Purusharth Malik

+919084356064 | purusharth19malik@gmail.com | LinkedIn | GitHub | Blog

Education

CHRIST (Deemed to be University) Bangalore, Karnataka
Master of Sciences (Artificial Intelligence and Machine Learning), 3.92/4.0 *July 2023 – May 2025*

University of Delhi New Delhi
Bachelor of Science, Honours in Mathematics, 8.27/10.0 *August 2019 – May 2022*

Experience

AI Engineer June 2025 – Present
Swiggy *Bangalore, Karnataka*

- Technology Stack: Python, LangChain, LangGraph, PyTorch, HuggingFace, DSPy
- Optimized agentic LLM workflows for production environments, focusing on latency reduction and response quality.
- Developed a custom evaluation framework utilizing RLHF and RAG metrics to systematically measure and improve customer-facing LLM performance.
- Applied DSPy to declarative LLM programming, systematically optimizing the prompting and composition of multi-turn conversational agents to boost overall dialogue accuracy and reliability.

Software Engineer - Artificial Intelligence Feburary 2025 – May 2025
Metrum AI *Bangalore, Karnataka*

- Technology Stack: Python, PyTorch, Docker, LangGraph, AWS, LangChain, HuggingFace, vLLM, SGLang
- Engineered GPU-accelerated AI applications for hardware capability benchmarking in collaboration with AMD and Dell.
- Implemented Stable Diffusion and custom latent-space models to develop digital avatar plug-ins, streamlining asset generation via prompt engineering.
- Benchmarked and deployed high-throughput LLM serving infrastructure leveraging vLLM with PagedAttention and SGLang's flexible LLM orchestration to achieve a 3.5× increase in QPS (Queries Per Second) for inference APIs under peak load.

Machine Learning Engineer October 2024 – January 2025
SMOOR Chocolates *Bangalore, Karnataka*

- Technology Stack - Python, AWS, HuggingFace, PyTorch
- Worked with the finance team to automate inventory management pan India
- Worked on an agentic framework to improve the customer service pipeline

Data Scientist August 2024 – October 2024
NayaOne *Remote, London*

- Technology Stack - Python, PyTorch, HuggingFace, AWS EC2, Docker, Kubernetes, CrewAI
- Provided automation for sales and marketing teams
- Created multi-agent systems to perform trend analysis

LLM Engineer May 2024 – August 2024
Iolite Technologies *Bangalore, Karnataka*

- Technology Stack - Python, PyTorch, HuggingFace, Flask, AWS EC2, React.js
- Created an LLM-driven chatbot for the university's ERP system
- Finetuned multiple pre-trained LLMs after performing quantization using LoRA

Projects

Swin-VQA | *Python, PyTorch, Transformers, OpenCV, Pillow* Source Code

- Created a visual question-answering system for assistance in the field of radiology.
- Used CLIP for visual encoding and a fine-tuned Llama-2 model for text encoding.
- Created custom decoder layers inspired by DeiT and Swin Transformers.

AutoQuote | *Python, PyTorch, LangChain, MongoDB, Flask, JavaScript, HTML, CSS, Pinecone* Source Code

- A web application that is able to automate the process of generating quotations for high-end hotel chains.
- Integrated language models with MongoDB and Vector DB to generate a quoted bill-of-quantity in the form of an Excel document.
- The project is in collaboration with Shanti Metal Industries and is currently in production.

Technical Skills

Python, PyTorch, Tensorflow, HuggingFace, Accelerate, TRL, vLLM, SGLang, TensorRT, ONNX, DSPy, LangChain, LangGraph, LlamaIndex, OpenCV, Docker, Kubernetes, Flask, MongoDB, MySQL, Redis, AWS, Milvus, Pinecone