

KETAN SARDA

Irvine, CA | (949)-992-4730 | sardak@uci.edu | [LinkedIn](#) | [GitHub](#) | [Medium](#) | [Portfolio](#)

SUMMARY

2+ years of experience across Lenovo, UCI BAI, and Persistent Systems delivering AI, cloud, and full-stack solutions
A quick learner passionate about creating scalable, intelligent, and production-ready solutions across the AI stack.

EDUCATION

University of California, Irvine

Master of Computer Science

GPA: 3.95

Dec 2025

MES College of Engineering, Pune, India

Bachelor of Engineering, Computer Engineering (Honors in Data Science & Statistics)

GPA: 3.84

May 2022

WORK EXPERIENCE

Lenovo

Morrisville, NC

AI Architecture Intern

June 2025 – Aug 2025

- Designed and implemented AI-powered automation pipelines for hybrid HPC infrastructure using Python, Lenovo Confluent, and xClarity APIs, reducing manual intervention by 80%
- Integrated an MCP server to provide dynamic, tool-aware context to LLMs, incorporating reinforcement and feedback loops to continuously improve accuracy and reduce hallucinations by 20%
- Led vCenter infra automation with Terraform and schema validation, cutting configuration errors by 95%
- Automated internal billing and approval processes via Python RPA + Jira, reducing cycle time by 60%
- Documented internal AI automation frameworks and mentored an intern on full-stack AI development

UCI Beall Applied Innovation

Irvine, CA

AI Engineering Intern

Nov 2024 – June 2025

- Built an Amazon-sponsored researcher discovery platform using RAG pipelines (LangChain + LLMs)
- Engineered a secure React front-end with Amazon Cognito authentication and AWS Lambda backend integration
- Boosted LLM accuracy to 95% while cutting costs 80% via optimized context management
- Ingested 400K+ publications, patents, and grants into Amazon S3 with metadata indexed in PostgreSQL
- Built Spark + AWS Glue ETL pipelines to process terabytes of research data, reducing retrieval time by 40%
- Achieved 50% faster retrieval by integrating OpenSearch with NoSQL indexing for automated query resolution

Persistent Systems

Pune, India

Senior Software Engineer

June 2022 – Aug 2024

- Built an OpenAI chatbot to automate test-case generation for Python, TypeScript cutting QA time by 40%
- Applied prompt tuning and feedback-driven reinforcement loops to refine responses, boosting 30% accuracy
- Reduced API response time by 66.67% (from 5.7s to 1.9s) by query optimizing in Java and MySQL
- Improved Angular performance, reducing API calls by 25% through efficient model, service, and pipe updates, resulting in a more responsive UI and potential cost savings of around 12%
- Automated 150+ UI test cases with Protractor and Jasmine, achieving over 95% coverage
- Architected and implemented a suite of 5 reusable web components with React, JavaScript, HTML, and Node JS, accelerating development by 40% across 7 applications within a monorepository

PROJECTS

LLM AccelGo (Go, CUDA, TensorRT, GPUs)

Aug 2025 - Nov 2025

- Built a GPU-accelerated LLM fine-tuning and inference framework achieving 2.2× speed-up, 65% lower VRAM usage, and <60 ms latency through mixed-precision, LoRA adapters, and FlashAttention.
- Developed distributed training APIs in Go on DigitalOcean GPUs with TensorRT integration and added quantitative evaluation using RAGAS and OpenEvals

[GitHub](#)

Luminix (Next.js, Gemini, PostgreSQL, and AWS)

Jan 2025 - May 2025

- Developed Luminix, an AI-powered platform that streamlines academic research with smart literature search, summarization, citation management, and PDF analysis by enhancing research efficiency by 60%. Accelerating synthesis and citation workflows across 200M+ publications indexed via Semantic Scholar

[Linktree](#)

Patient Risk Stratification (Angular, Python, Flask, & SQL)

Jan 2023 - Apr 2023

- Engineered a web platform analyzing over 46K+ patient records for chronic conditions and emergency visits, calculating risk scores and predicting health outcomes (using SVM, KNN, Random Forest models)

[Github](#)

SKILLS

Languages: Python, Java, C++, SQL, JavaScript, TypeScript, HTML & CSS

Tools: LangChain, PyTorch, TensorFlow, React, Node.js, Spring Boot, AWS, GCP, Spark, Docker, Kubernetes, Jenkins

Technologies: Generative AI, RAG & Agentic Workflows, Cloud & Distributed Systems, MLOps, CI/CD