

# Sai Paresh Karyekar

 Sunnyvale, CA  (404) 200-3907  [sai.karyekar@gmail.com](mailto:sai.karyekar@gmail.com)  [github.com/saikaryekar](https://github.com/saikaryekar)  [linkedin.com/in/sai-karyekar](https://linkedin.com/in/sai-karyekar)

## EDUCATION

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

## EXPERIENCE

<b>Amazon, Palo Alto   Software Development Engineer</b>	<b>June 2025 - Present</b>
<ul style="list-style-type: none"><li>Scaled <b>Visual Search platform</b> to 10+ new marketplaces by provisioning infrastructure with AWS CDK, establishing secure cross-account access, and automating CI/CD deployments.</li><li>Led <b>production migration</b> of a core request aggregation service from Dublin to Spain, ensuring zero downtime through rigorous testing and monitoring.</li><li>Designed and documented <b>system architecture</b> for integrating Visual and Core Search, including service decoupling and scalable API design; mentored an intern on an offline batch localization pipeline.</li><li>Enhanced platform security by deploying a <b>MapToken-based hybrid authentication system</b>, A/B tested to mitigate bot-driven DDoS traffic and block unauthorized access without affecting customer experience.</li></ul>	
<b>NVIDIA, Santa Clara   Software Engineering Intern</b>	<b>June 2024 - Aug 2024</b>

- 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 retrieval
- Spawned **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

<b>Enhancing Mathematical Reasoning in Small Language Models (SLMs)</b> 	<b>Sept 2024 - Dec 2024</b>
<ul style="list-style-type: none"><li>Fine-tuned a T5-small transformer using <b>LoRA</b> and <b>Chain-of-Thought prompting</b>, improving reasoning accuracy by 12%.</li><li>Exposed inference via Flask / FastAPI APIs, enabling integration with downstream applications.</li></ul>	

<b>ApplyCation: Automated Job Application Platform</b> 	<b>Oct 2024</b>
<ul style="list-style-type: none"><li>Built a Flask-based backend with Selenium automation to auto-fill job forms, cutting application time by 50%.</li><li>Integrated secure user authentication and <b>GCP Firestore</b> storage with a lightweight Streamlit UI.</li></ul>	

## SKILLS

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