Ankita Sethi

Santa Clara, CA | +1(530) 407-7394 | ankitasethi282@gmail.com | LinkedIn | GitHub | Portfolio

Summary

Software engineer focused on backend systems and applied AI, building scalable APIs and production data pipelines. Experienced in real-time inference, vector search, and large-scale data processing across cloud environments. Strong in building reliable, low-latency systems with LLM workflows and data pipelines, with exposure to Bazel-based build systems.

Professional Experience

Software Engineer (Founding Engineer) | *Aurora IQ Inc.*| *Remote (U.S)* Feb 2026 – Present

- Developed backend services using **Python** and **REST APIs** for image ingestion workflows handling 2K+ assets.
- Designed scalable data pipelines using **PostgreSQL** and **AWS S3**, improving retrieval efficiency by ~30%.
- Optimized **API** latency from ~400 ms to ~200 ms through query tuning and indexing.
- Built embedding and similarity search systems over **10K+ images**, enabling scalable recommendations.
- Implemented event-driven pipelines supporting real-time ML inference and improving throughput under concurrent workloads.

Software Engineer (Research Assistant) | *Stony Brook University* | *Remote (U.S)* July 2025 – Present

- Engineered end-to-end content segmentation using **Python**, **Django** and **Redis** caching, reducing response latency by ~45%.
- Constructed **asynchronous** and **batch processing pipelines** handling **100s** of pages per run, improving system stability.
- Integrated **Gemini 2.5** and **LangChain** with structured evaluation workflows, improving segmentation accuracy by 18%.

Graduate Research Assistant | *Stony Brook University* | *Stony Brook, NY* Aug 2024 – May 2025

- Led ML driven accessibility research by integrating **LLMs** with NVDA to interpret documents, charts and web content, improving usability and accessibility through iterative model evaluation.
- Built end-to-end systems including Chrome extensions, backend pipelines, improving reliability & user accessibility workflows.

Data Engineer | *Accenture Private Limited* | *Bangalore, India* Jun 2021 – Aug 2023

- Engineered large scale **ETL** pipelines using **Talend** and **Snowflake**, improving data throughput by 35% and reducing latency.
- Designed workflow **automation** and pipeline orchestration, reducing manual intervention and improving reliability.
- Automated structured data extraction by deploying **BERT** based **NLP** classifiers processing **2M+ unstructured records**.
- Built 15+ **Tableau** dashboards to monitor **KPIs**, pipeline health, anomalies & data quality, cutting issue identification time by 40%.
- Optimized **SQL** queries and **Snowflake** schemas with validations & alerts, reducing compute cost & removing manual QA steps.
- Developed a **Flask** based internal **chatbot** handling 50+ daily service tickets, improving request routing and support efficiency.
- Implemented CI/CD pipelines using Jenkins and Git workflows for automated testing, build validation and reliable deployments
- Collaborated with global teams, mentored junior engineers, improving delivery consistency through **Agile** and **Jira** workflows.

Technical Skills

Languages: Python, Java, C++, SQL
Backend & Systems: FastAPI, REST APIs, Django, Node.js, GraphQL, Microservices
Data & Infrastructure: Snowflake, PostgreSQL, ETL Pipelines, Data Modeling, Redis, AWS (S3, Lambda, ECS), Docker, Linux
DevOps & Workflow: CI/CD (Jenkins), Git, Kubernetes, Bazel-based build systems, workflow orchestration
Applied AI / ML: RAG, LLM Integration and evaluation, NLP, Transformers, Vector Embeddings, Computer Vision (CLIP, DINOv2)
ML & Data: PyTorch, Hugging Face, Scikit-learn, NumPy, Pandas, spaCy, OpenCV

Projects and Research

Member Aware Question Answering System Nov 2025

- Built a retrieval-augmented QA system using Gemini 2.5, Django and JSON retrieval for more than 3.5K member messages.
- Implemented context ranking, update detection and prompt construction with median latency under 2 seconds.
- Structured scalable pipeline for message extraction, feature preparation, deployed the service on Render with ongoing latency tracking.

Multi-Agent Approach for Detecting Hallucination in LLMs Sep 2024

- Formulated a multi-agent framework to detect hallucinations in LLM outputs using concept inference and NER, improving factual reliability by 18% in evaluation benchmarks.
- Scaled domain coverage from 7 to 57 via self-familiarity scoring, increasing detection consistency across varied text types by 22%.

Smart Energy Assistant for Smart Homes Feb 2024

- Engineered a home energy assistant using **ARIMA** model to forecast appliance usage based on time-series and weather trends, reducing prediction error by 28% across 1000+ observations.
- Integrated GPT-4 for anomaly detection and achieved $R^2 = 0.737$ through walk-forward validation.

Education

Stony Brook University | Stony Brook, NY | *Master of Science in Computer Science* Aug 2023 - May 2025

GITAM University | Visakhapatnam, India | *Bachelor of Technology in Computer Science and Engineering* Jun 2017 - Jun 2021