Vishnu M

Machine Learning Engineer

github : vishnu80152

Portfolio : https://vishnu-ai-nexus-sphere.lovable.app/

gmail : vishnu80152@gmail.com

linkedin : vishnu-m-015459324

+91 8015255825

Coimbatore

PROFILE

I am a **Machine Learning Engineer** with expertise in **deep learning (CNNs, RNNs), NLP, model quantization, and RAG-based chatbot development**. Skilled in deploying **ML solutions** using **TensorFlow, PyTorch, and LangChain**, I also have experience in **building Flask applications, creating RESTful APIs, and designing backend pipelines** for efficient automation.
I am actively seeking **ML Engineer, AI Engineer, or related roles** where I can apply my skills to develop innovative AI solutions and contribute to impactful projects.

WORK EXPERIENCE

Machine Learning Engineer | Pinaca Technologies (06/2024–Present)

- **Engineered** end-to-end RAG-based chatbot with **95% query accuracy**, integrating document retrieval and answer generation, and **automated** workflow execution that **reduced manual oversight by 70%**.
- **Built** modular NLP pipelines—including translation, speech-to-text, and speech-to-speech—**expanding multilingual support by 40%** and streamlining communication workflows.
- **Applied** 8-bit quantization to Transformers and LLMs, **halving model size** and **increasing inference throughput ~40%**, enabling scalable edge deployment.
- **Deployed** production-grade CNN/RNN vision models at **90–95% accuracy**, delivering client-ready computer vision solutions.
- **Engineered** audio-processing suite (diarization, denoising, speaker ID, background removal), **boosting transcript accuracy by 25%** in noisy conditions.
- **Automated** benchmarking pipelines, **cutting model validation time by 80%** and accelerating the model optimization process.
- **Deployed and optimized** GeoCLIP for image-to-location tasks, **reducing latency by 30%** and improving geospatial tagging accuracy.
- **Optimized** LLM fine-tuning and prompt strategies, **increasing NLP performance by 30%** across summarization and Q&A tasks.
- **Crafted** advanced prompt-engineering frameworks and autonomous AI agents using LangChain, **orchestrating intelligent chatbot workflows** and reducing manual prompts by 70%.

Machine Learning Engineer Intern | Pinaca Technologies (06/2023–06/2024)

- **Conducted** end-to-end EDA and built ensemble classifiers (XGBoost, Random Forest, SVM) achieving **95% accuracy**, delivering robust predictive insights for downstream systems.
- **Developed** an OCR pipeline for handwritten logs with **90% precision**, accelerating data digitization and reducing manual extraction effort.
- **Built** a Chinese-to-English translation module, **speeding data processing by 40%**, improving throughput for multilingual datasets.
- **Engineered** NER models achieving **92% accuracy**, automating extraction of domain-specific entities and key insights.
- **Designed** data-parsing automation tools, **reducing manual processing time by 60%**, boosting overall pipeline efficiency.

EDUCATION

2020 – 2024

- **B.Tech Artificial Intelligence and Data Science -** *Sri Eshwar College Of Engineering* | ***CGPA** : 8.3/10* | **First Class**

SKILLS

Languages: Python | R | Java

Tools: Tensorflow | Pytorch | Keras | Pandas | Linux | Git | Docker | Numpy | Scikit learn | Langchain | Langflow | VScode | Flask | SQL | MongoDB | Qdrant | Huggingface | Ollama | AI Tools | Neo4j | Power BI | Tablue | Plotly

Domains: Machine Learning | Deeplearning | LLM's | Data Science | RAG System | Backend | MLops | NER | Transformers | Model optimization | Cloud | Database | Prompt Engineering | Pipelines

CERTIFICATES

1 . Machine Learning and Data science

2 . Ethical hacking

PROJECTS

RAG & General Chatbots

- **Developed** a RAG-based chatbot that parsed PDFs, Word docs, and text files with **95% accuracy**, producing concise, user-tailored summaries.
- **Engineered** multi-task prompt strategies and fine-tuned RAG pipelines, **cutting retrieval latency by ~30%** while maintaining response quality.
- **Built** autonomous AI agents using LangChain that orchestrate end-to-end workflows—dynamic retrieval, context reformatting, and response generation—**reducing manual intervention by 70%**.

Agentic RAG System

- **Designed and deployed** an agentic RAG framework where autonomous LangChain agents dynamically managed retrieval, context building, and generation—yielding **highly accurate, analytically-rich answers** and reducing prompt engineering overhead by 60%.

Model Quantization

- **Applied** 8-bit quantization to Transformer and LLM architectures, **halving model size** and improving inference throughput by **40%**, enabling efficient deployment on edge devices.

Automated Model Testing

- **Engineered** an end-to-end automated validation pipeline (CI/CD style), **cutting manual testing effort by 70%** and accelerating model rollout cycles by 50%.

NLP & Multimodal Suite

- **Built** a multilingual translation engine, extending support across 5 languages and **increasing data throughput by 45%**.
- **Developed** text summarization and NER pipelines achieving **90%+ accuracy**, enabling automated extraction of key insights.
- **Created** an integrated speech-captioning and speech-processing stack (ASR → translation → caption), enabling seamless multimodal content delivery.

OCR Data Extraction

- **Developed** an OCR pipeline for handwritten logs, **achieving 90% precision** and accelerating digitization throughput by 60%.

Predictive Classification System

- **Engineered** a real-time ensemble classifier (XGBoost, Random Forest, SVM) to augment OCR extractors, **boosting OCR accuracy by 15%** and reducing data errors in production.

PUBLICATIONS

- **Predictive Maintenance of Machine Tools, IEEE**
- **Emergency Medical System using ML, IEEE**
- **An Efficient Driver Drowsiness Detection Using Deep Learning, IEEE**