# Santosh Chaluvaraju

Machine Learning Engineer | Ex-Machine Learning Engineer Intern at Object Computing

**Chicago, IL, 60616 | Ph: +1 312 478 2361 | email: schaluvaraju@hawk.iit.edu | Portfolio | LinkedIn**

AI Engineer with 4+ years of work experience having robust leadership skills and in-depth AI system comprehension. Proficient in Large Language Model development and, crafting scalable end-to-end AI pipelines, from pure research to minimum viable products to scalable production ready deployments.

## EDUCATION

**Masters Of Science in Data Science** | Jan 2022 - Dec 2023 | Illinois Institute of Technology (**GPA: 3.7/4.0**)

## RELEVANT COURSEWORK

**Applied Statistics, Machine Learning, Deep Learning, Big Data Technologies, NLP, Data Structures and Algorithms**

## PROFESSIONAL EXPERIENCE

**OBJECT COMPUTING** | St. Louis, USA | ***MACHINE LEARNING ENGINEER INTERN*** **Jun 2023 - Aug 2023**

- Leveraged Google Cloud Platform to segment images at scale using Segment Anything Model (SAM) from META.
- Developed Custom prediction routines and containers to deploy SAM to an endpoint for Online Predictions using VertexAI.
- Built Batch Prediction pipelines using Kubeflow to segment images in batches and tested against batch size of up to 1000 images.
- Finally documented and Open Sourced the work in Github repo and also gave a presentation on the usage and impact.

**WERIZE (Financial Tech)** | INDIA | ***SOFTWARE ENGINEER*** **Feb 2021 - Dec 2021**

- Built a Bot model to emulate credit manual underwriting process, resulting in a 60% improvement in operational efficiency.
- Integrated Google Auth to the web app, added API filters in backend to grant access to users only within the organization.
- Achieved decoupling across microservices in processing daily EMI collections by integrating AWS Simple Queue Service(SQS).

**HIGH PEAK SOFTWARE** | INDIA | ***SOFTWARE ENGINEER*** **Aug 2018 - Feb 2021**

- Designed and developed server side RESTful web services and also involved in database design for the ADP project.
- Built a messaging service facilitating seamless client-end user communication, resulting in an 50% efficiency boost.
- Lead a team of five and took self initiatives to resolve blockers and develop ownership consciousness within the team.

**WINIX TECHNOLOGIES** | INDIA | ***SOFTWARE ENGINEER*** **Jun 2017 - Aug 2018**

- WIP, A comprehensive Mesh Network Technology. Worked on pairing, discovery, and network association parts of WIP.
- Integrated Luci web interface with WIP and implemented device management and firmware upgrade of devices.
- Actively involved in Bug Fixing and end to end testing of the platform before every release.

## SKILLS

**Programming and Database -** Python, R, C, C++, Java, Javascript, SQL, MongoDB, DynamoDB
**ML Algorithms** - SVM, Regression, Random Forest, XGBoost, PCA, Decision Trees, Neural Nets, Transformers
**MLOps -** Kubeflow, Airflow, Docker, AWS(Amazon Web Services), GCP(Google Cloud Platform), Sagemaker, VertexAI
**Frameworks & Libraries -** Sklearn, Spark, Hadoop, Kafka, Pytorch, TensorFlow, Spring, FAST API, LangChain, Hugging Face
**Project Management & Version control -** Jira, Asana, Confluence, Git, DVC

## CERTIFICATION

**Generative AI with Large Language Model - [Certificate] -** DeepLearning.AI - [Oct 2023]
**AWS Certified Cloud Practitioner - [Certificate] -** Amazon Web Services - [Validity: Sep 2023 - Sep 2026]
**Neural Networks and Deep Learning** - **[Certificate]** - DeepLearning.AI - [May 2023]
**Natural Language Processing with Python** - **[Certificate]** - Udemy - [Apr 2023]

## PROJECTS

**Fine-Tune FLAN-T5 Model with Reinforcement Learning (PPO) - [Project]** **Oct 2023**
Fine-tune a FLAN-T5 model to generate less toxic content with Meta AI's hate speech reward model. Proximal Policy Optimization (PPO) is used to fine-tune and reduce the model's toxicity.

**Fine-Tune a FLAN-T5 Generative-AI Model with PEFT- [Project]** **Sep 2023**
Fine-Tune an existing Large Language Model FLAN-T5 from Hugging Face for Enhanced Dialogue Summarization.

**META SAM Model workflow in Google Cloud Platform - [Project]** **Aug 2023**
Deploy SAM on GCP using VertexAI for Online Predictions. Built Batch Prediction Pipeline using Kubeflow.

**Chat Bot, Based on End to End Memory Networks - [Project]** **Mar 2023**
Question Answer model trained on Facebook Babi Dataset, Based on the Paper End to End Memory Networks.

**Scalable End to End Machine Learning Data Pipeline with Kafka Streaming in AWS - [Project]** **Nov 2022**
Simplifying MLOps, A Multi-threaded python app capable of handling end to end ML Workflow from data engineering to predictions.