

# Sai Krishna Reddy Daka

[saikrishnareddydeka14.github.io](https://saikrishnareddydeka14.github.io) | [linkedin.com/in/sai-krishna-d-256753229](https://linkedin.com/in/sai-krishna-d-256753229) | [saikrishnareddy1422@gmail.com](mailto:saikrishnareddy1422@gmail.com) | +1 (602) 576-7786

Versatile software engineer with experience in full-stack development, AI/ML integration—including LLM-based solutions—and scalable backend systems. Skilled in building intelligent, end-to-end applications using modern web frameworks, cloud-native technologies, and data-driven design. Passionate about writing clean, maintainable code and delivering high-impact, production-ready software.

## Education

<b>Arizona State University</b>	<b>Tempe, Arizona, USA</b>
<b>Computer Science, M.S. (GPA: 4.0 GPA)</b>	<b>Expected Dec 2025</b>
<ul style="list-style-type: none"><li>Coursework in Data Mining, Data Processing, Cloud Computing, Machine Learning, Data Visualization, Software Engineering &amp; Testing</li><li>New American University Scholarship Awardee (Merit-Based Scholarship)</li></ul>	

## Experience

<b>Arizona State University</b>	<b>Tempe, Arizona, USA</b>
<b>Graduate Teaching Assistant</b>	<b>Aug 2024 – Dec 2025</b>
<ul style="list-style-type: none"><li>Graded assignments and evaluated student deliverables for 380+ students in CSE 565 (Software Verification, Validation, and Testing) and CSE 566 (Software Project, Process, and Quality Management) courses, ensuring precision and adherence to academic integrity standards.</li><li>Supported faculty by managing course logistics, resolving 80+ student queries, and delivering timely feedback on software engineering methodologies, testing strategies, and quality assurance practices.</li></ul>	
<b>Tata Consultancy Services (TCS)</b>	<b>Chennai, India</b>
<b>Software Development Engineer Intern</b>	<b>Feb 2023 – Apr 2023</b>
<ul style="list-style-type: none"><li>Developed a scalable internal framework to handle high volumes of asynchronous requests from microservices by batching, storing them in PostgreSQL, and forwarding to a target component—reducing processing latency by 40% and improving system reliability.</li><li>Optimized PostgreSQL database queries and REST API endpoints, significantly reducing average response time by 60% and enabling the system to handle 3x more concurrent requests without any performance degradation or service downtime.</li><li>Implemented comprehensive automated test scripts using Python’s PyTest framework, achieving 94% code coverage and reducing QA testing time by 15+ hours per week while proactively identifying critical bugs before production.</li></ul>	

## Projects

<b>Edge-Cloud AI-Powered Face Recognition System using AWS Greengrass, Lambda &amp; ECR</b>	<b>Mar 2025 – May 2025</b>
<ul style="list-style-type: none"><li>Designed and deployed a real-time face recognition system using MTCNN on AWS IoT Greengrass and FaceNet model on AWS Lambda, achieving 98.7% accuracy; containerized models with Docker and deployed via Amazon ECR.</li><li>Established a secure, event-driven messaging pipeline using MQTT and TLS, with intelligent edge-side filtering to discard “No-Face” frames, reducing cloud workload by 40% and cutting Lambda invocation costs by 35%.</li></ul>	
<b>Scalable AI-Powered Face Recognition System with Custom Event-Driven Autoscaling on AWS</b>	<b>Jan 2025 – Mar 2025</b>
<ul style="list-style-type: none"><li>Engineered a decoupled, event-driven face recognition system on AWS (S3, SQS, EC2) using MTCNN for face detection and a PyTorch-based FaceNet model for face recognition, achieving 98.7% accuracy and supporting 100+ concurrent requests.</li><li>Devised a custom autoscaling controller that adjusted EC2 provisioning based on SQS queue depth, scaling to 15 EC2 nodes, reducing idle time by 94%, and completing 100-request workloads in 96 seconds with a 2.5x throughput boost over static provisioning.</li></ul>	
<b>Multi-Modal Retrieval-Augmented Generation (RAG) System with LLM for Advanced Analysis</b>	<b>Aug 2024 – Nov 2024</b>
<ul style="list-style-type: none"><li>Built a multi-modal RAG system where users upload research paper PDFs via a web interface; files are securely stored in Amazon S3, with metadata indexed in Amazon DynamoDB to enable fast, structured retrieval of document content (text, tables, images).</li><li>Integrated the LLaMA-3.2-90B-Vision-Instruct model for image captioning and semantic answering, and applied FAISS for vector-based indexing, enabling real-time multi-modal query responses with 96% accuracy and reducing research effort by 87%.</li></ul>	

## Technical Skills

- Programming Languages:** Java, Python, JavaScript, C#, C, C++, SQL, PHP, HTML
- Cloud & DevOps:** AWS (Lambda, S3, EC2, Greengrass, IoT Core, SQS, ECS, ECR, DynamoDB), Docker, Git, CI/CD pipelines
- AI/ML & LLM Tools:** PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, LoRA Fine-Tuning, OpenAI API, LLaMA, Gemma
- Web Frameworks & API Development:** ReactJS, Django, Flask, CSS, Tailwind CSS, Bootstrap
- Database Technologies:** PostgreSQL, MySQL, Oracle, MongoDB, Microsoft SQL Server, SQLite, DynamoDB
- Software Engineering Practices & Tools:** Agile (Scrum, Kanban), SDLC, Jira, Unit Testing (JUnit, PyTest), Selenium, Test Automation

## Certifications

**Microsoft** – Career Essentials in Generative AI by Microsoft and LinkedIn, 2024 | **Oracle** – Oracle Cloud Infrastructure AI Certified Foundations Associate, 2024 | **Oracle** – Oracle Cloud Data Management Certified Foundations Associate, 2024 | **Forage** – Goldman Sachs Software Engineering Job Simulation, 2024