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 **May 2025**
Master of Science in Electrical and Computer Engineering *GPA: 3.9/4*

Veermata Jijabai Technological Institute (VJTI), India **June 2023**
Bachelor of Technology in Electronics & Telecommunication Engineering *CGPA: 9.62/10*

Experience

Amazon, Palo Alto | *Software Development Engineer* **June 2025 - Present**

- Scaled Amazon **Visual Search infrastructure (25M+ MAU, 1K+ TPS)** to 10+ global marketplaces using AWS CloudFormation (IaC) and CloudWatch-based CI/CD pipelines for automated deployment and monitoring.
- Led cross-region **production migration** of a core request aggregation service with zero downtime through load testing, auto-scaling validation, and canary rollouts across Python and Node.js (TypeScript) services.
- Designed the **system architecture** for integrating Visual and Text Search, enabling scalable APIs and fault-tolerant service decoupling.
- Enhanced platform reliability and security by deploying a **MapToken-based hybrid authentication system** and improving observability via Log4j logging, CloudWatch metrics, and alerting.

NVIDIA, Santa Clara | *Software Engineering Intern* **June 2024 - Aug 2024**

- Built an internal **Generative AI content generation tool** using RAG, reducing blog creation time by 2.5×.
- Developed and deployed a **vector search pipeline** with Milvus for efficient document retrieval.
- Configured and secured **AWS EC2 infrastructure** with firewall rules, VPN access, and load-balanced scaling.
- Containerized and orchestrated deployment with **Docker** and AWS ECS for scalable production rollout.

Projects

ApplyCation: Automated Job Application Platform **Oct 2024**

- Developed a Flask backend with Selenium automation and **Claude API** for keyword-based resume tailoring.
- Leveraged **GCP Storage** for persistent user data and deployed a Streamlit UI for user input and job tracking.

Enhancing Mathematical Reasoning in Small Language Models (SLMs) **Sept 2024 - Dec 2024**

- Fine-tuned a T5-small transformer using **LoRA and Chain-of-Thought prompting**, improving model reasoning accuracy by 12%.
- Exposed inference via Flask APIs for seamless integration with downstream applications.

Skills

Languages: Python, Java, TypeScript, SQL
Systems Backend & Cloud: Flask, Node.js, REST/gRPC APIs, GCP, Docker, Distributed Systems
Infra & AI: OpenSearch, Terraform (IaC), CloudWatch, PostgreSQL, DynamoDB, Redis, PyTorch, TensorFlow, LangChain, MLflow, OpenCV