

Sai Paresh Karyekar

🏠 Sunnyvale, CA ☎ (404) 200-3907 ✉ saikaryekar@gmail.com 🌐 github.com/saikaryekar 🔗 linkedin.com/in/sai-karyekar

EDUCATION

Georgia Institute of Technology, Atlanta, GA <i>Master of Science in Electrical and Computer Engineering</i>	May 2025 <i>GPA: 3.9/4</i>
Veermata Jijabai Technological Institute (VJTI), India <i>Bachelor of Technology in Electronics & Telecommunication Engineering</i>	June 2023 <i>CGPA: 9.62/10</i>

EXPERIENCE

Amazon, Palo Alto <i>Software Development Engineer</i>	June 2025 - Present
<ul style="list-style-type: none">Scaled Visual Search platform to 10+ new marketplaces by provisioning infrastructure with AWS CDK, establishing secure cross-account access, and automating CI/CD deployments.Led production migration of a core request aggregation service from Dublin to Spain, ensuring zero downtime through rigorous testing and monitoring.Designed and documented system architecture for integrating Visual and Core Search, including service decoupling and scalable API design; mentored an intern on an offline batch localization pipeline.Enhanced platform security by deploying a MapToken-based hybrid authentication system, A/B tested to mitigate bot-driven DDoS traffic and block unauthorized access without affecting customer experience.	
NVIDIA, Santa Clara <i>Software Engineering Intern</i>	June 2024 - Aug 2024
<ul style="list-style-type: none">Built an internal Generative AI-powered content generation tool using RAG, reducing blog creation time by 2.5x.Developed and deployed a vector search pipeline using Milvus for efficient document retrievalSpawned AWS EC2 instances with firewall rules and VPN access, ensuring secure access with load balancing.Containerized the application using Docker and orchestrated deployment via AWS ECS for scalability.	

PROJECTS

Enhancing Mathematical Reasoning in Small Language Models (SLMs) 🔗	Sept 2024 - Dec 2024
<ul style="list-style-type: none">Fine-tuned a T5-small transformer using LoRA and Chain-of-Thought prompting, improving reasoning accuracy by 12%.Exposed inference via Flask / FastAPI APIs, enabling integration with downstream applications.	
ApplyCation: Automated Job Application Platform 🔗	Oct 2024
<ul style="list-style-type: none">Built a Flask-based backend with Selenium automation to auto-fill job forms, cutting application time by 50%.Integrated secure user authentication and GCP Firestore storage with a lightweight Streamlit UI.	

SKILLS

Languages:	Python, Java, SQL
Backend & Cloud:	Flask, REST APIs, AWS, Docker, GCP, CI/CD, Distributed Systems
Machine Learning & AI	PyTorch, TensorFlow, LLM, RAG, LangGraph, OpenCV, LangChain