

# KETAN SARDA

Irvine, CA | (949)-992-4730 | [sardak@uci.edu](mailto: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

<b>University of California, Irvine</b> <i>Master of Computer Science</i>	<b>GPA: 3.95</b> <i>Dec 2025</i>
<b>MES College of Engineering, Pune, India</b> <i>Bachelor of Engineering, Computer Engineering (Honors in Data Science &amp; Statistics)</i>	<b>GPA: 3.84</b> <i>May 2022</i>

## WORK EXPERIENCE

<b>Lenovo</b> <i>AI Architecture Intern</i>	<b>Morrisville, NC</b> <i>June 2025 – Aug 2025</i>
<ul style="list-style-type: none"><li>Designed and implemented AI-powered automation pipelines for hybrid HPC infrastructure using Python, Lenovo Confluent, and xClarity APIs, reducing manual intervention by 80%</li><li>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%</li><li>Led vCenter infra automation with Terraform and schema validation, cutting configuration errors by 95%</li><li>Automated internal billing and approval processes via Python RPA + Jira, reducing cycle time by 60%</li><li>Documented internal AI automation frameworks and mentored an intern on full-stack AI development</li></ul>	
<b>UCI Beall Applied Innovation</b> <i>AI Engineering Intern</i>	<b>Irvine, CA</b> <i>Nov 2024 – June 2025</i>
<ul style="list-style-type: none"><li>Built an Amazon-sponsored researcher discovery platform using RAG pipelines (LangChain + LLMs)</li><li>Engineered a secure React front-end with Amazon Cognito authentication and AWS Lambda backend integration</li><li>Boosted LLM accuracy to 95% while cutting costs 80% via optimized context management</li><li>Ingested 400K+ publications, patents, and grants into Amazon S3 with metadata indexed in PostgreSQL</li><li>Built Spark + AWS Glue ETL pipelines to process terabytes of research data, reducing retrieval time by 40%</li><li>Achieved 50% faster retrieval by integrating OpenSearch with NoSQL indexing for automated query resolution</li></ul>	
<b>Persistent Systems</b> <i>Senior Software Engineer</i>	<b>Pune, India</b> <i>June 2022 – Aug 2024</i>
<ul style="list-style-type: none"><li>Built an OpenAI chatbot to automate test-case generation for Python, TypeScript cutting QA time by 40%</li><li>Applied prompt tuning and feedback-driven reinforcement loops to refine responses, boosting 30% accuracy</li><li>Reduced API response time by 66.67% (from 5.7s to 1.9s) by query optimizing in Java and MySQL</li><li>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%</li><li>Automated 150+ UI test cases with Protractor and Jasmine, achieving over 95% coverage</li><li>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</li></ul>	

## PROJECTS

<b>LLM AccelGo</b> (Go, CUDA, TensorRT, GPUs)	<i>Aug 2025 - Nov 2025</i>
<ul style="list-style-type: none"><li>Built a GPU-accelerated LLM fine-tuning and inference framework achieving <math>2.2\times</math> speed-up, 65% lower VRAM usage, and &lt;60 ms latency through mixed-precision, LoRA adapters, and FlashAttention.</li><li>Developed distributed training APIs in Go on DigitalOcean GPUs with TensorRT integration and added quantitative evaluation using RAGAS and OpenEvals</li></ul>	<a href="#">GitHub</a>
<b>Luminix</b> (Next.js, Gemini, PostgreSQL, and AWS)	<i>Jan 2025 - May 2025</i>
<ul style="list-style-type: none"><li>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</li></ul>	<a href="#">Linktree</a>
<b>Patient Risk Stratification</b> (Angular, Python, Flask, & SQL)	<i>Jan 2023 - Apr 2023</i>
<ul style="list-style-type: none"><li>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)</li></ul>	<a href="#">Github</a>

## 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