neo4j, PostgreSQL, MongoDB, SQLite, Snowflake
- **Cloud & MLOps:** AWS (SageMaker, Bedrock, Lambda, EC2, S3), Docker, Kubernetes, Helm, MLflow, Jenkins
- **Programming Languages & Tools:** Python, C/C++, SQL, GraphQL, FastAPI, REST APIs, Git, Linux/Unix

SELECTED TECHNICAL PROJECTS

MedQuery: Surgical Knowledge Assistant | LangGraph, LangChain, FAISS, OpenAI API| Project Link

- **Building** a medical query **classification benchmark** with structured output extraction across intent, entities, and relationships, evaluating **10K annotated synthetic queries** across **4 difficulty levels** and **9 categories**.
- **Benchmarking** local quantised models (MLX, llama.cpp) against **API baselines** (OpenAI, Anthropic) to evaluate **cost-latency-quality tradeoffs** for medical query classification via a **backend-agnostic** evaluation framework.

HybridFlow: Multi-Backend Knowledge Layer | Qdrant, Neo4j, SQLite, FastAPI | Project Link

- **Building multi-backend retrieval** across Qdrant (**36K paragraphs**), Neo4j (**107K nodes**), and SQLite to index **3 surgical textbooks** with PubMedBERT embeddings.
- **Designing agentic retrieval** layer exposing **10 tool functions** with dynamic context expansion, enabling **multi-step reasoning** over retrieved content, via similarity matching and graph traversal at **<12ms p50**.

Yelp Recommendation & Sentiment Platform | PySpark, FastAPI, MLflow, Docker, Redis | Project Link

- **Built RESTful microservices** for collaborative filtering and sentiment classification, serving **5M+ interactions** at **<100ms latency** via Spark ALS, Redis, and MLflow.