PRASAD P R

+919740146523 | shenoy.prasad20@gmail.com | Bengaluru, Karnataka, India | LinkedIn | GitHub | Website

SUMMARY

Accomplished **AI Engineer** with **8+ years of expertise** in designing, developing, and deploying both traditional AI/ML and LLM based solutions at scale. Proven success in **optimizing LLMs, generative AI, and computer vision models** for production, with deep proficiency in **Python, PyTorch, TensorFlow, Langchain and OpenVINO**. Experienced in building **scalable pipelines, RAG pipelines, agentic workflows, model optimization frameworks, and AI-driven applications** across NLP, RAG. Adept at **leading AI projects end-to-end**, mentoring engineers, and delivering solutions that drive measurable business impact.

SKILLS

Programming Languages: Python, JavaScript, C++
Machine Learning Techniques: Deep Learning, Generative AI, LLM, NLP
Data Processing and Analysis: Computer Vision, Image and Video Processing, Quantitative Analysis, Data Visualization Techniques, NoSQL
Software Development: Application Design, Microservices, Git, Docker, Kubernetes
Generative AI & LLMs: LangChain, Hugging Face Transformers, OpenAI API, Azure OpenAI, Agentic AI, RAG (Retrieval-Augmented Generation), Prompt Engineering, Vector Databases (Pinecone, FAISS, Weaviate).
Frameworks & Libraries: Pytorch, TensorFlow, OpenVINO, ONNX.
MLOps & Deployment: Fast API for AI microservices, Docker, Kubernetes, Django REST Framework
Databases: PGSQL, MongoDB

PROFESSIONAL EXPERIENCE

AI Software Development Engineer, **Intel Corporation** *Jun '22 — Present*

- Achieved a **10x performance enhancement** on Intel GPU by optimizing Huggingface transformer models (**Whisper and M2M100**).
- Implemented an **automated benchmarking pipeline** of Transformer models across Intel platforms, reducing the time per machine from 5 hours to 30 minutes.
- **Optimized model performance** using ONNX runtime and OpenVINO, **reducing analysis time by 50%** from days to minutes.
- Developed a **Real-Time 3D Camera Calibration strategy**, reducing inference time by 80% **using Python, PyTorch, Open3D, and Docker.**
- Enhanced AI models, boosting training and inference speeds by 30%, and incorporated the solution into an IoT-enabled device with features for **MQTT and video feed operations**.
- Created a real-time visualization application and an efficient inference pipeline for **segmentation and labelling of point clouds** utilizing **Python, Intel RealSense Sensor, PyTorch, and Open3D,** leading to a more than 50% decrease in inference time.

System Software Development Engineer, Intel Corporation *Aug '18 --- Jun '22*

- Led the development of a **MERN stack Solution Delivery System**, that garnered over 1000+ downloads.
- Improved access time by over 30% through remodelling the architecture from **monolithic to microservices based-architecture**.
- Guided the development of a management portal, integrating CICD tools, reducing maintenance time by 80%.

INTERNSHIPS

Software Engineering Intern, Intel Corporation *Oct '17 --- Aug '18*

- Contributed code to 2 repositories in the GitHub Archives program as a part of **Hyperledger Sawtooth Enterprise Blockchain**.
- Built **AngularJS UI** for blockchain visualization and monitoring
- Developed POCs for **Stereo Vision, Computer Vision, and AI-based node analytics**.

EDUCATION

Liverpool John Moores University - **Liverpool** & upGrad *Apr '25 -- Present*
M.Sc. in Artificial Intelligence and Machine Learning

IIIT-B & upGrad *Feb '24 — Apr '25*
Executive PG Diploma in Artificial Intelligence and Machine Learning *3.63/4 CGPA*

JSS Academy of Technical Education, Bangalore *Aug '14 — Jul '18*
Bachelor of Engineering (B.E.), Computer Science & Engineering Grade – Distinction

KEY PROJECTS (Above and Beyond my Work)

Find Me A Cube (Internal tool)

Objective: No cost solution for identifying empty cubes in a Hybrid Model. Also provide cube usage analytics. **Solution:** Cloud hosted solution using Django REST API as a backend, React JS based frontend and Mongo DB database. **Result:** Faster, hassle free cube identification for employees the clean frontend and faster backend.

Automated Smart Reports using Generative AI (Internal Tool)

Objective: Speed up report collation, reduce manual report generation based on multiple mails. **Solution:** Using Gen AI & prompt engineering, extract required information from multiple reports and combine the information into a single status report. **Result:** Reduced the managers time spent on report generation from 5 days to 30 minutes.

Smart Hiring and Resume Screening using Generative AI (Internal Tool)

Objective: Help hiring manager screen suitable resumes based on the job description required for the role. **Solution:** Using LLM models, understanding the data of the resume and then trying to perform a similarity mapping with the job description to obtain the top resumes that are suitable for the job. **Result:** Helped managers reduce the resume screening time from hours to minutes.

Enterprise RAG Assistant (Internal Tool)

Objective: Having an assistant to look through a huge set of enterprise documents and return suitable documents matching the keywords. **Solution:** Retrieval-augmented generation system using LangChain + PGSQL, hosted on IBM Cloud Foundry using Docker images to enable natural language querying across 10k+ enterprise documents. **Result:** Top 'n' number of documents are shared back to the user based on their query.

Finance and Stock Markets

Objective: Fetch the financial details of a company based on their ticker and perform the fundamental analysis. **Solution:** Using Agentic AI, fetch the company data and using Python pandas perform fundamental analysis of the company to suggest if the company stock can be purchased or not. All this is performed using phidata module and **agentic AI**

AWARDS & CERTIFICATIONS

Divisional Recognition Award (DRA)	Intel
Innovation Star	Intel
Deep Learning With Pytorch Image Segmentation	Coursera
LangChain- Develop LLM powered applications with LangChain	Udemy
Generative AI with Large Language Models	Coursera