Arvind Sharma

+91-9176076208 | karvind1998@gmail.com | linkedin.com/in/arvindsharma18 | github.com/ArvindSharma18

EDUCATION

SASTRA Deemed to be University — **Thanjavur, Tamil Nadu**
Bachelor of Technology in Computer Science Engineering — *June 2020*

- **Cumulative GPA:** 8.5411
- **Awards & Honors:** Dean's Merit List(Top 10% in CS Department) for Academic Year 2016-17,2017-18,2018-19,2019-20;

TECHNICAL SKILLS

Languages: Python, Java, YAML, C++

Machine Learning Packages: NumPy, Pandas, scikit-learn, Tensorflow, Pytorch , keras, transformers, trl, unsloth, accelerate, ray, deepspeed, FSDP, flash-attn, LLaMA-factory, axolotl, torchtitan, torchao, langchain, llama-index, llama.cpp, vllm, ollama, lorax, DSpy, whisper.cpp, crewAI, guidance, outlines, opencv, diffusers, ultralytics, sentence-transformers, spaCY, h2o, bentoML, dask, MLFlow, Kubeflow, qdrant, faiss, haystack, chroma, milvus

Frameworks: Docker, Kubernetes, Istio, FastAPI, Springboot, Postgresql, Git, Helm Charts, Ansible, Jenkins, Redis, Terraform, Kafka, RabbitMQ, celery, gradio, streamlit, ReactJS

Cloud Platform: AWS (EKS, EC2, Cloud Formation, ECR, IAM, S3), GCP (GKE, Buckets, VM, Cloud Vision API), Azure (AKS, VMs, Disk, OpenAI)

Developer Tools: Chat-GPT, MS-Office Tools, Adobe Express, VSCode, WSL, Anaconda, GitLab/Github, Swagger editor, Postman

WORK EXPERIENCE (4+ years)

Tata Consultancy Services — Chennai, Tamil Nadu

Machine Learning Engineer/ Lead Developer — *October 2020 – Present*

- *CTO Business Unit(August 2022 – Present)*
  - Machine Learning Engineer and Lead Developer, handling requirement analysis, design and delivery of MVPs for a transformative AI powered product at scale. Expertise in scaling LLMs on multi GPU (A100) environment, fine tuning and aligning LLM models on domain data, creating LLM powered Agents and complex pipelines using prompt engineering.
  - Consulting, Ideating and Building Generative AI solutions/pipelines use cases and POCs for various Business domains/verticals.
  - Use cases include fine tuning and aligning open source models on custom datasets, designing and leading a team to create a workflow using OpenAI and fine-tuned Stable Diffusion XL model with refiner, Domain pipeline automation using GPT.
  - POCs include Talking Avatars, Sales Bot and Automation using LLM powered Agents, Prompt Engineering with Open Source model on local consumer machines and consumer GPUs.
  - Worked with AI open source projects like Auto GPT, Stable Diffusion Web UI, llama.cpp, vLLM, crewAI, Langchain, TRL etc.
  - Market Research and Data Analysis on business related data to enable strategic business choices on emerging technology.
  - Studying and Presenting new Generative AI research and Solutions to multiple stakeholders. I have written multiple whitepapers and submitted within internal forums.
- *Corporate Labs (STIL & CMI Rapid Labs), TCS (July 2021 – July 2022)*
  - Developed and documented Sales based use cases using NLP techniques as part of Sales and Technology Innovation Lab (STIL)
  - In Rapid Labs, developed rapid state of the art Artificial Intelligence solutions (NLP, Image Processing/Computer Vision, Graph NN) as part of POCs and use cases for various Clients across Communication Media and Information Unit of TCS.
  - POCs include Speaker Diarization, Automated Image Annotation using Vision API, Q&A tasks using Transformers.
  - Gained competency in 5G core technology and worked on orchestration of application containers using Kubernetes(EKS).
  - Use cases include Edge Video Analytics application, Application platform for Ecosystem based solution. I took responsibility for Development, Automation of Infrastructure using Cloud Formation and Ansible, Container Orchestration using Kubernetes(EKS) and Network Mesh using Istio for the mentioned use cases.
  - Studying and presenting multiple emerging technology across domains and open source libraries to multiple stakeholders.
- *Telefonica, UK (December 2020 – June 2021)*
  - Took responsibility for monitoring of Big Data modules, which provides key insights about the customers.
  - Part of production upgrade Team, took responsibility for Analytics application's compatibility with upgraded environment. Achieved zero data loss.

Universitat Autònoma de Barcelona — Barcelona, Spain

Research Intern (Exchange Student), Centre de Visió per Computador — *February 2020 – July2020*

- *Acquisition of Hyper Spectral Images and Pre-Processing*
  - Collection and processing of multiple bands belonging to different spectral intensities to form an image band cube.
- *Semantic Segmentation Network for Classification*
  - Trained U-Net model with band cube of multi and hyperspectral images and its corresponding annotated Images as inputs.

BroadSoft **Chennai, Tamil Nadu**

Developer Intern *November 2017 – December 2017*

- Got hands on experience with Software Development Life Cycle. Worked under development team.

PROJECTS

- **Clinical Trial LLMs**
  - Fine tuned small language models (Llama3, Phi 3 mini, Qwen 2.5 3B) on consumer GPUs using custom dataset assorted for clinical trials. Used deepspeed, and data parallelism to speed up the process on multi GPU setup.
  - Aligned the fine tuned models using techniques like ORPO and DPO using custom dataset collected from fine tuned models.
  - Published some models on Huggingface repo under open source license.
- **Parallel LLM Inference Pipeline**
  - Setup multiple scripts using different techniques like batching, parallel inference etc. for fast LLM inference on multi GPU environment.
- **Omni-Model**
  - A small project demonstrating multi-modal capability on CPU machine using ensemble of models namely whisper, x-gen-mm-phi3 and nemo tts model. The flow will take in a video as input and provide audio as output. Completed end to end in 2 days.
- **Automatic Image Masking**
  - Automatic Image mask generation using a pipeline utilizing SAM and Dino model. Entire process was demonstrated on Nvidia GTX 1660-Ti 8 GB graphic card laptop.
- **Executive Summary**
  - Created a pipeline that generated an executive summary of a document. Used Falcon 40 B model and Flan-T5 model for generating summary. Utilized custom Map-Reduce and Refiner setup to achieve coherent summary.
- **Appointment Scheduler**
  - Created an end-to-end application for scheduling appointment. Used Springboot for server and ReactJS for frontend.

ADDITIONAL

- A keen writer having published technical and business articles in both private and public forums[1, 2]
- Cricket player having represented school teams and club teams in State level tournament. Football player having represented school for Inter-School tournament. Follow Badminton, Tennis and Mixed Martial Arts, Motor Sports closely.
- Hobbies include Sketching, Movies and Anime, Photography, Music and Books.