Sharath Pai

+91-7045883710 | sharathpai107@gmail.com | linkedin.com/in/sharathpai107 | github.com/Sharath1036

Education

B.Tech in Electronics and Telecommunication — Dec 2020 – May 2024
Dwarkadas J. Sanghvi College of Engineering — GPA: 7.66

Higher Secondary Education — Jul 2018 – Mar 2020
Patkar Varde College of Science, Commerce and Arts — Percentage: 83.23

Experience

Software Development Engineer, LLMs | Salk AI — Oct 2024 - Feb 2025
Experience Letter — Nagpur, India

- Benchmarked performance of **LLMs, VLMs, STTMs** with **prompt engineering** approaches like **one-shot, few-shots, CoT**.
- Developed **Proof of Concepts** (POCs) involving **LLMs**, **OpenCV** and **NLP** based practices.
- Traced the model API calls on **Langfuse** by integrating LLMs with **LiteLLM** and implementing **fallback** strategies.
- Hosted open source models using **vLLM** with on **NVIDIA A100 GPU** and tested their performance with their endpoints.
- Integrated AI into **development** environment by creating their **RESTful backend APIs** and user-friendly frontend interface.
- Created automation scripts using **GitHub Actions CI/CD** and deployed apps into **production** on an **Amazon EC2** instance.

AI/ML Research Intern | Avignon Université — Jun 2024 - Jul 2024
Internship Certificate — Avignon, France

- Conducted research in **Bioinformatics** to analyze the impact of UV rays on living organisms using ML algorithms.
- Developed datasets through extracting data from sources like **NCBI** and **Springer**.
- Tuned ML models to predict UV ray impact, leveraging **ANN architectures** with **BatchNormalization** and **Dropout layers**.
- Implemented predictive modeling techniques, ensuring model accuracy with **Early Stopping** and **model-saving callbacks**.
- Integrated best-performing models into a **Django-based web application** for real-time UV ray impact prediction.

Projects

Video Customer Identification Process — Oct 2024 – Nov 2024
Tech Stacks: Python, OpenCV, HuggingFace, YOLOv8, NVIDIA A100 GPU, Next.js, FastAPI, MongoDB

- Reduced KYC onboarding time to 30 seconds, **cutting drop-offs by 35%**, enabling **5,000+ automated verifications/month**.
- Prevented fraud using blink detection and anti-spoofing using **OpenCV**, **reducing fake submissions by 60%**.
- Enabled multilingual support via **Whisper V3**, removing language barriers and **transcription accuracy upto 90%**.
- Automated document processing with **YOLOv8** & **Qwen-VL** with **95% OCR accuracy**, preventing human verification.
- Stored the KYC data in a **MongoDB database**, also displayed on the **Next.js frontend**.

Multi-Agentic RAG based Question & Answering — Mar 2025
Tech Stacks: Python, Agno, MongoDB, OpenAI, Ollama, FastAPI, Docker, GitHub Actions, Amazon EC2 — Project Link

- Built a pipeline using **Agno** for processing texts from **knowledge bases** like PDF URLs, Wikipedia and Web URLs.
- Initialized knowledge base and performed text **embedding** using **Ollama's openhermes:v2.5** model.
- Stored the embeddings in a **MongoDB database** consisting of **1536 dimensions** and configured **Vector Search** in Atlas.
- Utilized **OpenAI** Chat for generating a response from the vector database through search index.
- Developed **RESTful APIs** of each agent in FastAPI and containerized the application stack using **Docker**.
- Created a **CI/CD Pipeline** that pushes the code to **Dockerhub** and runs the Docker container on an **EC2 machine**.

College Reviews Sentiment Analysis using Transformers — Sept 2024
Tech Stacks: Python, spaCy, Transformers, PyTorch, Matplotlib — Project Link

- Applied preprocessing techniques like **tokenization**, **stop-word removal**, and **text normalization** to prepare data for analysis.
- Applied **DistilBERT** for sentiment analysis reducing **60% less parameters** compared to traditional BERT models.
- Developed a pipeline with **vectors** and **classification models** to benchmark performance and compare with DistilBERT.
- Achieved an **85% accuracy** in sentiment classification using DistilBERT with efficient text preprocessing and model tuning.

Publications

Predictive ML for Educational Decisions | 12th International SMART 2023 Conference — Dec 2023
Authors: S Pai, A Wahedna, H Shaikh, S Gosavi, Prof. T Sawant — Publication Link

- Authored a research paper for analysing trends revolving Engineering admissions using **supervised learning** algorithms.
- Created a custom dataset from scratch involving around **400 observations** of 17 variables for training and test sets.
- Using **Regression**, we predict cutoffs of colleges and in **Classification**, we identify probability of student choosing a college.
- Evaluation was done by finding the significant features and noting its effect on metrics (MSE, Standard Error, Accuracy).
- The **Bagging** model yielded the least **MSE of 162.1** and the **XGB** model yielded the best **accuracy of 93.38%**.
- A **Tableau dashboard** was prepared after preprocessing for easy analysis of features.

Technical Skills

Programming: Python, R, JavaScript, SQL, LaTeX
DevOps: Docker, GitHub Actions (CI/CD), AWS (EC2, Sagemaker)
Data Science: Tensorflow, NLTK, spaCy, PyTorch, OpenCV, Langchain, NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn
Development: Django, FastAPI, React JS, Next.js, Handlebars, HTML, CSS, Bootstrap, Tailwind
Databases/Vector DBs: MySQL, MongoDB, Qdrant, ChromaDB, Pinecone
Operating Systems: Windows, Linux