Rithik Sachdev

rithiksachdev3@gmail.com | (412) 224-8863 | linkedin.com/in/rithik-sachdev| Pittsburgh, PA | rithiksachdev.github.io

Summary: Full-Stack Software Engineer with interest in Infrastructure and Cloud Computing starting January 2025

EDUCATION

Carnegie Mellon University **Pittsburgh, Pennsylvania**
Master of Software Engineering in Scalable Systems **December 2024**
Courses: Design Patterns, Intro to ML, DevOps and CI, API Design, ML in Production **GPA 4.00**

Shri G.S. Institute of Technology and Science **Indore, India**
Bachelor of Technology, Computer Science and Engineering **July 2021**
Courses: Data Structures, Algorithms, Operating Systems, Cloud Computing, Advanced Databases **GPA: 3.62**

SKILLS

Programming Languages: Java, Python, TypeScript, JavaScript, SQL
Frameworks and Database: Spring Boot, React, Node.js, Flask, MySQL, Postgres, MongoDB, DynamoDB, Cassandra, Redis, Elasticsearch
Cloud Technologies and Infrastructure: AWS, Azure, GCP, Docker, Ansible, Istio, Kubernetes, Terraform, New Relic, Grafana, Prometheus
Tools and Techniques: Jmeter, Kafka, Postman, gRPC, GraphQL, Git, Jira

WORK EXPERIENCE

WAVLab, CMU, Language Technology Institute **Pittsburgh, PA**
Research Assistant **May 2024 - Present**

- Conducting research and preparing a paper on improving accuracy in automatic speech recognition using LLMs and prompt engineering, tested on datasets such as WSJ, CHiME, and Common Voice (CV), achieving significant performance enhancements.
- Utilizing an evolutionary algorithm on CUDA parallel processing to select the optimal prompts on a population of 10 different prompts, resulting in reduction in word error rates for automatic speech recognition systems.

Teel Lab, CMU, School of Computer Science **Pittsburgh, PA**
Research Assistant, Full-Stack Software Engineering Programmer **May 2024 - Present**

- Developing a messaging service on the Sail 2.0 platform to assist instructors in publishing notifications through banners and announcements, ensuring timely and effective communication.
- Creating end-to-end test cases using Cypress and building integration tests to improve reliability and functionality of system.

Ludo Lab, CMU, Human-Computer Interaction Institute **Pittsburgh, PA**
Research Assistant **January 2024 - May 2024**

- Improved accessibility for 1,000+ users by building an extension with language translation, keyboard shifts, and interactive buttons for disabled users, deploying a Pub/Sub system with Node.js, Redis, and S3.
- Reduced latency by 25% by performing web sockets and data compression strategies for data transfer.
- Implemented horizontal scaling using AWS EC2, boosting system capacity to support 5,000+ concurrent users.

Nextuple Inc **Bangalore, India**
Software Engineer **August 2021 - July 2023**

- Spearheaded efforts to provide training to interns working on both frontend and backend services, enhanced intern productivity and skill development, directly contributing to two interns securing full-time positions post-internship.
- Upgraded performance by 40% by building a custom load balancer for leveraging multiple reader pods in AWS.
- Reduced hosting expenses by 15% by reconfiguring data upload micro-services to support both Azure Cloud Services and Amazon Web Services, optimizing resource allocation.
- Decreased the API response time by 30 milliseconds using near cache in spring boot for transit and item microservice.
- Developed a robust Kafka consumer by leveraging Avro serialization responsible for processing 25 million records in 2 hours and storing into Cassandra and Redis for better latency in contrast to 6 hours originally due to invalid data ingestion.
- Designed and built a New Relic dashboard with Elasticsearch, setting up alerts for system health, API performance, cache rates, memory usage, and gateway latency, achieving greater effectiveness in anomaly detection and preventing service crashes.
- Increased system reliability by leading efforts in debugging and troubleshooting software issues and providing production on-call support.

RELEVANT PROJECTS AND RESEARCH PUBLICATIONS

Movie Recommendation System ***January 2024***

- Led a team to develop a production-grade movie recommendation service for over 1 million users with a 50ms response time, accomplishing best performance among a class of 100 students.
- Built ETL data pipelines to ingest Kafka data streams from over 1 million users and extract content consumption behavior into a time series database using Prometheus.
- Accomplished comprehensive monitoring and alerts using Grafana, enabling real-time visibility into system metrics and proactive performance optimization with a 5-second refresh rate.
- Created MLOps CI/CD Jenkins pipelines to assess recommendation model, data, and code quality, attaining at least 90% test coverage for all service components prior to deployment.

Comparison of Detection of Distributed Denial of Service attacks using Machine Learning | Publication [Link] ***August 2021***

- Co-authored a research paper presenting a comparative analysis of supervised learning algorithms for DDoS attack detection, using CIC-IDS 2017 dataset, and publishing complex findings.