Ayush Ranjan

San Francisco, CA, USA | aranjan1@ucsc.edu | (831) 266-5973
linkedin.com/in/ayuranjan | github.com/ayuranjan | ayuranjan.github.io

Summary

Backend Engineer with **3.5+** years of experience, including **2.5+ years in production backend engineering** and **1 year in applied AI and agent development**. Experienced in building and operating **distributed microservices at Airbnb** and **monolithic enterprise backend systems at Mercedes-Benz**. In academic labs and independent projects, developed **LLM-based agents, vector search systems, multimodal AI pipelines, and MCP-compatible developer tools** for code and repository understanding. Strong background in **backend architecture, distributed systems, databases, and AI-augmented infrastructure**.

Education

University of California, Santa Cruz **Sep 2023 – Aug 2025**
Master of Science (MS) in Computer Science CGPA: 3.92/4

- **Relevant Coursework:** Analysis of Algorithms, Design and Implementation of Database Systems, Deep Learning for Advanced Computer Vision, Artificial Intelligence (AI), Applied ML: Deep Learning (DL), Computer Networks
- **Teaching Assistant Roles: 4x TA for Database Systems (CSE-180/182)**: led labs/projects on **SQL**, **transaction management**, **indexing**, **stored functions**, **PL/pgSQL**, **ETL workflows**, and **query optimization**. **1x TA for Software Engineering (CSE-115A)**: mentored student teams on **Agile development**, version control, and team collaboration practices.

Manipal University, Jaipur **July 2017 – May 2021**
Bachelor of Technology (B.Tech.) in Information Technology

- **Relevant Coursework:** Operating Systems, Data Mining and Warehousing, Data Science, Cryptography and Network Security, Advanced Data Structures, Natural Language Processing, Advance Machine Learning Techniques

Work Experience

Altimetrik (Client: Airbnb) **San Francisco, CA**
Backend Developer | **Dropwizard, Kafka, Thrift RPC, MySQL, Redis, Bazel, K8s** *Sep 2025 – Present*

- Engineered high-concurrency **Java microservices** for Airbnb Payments Incentives, processing **2.55B requests/month (avg 5.06k QPS)** using **Kafka, Thrift RPC, Dropwizard, and Kubernetes**.
- Optimized developer workflows by contributing to internal **Claude Code** capabilities; engineered **custom commands and skills** to automate Payments-specific codebase navigation, analysis, and boilerplate generation.
- Co-architected **Thrift RPC endpoints** that surfaced **$120M** in dormant user credits and integrated results into automated re-engagement email pipelines.
- Developed a robust event-driven **Producer-Consumer architecture** using **Kafka** and **Tempo** to automate the lifecycle of Airbnb Virtual Credit Cards (VCC), ensuring 100% automated refund reconciliation upon card expiration.

Information Retrieval and Knowledge Management Lab, UCSC **Santa Cruz, CA**
AI Engineer Intern | **Python, Flask, LangGraph, Whisper, Pinecone, Hugging Face** *July 2024 – Sep 2024*

- Partnered with a **stealth hardware startup** to develop a **0-to-1 multimodal AI agent** for smart wearable devices (camera-integrated earphones), implementing wake word detection, intent classification, and real-time audio-visual processing for calorie estimation, emergency response and video summarization.
- Designed intelligent query routing system with **95% accuracy** in classifying continuous vs. new queries, integrating **Dialogflow for 8+ pre-built workflows** (calorie estimation, contact calling, emergency location services) and custom **LangGraph agents** for open-domain conversations.
- Engineered real-time multimodal data fusion system combining **audio transcription** (Whisper), **computer vision** (food segmentation, depth estimation), and vector similarity search for intelligent fallback routing to **external tools (web search, OCR)** when confidence scores dropped below 0.8 threshold.
- Developed a multi-threaded memory manager to asynchronously encode and cache historical observations (images, transcripts) into vector embeddings using **Hugging Face transformers**, with storage in **Pinecone**.

Capgemini (Client: Mercedes-Benz) **Mumbai, India**
Backend Developer | **Java, JDBC, SQL, IBM Db2, Data Modelling** *July 2021 – Aug 2023*

- Maintained **XDIS**, a **SOAP**-based diagnostic tool structured around a three-tier monolithic Java architecture used by **Mercedes-Benz** service teams for vehicle automation and diagnostics.
- Reduced XML migration time by **67% across 50+ service teams** by migrating to a **delta-based Db2 migration** strategy.
- Optimized export testing by creating a wrapper around the Autosar framework and implementing an **XML file** import strategy, **reducing overall testing time by 40%** and speeding up export time for **individual modules by 17% on average**.

- **Headed the Data Modeling Team**, focusing on backend schema evolution for **vehicle network topology change requests** (e.g., ECU reconfigurations, bus architecture edits).
- Achieved **3rd Place at Innocircle 2022**, Mercedes-Benz Internal Innovation Forum by implementing a micro frontend architecture that enabled users to modify their vehicle network topology and review changes, **reducing process time by over 50%**.
- Developed an **AI-assisted validation system** for over 2,500 historical Change Requests by embedding symbolic vehicle network topologies using custom **Word2Vec** and **Sentence-BERT models**, which flagged rare configurations and recommended optimal topologies, improving validation accuracy.

Capgemini **Mumbai, India**

Full-Stack Developer Intern | **Spring Boot, React, Redux, MySQL, JUnit** *Jan 2021 – May 2021*

- Developed a full-stack Medical Portal using **Spring Boot** and **React/Redux**, engineering 17 RESTful endpoints documented with **Swagger** to streamline feature delivery and team collaboration.
- Architected a secure data layer using **MySQL** and **JPA/Hibernate**, implementing **Spring Security** (JWT) with role-based access control and achieving high code reliability through automated **JUnit** and **Jest** testing
- Streamlined deployment by containerizing the application with **Docker**, establishing a **GitHub Actions CI/CD** pipeline, and orchestrating local releases via **Kubernetes** (Minikube) to ensure consistent environments for stakeholders and QA.

Selected Projects

Fathom : Code-Aware Search Engine | ***Python, FastAPI, pgvector, Java, Protobuf*** | GitHub **Dec 2025 – Jan 2026**

- Built a **file-discovery system** for coding agents (Claude Code, Gemini CLI) that combines lexical (Zoekt), semantic (pgvector), and structural (SCIP) search to **precisely locate relevant files on first query-reducing redundant file reads and associated token costs**.
- Engineered a **graph-augmented retrieval pipeline** using SCIP cross-references to automatically surface dependent files, callers, and superclasses alongside search results-**eliminating exploratory navigation steps that consume agent context windows**.
- Deployed as an **MCP-compatible tool server** with FastAPI, enabling agents to execute go-to-definition, find-references, and semantic file search in **single tool calls instead of multiple sequential file reads**, improving task completion speed and cost efficiency.

PgVector+ | ***C/C++, PL/pgSQL, Database Systems, Vector Similarity*** | Ppt **Jan 2024 – Mar 2024**

- Designed and built a **custom PostgreSQL extension** on top of pgvector to support hybrid similarity–dissimilarity search and low-level query composition, bridging gaps in vector DB functionality seen in systems like Pinecone and Qdrant.
- Prototyped a `compound_similarity()` operator in **PL/pgSQL** to support similarity search queries like "similar to X, but unlike Y" using cosine and inner-product thresholds.
- Prototyped PL/pgSQL-based `search_similar_vectors()` function to simulate centroid-based multi-query composition and validate set-based similarity retrieval.
- Earned an **A grade** in CSE 215: Design and Implementation of Database Systems for system-level innovation in vector search.

Unveiling Glitches in CLIP | ***Hugging Face, Python, Vector Database, Prompt Engineering*** | arXiv **Jan 2024 – Mar 2024**

- Conducted in-depth analysis of the CLIP model's image comprehension capabilities. Identified and documented **14 systemic faults**, including **four novel faults**, impacting CLIP's interpretation of images using **two novel methodologies**.
- Implemented the Discrepancy Analysis Framework (**DAF**) to analyze discrepancies in image similarity rankings between CLIP and **DINOv2** and utilized **OpenAI's GPT API** to identify and analyze faults systematically. Utilized the Transformative Caption Analysis for CLIP (**TCAC**) approach to evaluate CLIP's response to transformations applied to images.
- Achieved **A+ grade** in CSE 290D Neural Computation at UCSC for this project.

Technical Skills

Programming Languages: Python, Java, TypeScript, JavaScript, C/C++, Go, Rust, SQL
AI/ML Tooling: NLTK, spaCy, transformers library, Sentence-BERT, Word2Vec, MCP Protocols, MLflow, Prompt Engineering
Backend Development: Spring Boot, REST APIs, GraphQL, FastAPI, Flask, PL/pgSQL, JDBC, Node.js, Express.js
Frontend Development: React, Next.js, Redux, Axios, HTML/CSS, Swagger / OpenAPI, JWT, RBAC
Databases: PostgreSQL, MySQL, Oracle, IBM Db2, MongoDB, Redis, Vector Databases (pgvector, Pinecone, Milvus, Qdrant)
Data Engineering: ETL (Java/XML), Liquibase, Pandas, NumPy, Apache Kafka, Apache Spark
DevOps & MLOps: Docker, Docker Compose, Kubernetes, Minikube, Git, GitHub Actions, Jenkins, Maven, Gradle, Terraform
Testing and Automation: JUnit, Mockito, Jest, React Testing Library, Playwright, TDD, BDD
Cloud and Distributed Systems: AWS (EC2, S3, RDS, Lambda, DynamoDB), Amazon Bedrock, Google Cloud(GCP), Azure
Practices & Miscellaneous: Agile/SAFe, CI/CD, SDLC Lifecycle, Prometheus, Grafana, Apache Airflow