

SOHAM SONAR

soham.sonar427@gmail.com ◊ [LinkedIn](#) ◊ United States ◊ +1 (312) 975-7439 ◊ [GitHub](#) ◊ [Portfolio](#)

EDUCATION

Master of Computer Science Illinois Institute of Technology, Chicago, IL Relevant Coursework: Machine learning, Cloud computing, Design and analysis of algorithms, Advanced operating systems	August 2023 - May 2025
Bachelor of Computer Engineering Savitribai Phule Pune University Relevant Coursework: Data structures and algorithm, Object oriented programming, Advanced database organization, Big data	August 2018 - July 2022

EXPERIENCE

Research Assistant Gnosis Research Center - Illinois Institute of Technology	March 2025 - September 2025 <i>Chicago, IL</i>
<ul style="list-style-type: none">Developed an integrated Agentic AI platform leveraging multi agent orchestration to automate end-to-end workflows across 40+ node clusters, enabling autonomous task execution and intelligent operations coordination.Enhanced the performance of open source projects (IOWarp, Chronolog), by designing REST APIs and integrating an intuitive assistant for data analytics and AI driven workflows, reducing average data retrieval latency by 40%.Automated CI/CD pipelines with GitHub Actions and Docker, automating build, lint testing, and deployment processes across on prem systems and scalable cloud environments for faster and more reliable application delivery.Built clean, reusable code while researching LLM based applications (Cursor, Claude), applying best practices in architecture, code reviews, unit testing, and scalability to ensure reliable enterprise scale AI systems.	
Machine Learning Intern Vosyn Inc.	January 2025 - April 2025 <i>Chicago, IL</i>
<ul style="list-style-type: none">Designed and optimized machine learning models using Vertex AI, Kubeflow and Tensorflow to improve real-time multilingual voice synthesis accuracy by 35%, ensuring seamless contextual translation across global markets.Integrated 10+ AI voice features into customer facing applications through continuous model development and A/B testing, enabling real time support and improving usability for non-technical users.Deployed ML models for real-time voice localization using Kubernetes & Cloud Run, optimizing inference via CPU/GPU benchmarking reducing latency by 20% and enabled scalable cross-platform integration.	
Executive Hexaware Technologies	March 2023 - June 2023 <i>Mumbai, India</i>
<ul style="list-style-type: none">Architected and debugged software applications for healthcare solutions using Python, and SQL, achieving 60% performance improvement through software architecture optimization.Streamlined data entry workflows, reducing manual workload and improving data processing efficiency by 30% through scripting and workflow automation.Collaborated with cross functional teams in an Agile Scrum environment, and led backlog grooming and sprint planning across software engineering and QA teams, reducing post-deployment defects by 30%.	

SKILLS

Programming Languages: Python, Java, C++, SQL, HTML, CSS, Bash/Shell Scripting.
AI Development: Autogens, FAISS, Finetuning, Github Copilot, Langchain, Vector DB, LLMs, RAG Models, Claude.
Machine Learning: Tensorflow, Pytorch, Scikit-learn, Huggingface, Keras, supervised & unsupervised models, MLflow.
Web/Software Development: Angular, React, Next.js, REST API, FastAPI, JIRA, Linux, Git, Agile, Scrum, VS Code.
Cloud/Big Data & Database: AWS, GCP, Azure, Docker, Kubernetes, Hadoop, Spark, Kafka, MySQL, PostgreSQL.

PROJECTS

Docquery [\[Link\]](#)

- Engineered an **AI powered knowledge** and document querying platform using **FastAPI, PostgreSQL, FAISS, and OpenAI embeddings**, enabling hybrid semantic–keyword retrieval with 30% higher answer accuracy than regular systems.
- Implemented scalable, production grade architecture with **Redis caching, async processing, supporting 200+ concurrent users** while reducing compute cost per query by 40% through intelligent caching and incremental indexing.

Enterprise IO Automation Framework [\[Link\]](#)

- Led the development of the **Scientific Model Context Protocol (MCP)** server framework, including Pandas, Parquet, Plot and HDF5 MCP servers, to **automate I/O and filesystem workflows** for local and cloud environments.
- Designed a **custom LLM client** using **Google Gen AI sdk** to coordinate **120+ simulation pipelines**, processing **multi-terabyte datasets** and significantly **reducing data access latency** across distributed systems.