AAKANKSH SINGH
+1 315-373-9940 | aakanksh.s10@gmail.com
LinkedIn: https://www.linkedin.com/in/aakankshsingh133/ | **GitHub:** https://github.com/Aakanksh94310
Portfolio: https://aakanksh94310.github.io/portfolio/

Software Engineer | LLM Inference, Performance Analysis, ML Systems | Python, C++
Software engineer with experience analyzing and optimizing large language model inference pipelines, focusing on latency, throughput, and reliability in distributed systems. Strong background in ML systems and performance experimentation for production workloads.

EDUCATION

Syracuse University – College of Engineering & Computer Science, Syracuse, NY
Master of Science in Computer Science (M.S.) | CGPA: 3.71/4.0 | Dec 2025
Relevant Coursework: Data Structures & Algorithms, NLP, Systems Programming, Operating Systems

SRM Institute of Science and Technology – SRMIST, Chennai, India
Bachelor of Technology in Electronics and Communication Engineering (B.Tech) | CGPA: 9.2/10 | June 2023
Relevant Coursework: Data Structures & Algorithms, Object-Oriented Design, System Design Fundamentals

TECHNICAL SKILLS

- Programming Languages: Python, Java, C++, SQL | Frameworks: PyTorch, TensorFlow, Keras, FastAPI, Flask, React, Node.js
- Machine Learning: Supervised Learning, Feature Engineering, Model Evaluation, Training Pipelines
- AI / LLM Systems: RAG, Vector Embeddings, LangChain, MCP, Prompt Engineering, LLM Evaluation
- Data & Distributed Systems: Apache Spark, Kafka, ETL, SQL Server | Testing & DevOps: Selenium, Jest, CI/CD, Git
- Cloud & Infrastructure: AWS (EC2, S3, IAM – foundational), Microservices
- Systems & Performance: Performance analysis, benchmarking, inference optimization, distributed systems

WORK EXPERIENCE

Software Developer Intern, Rightworks LLC, USA - Remote *September 2025 – December 2025*

- Analyzed and optimized distributed LLM inference pipelines using Apache Spark and vector embeddings, focusing on latency, throughput, and reliability.
- Designed scalable embedding ingestion pipelines with distributed processing and fallback mechanisms, increasing throughput by 18%.
- Built automated experimentation and benchmarking pipelines to evaluate inference latency, throughput, and embedding consistency, reducing model iteration time by 30%.
- Partnered with engineering and AI teams to translate inference evaluations into production-ready optimizations, improving system reliability and efficiency.
- Monitored and troubleshot pipeline failures, implementing fallback mechanisms and improving system reliability under production workloads.

Intern – Technology Track (Software Engineering), RIA Advisory LLC, USA – Remote *June 2025 – August 2025*

- Executed 200+ Selenium regression tests, improving coverage by 30% and reducing release defects by 25%.
- Performed cross-browser testing, logged issues in Jira, and verified fixes with developers to ensure stable releases.

Software Engineering and AI Intern, Ascendion , USA - Remote *May 2025 – July 2025*

- Developed LLM-powered tools including a RAG-based extractor and log analysis system, reducing debugging time by 40% and improving evaluation speed by 30%.
- Developed reusable NiceGUI frontends, improving usability and cutting interface development time by 50%.
- Contributed to improving inference efficiency in agent-based LLM systems, evaluating and improving inference efficiency through systematic performance analysis and optimization.

FULL-STACK DEVELOPER, InfluencivePress, Bengaluru *Jan 2023 – Dec 2023*

- Developed and optimized backend services and REST APIs using Node.js and MongoDB, improving system performance by 15% and increasing automated test coverage by 50+ cases.

PROJECT EXPERIENCE

Task Manager - Multi-Tenant SaaS Application *Jan 2023 – Feb 2023*

- Built a multi-tenant SaaS application using React, Node.js, and SQL Server with role-based access control.
- Implemented RESTful APIs and WebSocket-based real-time updates for scalable task management.

Facemask Detection System *Jan 2020 – March 2020*

- Built a machine learning–based computer vision classification system using Python, OpenCV, and TensorFlow for facemask detection.

LEADERSHIP

Student Organizer, AARUUSH, SRMIST, Chennai *Aug 2022 – Dec 2022*

- Increased participation in technical hackathons by 30% through CRM-driven outreach using HubSpot.