

UNNATI BHAVEN SHAH

New Brunswick, NJ | (213) 643-9377 | unnatishah.ix@gmail.com | [linkedin.com/in/unnati9](https://www.linkedin.com/in/unnati9) | github.com/unnatibshah

EDUCATION

University of Southern California Master of Science in Computer Science - Graduate Research Assistant: ML analysis of large-scale App Store and developer data for platform acquisitions research - Graduate Teaching Assistant: BUAD 310g Applied Business Statistics	Los Angeles, CA August 2022 – May 2024
University of Mumbai Bachelor of Engineering in Computer Engineering	Mumbai, India August 2018 – June 2022

WORK EXPERIENCE

Johnson & Johnson Generative AI Engineer	New Brunswick, NJ June 2025-Present
---	--

- Architected a stateful multi-agent system using A2A, LangGraph and Model Context Protocol, implementing cyclic graph workflows to handle complex intent routing for 50K+ daily queries, achieving 96% intent classification accuracy.
- Engineered a low-latency graph retrieval system using Neo4j and Qdrant; implemented multi-hop retrieval algorithms to improve search relevance by 20% and optimized vector lookup latency by 33%.
- Built a production-grade evaluation framework using OpenTelemetry to monitor system latency and drift; automated regression testing pipelines which reduced QA/evaluation cycles by 81%.
- Managed full-stack LLM operations across 12 production endpoints; implemented token-caching strategies and model quantization techniques to reduce monthly API infrastructure spend by 18%.

Machine Learning Engineer	June 2024-January 2025
----------------------------------	------------------------

- Scaled a RAG system using Azure OpenAI, DynamoDB, Bedrock, and Qdrant to 100K+ documents and 50K+ daily queries, achieving 96% relevance through multi-query retrieval and context compression optimization.
- Designed a hybrid LLM/rule-based parsing engine for unstructured data (Excel, PPT); unlocked 100K+ previously inaccessible documents for cross-source querying, boosting retrieval precision by 28% for critical business searches.
- Fine-tuned Mistral-7B model using QLoRA and 4-bit quantization on 150K+ HR policy documents, reducing training memory requirements by 75% while achieving production-grade accuracy for downstream policy retrieval and classification tasks.
- Established LLM-as-a-judge benchmarking framework with Claude Sonnet 3.5 to programmatically measure faithfulness, precision, and answer relevancy, standardizing deployment criteria across 8+ business units.

C3.ai Data Scientist (Generative AI)	Redwood City, CA January 2025-May 2025
---	---

- Modeled a Text to SQL generator with Llama 3-70B, implementing grammar-constrained decoding with GBNF parsers to achieve 94% SQL syntax accuracy across 5M+ financial transactions
- Engineered a reproducible evaluation suite for generated SQL using cosine-similarity scoring and benchmark query sets, improving code generation accuracy by 15%.
- Migrated core retrieval stack from ColBERT to pgvector with HNSW indexing, scaling the system to 2M+ records and improving query throughput by 40% while maintaining 94% semantic precision.
- Standardized an LLM Token Counter framework adopted by 10 internal teams, enabling real-time usage transparency and driving a 25% reduction in platform-wide LLM inference costs.

University of Southern California Machine Learning Researcher	Los Angeles, CA September 2022-June 2024
--	---

- Engineered a distributed text mining pipeline to process 71M+ user reviews, extracting 100+ unique semantic keywords per category to quantify "economies of scope" for a study published in the Strategic Management Journal.
- Built a parallelized fuzzy matching system to link 279,000+ developer profiles with acquisition data across disparate sources, achieving 99% accuracy using custom string distance algorithms (Jaro-Winkler, Levenshtein).
- Constructed high-dimensional SDK-to-Category sparse matrices for 12,000+ unique software kits, enabling the detection of technological synergies across the iOS ecosystem.

TECHNICAL SKILLS

Programming Languages: Python, SQL, Bash, JavaScript, TypeScript, Java
Generative AI: Multi-Agent Orchestration, Agentic Evaluation, Prompt Engineering, LLM Guardrails, Model Context Protocol, Fine-tuning (LoRA, QLoRA), Token Optimization, Vector Databases (Chroma, Qdrant, pgvector), Telemetry, RAG, GraphRAG (Neo4j)
AI & ML Frameworks: LangChain, LangGraph, A2A, FAISS, PyTorch, TensorFlow, Pandas, Numpy, Scikit-learn, NLTK, SpaCy
Web & Application Frameworks: FastAPI, Flask, Node.js, Angular, Streamlit, Flutter, Tableau
Cloud & DevOps: Azure AI, Azure Functions, AWS (S3, Lambda, Sagemaker, Bedrock), GCP, Docker, Kubernetes, Git, CI/CD

RESEARCH PUBLICATIONS

- The Entry-Deterring Effects of Synergies in Complementor Acquisitions: Evidence from Apple’s iOS App Store. Strategic Management Journal, 2024.
- Maintaining Privacy in Medical Imaging with Federated Learning, Deep Learning, Differential Privacy, and Encrypted Computation. IEEE I2CT, 2021.