Santosh Sawant

(+91) 7829674405

contactmeurgently@gmail.com

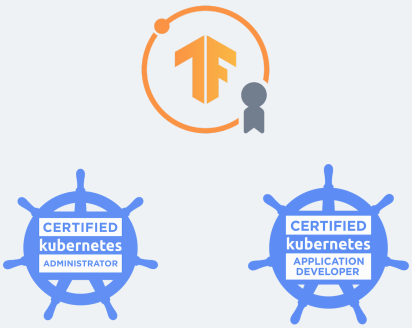
santoshsawant

ssawant.github.io

LLM Architect learning to innovate, optimize, and scale the next generation of large language models.

.

CERTIFICATION



SKILLS

Libs & Frameworks: *Pytorch, Tensorflow, Nvidia NeMo Hugging Face TRL, vLLM, DeepSpeed-FastGen, Autogen, CrewAI*

Programming Language : *python, golang, rust, Java, javascript and c++.*

WORK EXPERIENCE

Sr Solution Architect Generative AI @ Philips. (2024 - present)

Working towards enabling LLM powered Kennis AI search assistant with a goal of reducing 5 ~ 10 % of overall Philips customer care workload.

- *Developing a Multi Agent modular platform, which enables org wide internal search and actions.*
- *Fine-tuning LLM with a multi LoRA adaptor for Agent routing.*

Sr Architect Machine Learning @ Tredence Inc. (2022 - 2024)

Leading a cross-functional team to deliver a cutting-edge Generative AI platform and products in the Data Analytics, Healthcare and ESG domain.

- *Fine-tuning LLM with multi LoRA adaptor for domain specific task in Retail, CPG and Supply Chain domain*
- *Developed distributed cloud GPU training approaches for LLMs models, including data distribution editing, data quality improvements, and representation learning with self-supervision.*
- *Experience in reading papers and implementing algorithms described in papers to increase performance, quality, data management, and accuracy of AI systems.*
- *Designed lean proofs of concepts (POC) to answer targeted business questions using Gen.AI.*
- *Design, Developed and integrated various large-scale, distributed machine learning systems for production ready Gen.AI services.*

Lead Machine Learning Engineer @ Parallel Wireless (2021 - 2022)

Lead Machine Learning engineer at parallel wireless. Responsible for developing end-to-end machine learning pipeline along with defining and executing ML workflows.

- *Developed a key performance index forecasting system using LSTM with 70% accuracy; entailing reduced network downtime with improved log-collection triggering.*
- *Defined and executed specific ML workflow, which includes data collection, sampling, model building and training, metrics definition and evaluation.*

Principal Engineer @ OLA (2015 - 2021)

Lead engineering team at OLA. Responsible for developing and scaling various data, machine learning and application pipelines.

- *Develop services and components for deploying Airflow and flink clusters on kubernetes and onboarded/scaled various Data and BI spark and flink jobs pipelines.*
- *Develop services and components for feature store integration with existing data and ml pipeline*

SDET - 2 @ Zynga (2011 - 2014)

SDET - 2 @ Pengala (2010 - 2011)

SDET - 1 @ Dell - EMC (2007 - 2010)

WORK PROJECTS

Kennis AI Search Assistant (Philips.)

Working towards enabling LLM powered Kennis AI search assistant with a goal of reducing 5 ~ 10 % of overall Philips customer care workload..

Generative AI Products and Platforms for Data Analytics, Healthcare and ESG Domain (Tredence Inc.)

Research, design, develop and build various GenAI LLM base products and services in the following domain.

- *Trends - Copilot for retailers and CPG domain powered by finetune Llama 3 8B modal using multi LORA adaptor.*
- *YODAI - Your Own data analysis intellect. Utilize Multi agents LLM framework to develop and automate various retail, CPG & SCM data analysis workflows*
- *Data Whizz. Utilize agent framework to generate & execute SQL /py queries for quick insight of DB*
- *Healthcare - Differential DX and Prior Auth .*
- *ESG (Environment, Social and Governance) reporting*

KPI Early Warning System using time series forecasting (Parallel Wireless)

- *Created an LSTM autoencoder based Anomaly detection model to predict the outlier occurrence of KPIs with 70% accuracy; optimized run-time of model with PCA dimensionality reduction*

Cell Auto configuration using Deep Reinforcement Learning (Parallel Wireless)

Cell Auto configuration uses a deep reinforcement learning model to optimize the "tilt" and "power" settings of a cell ENodeB. Develop RLOps pipeline and integrate it with ns3 simulator (digital twin) and OpenRAN systems.