Tharani Tharan M

☎ +91-638-262-8726 ✉ tharanisak63@gmail.com in Tharanitharan-M

Summary

AI/ML - Application Engineer with an M.Tech from IIT Madras and expertise in time series forecasting, Python, AWS, and machine learning. Skilled in designing, building, and optimizing end-to-end ML systems to deliver actionable insights and improve operations. Experienced in fine-tuning large language models, building RAG agents, and automating SQL workflows. Passionate about creating impactful AI solutions in dynamic environments.

Technical Skills

Programming: Python, SQL, FastAPI, Streamlit.
Cloud & DevOps Tools: Docker, AWS (Boto3, S3, Athena, EC2, SageMaker), Airflow, Git.
ML & AI Frameworks: LangChain, LangGraph, Hugging Face, Statsmodels, Scikit-Learn, TensorFlow, SpaCy.
Data Analytics: NumPy, Pandas, SciPy, Power BI, MS SQL Server.

Professional Experience

Application Engineer - AI/ML **Mar 2025 – Present**
Ingenero Technologies Private Limited *Remote*

- Developed a **Prophet-based time series forecasting model** to predict power consumption with **MAPE under 10%**, enabling proactive energy management and operational efficiency.
- Implemented **covariate shift detection** using the **Kolmogorov–Smirnov (KS) test** and triggered model retraining pipelines when **data distribution shifted**, reducing resource usage by at least 15%.
- Performed data extraction and analysis of time-series data from the PI Database using **SQL queries via PI SQL Client**, deriving actionable insights to support operational decision-making.
- Extracted numerical data from design plots to **quantify deviation** from the ideal scenario in benchmarking analyses.

Modelling Engineer **June 2023 – Aug 2024**
AZG Consulting LLP - AsInt *Remote*

- Leveraged **ARIMA to model linear components** of time series data, achieving 95%+ stationarity confidence and optimizing (p, d, q) parameters via ACF/PACF plots and grid search.
- Refined **multi-layer LSTM on ARIMA residuals**, incorporating 5+ additional features and 270-step sliding windows, improving multi-step sales forecast accuracy by 15%.
- Deployed a **Dockerized FastAPI** app on **AWS EC2** for real-time predictions, loading model artifacts from **S3**.
- Formulated **AWS Data Lake S3 Intelligent-Tiering** & lifecycle rules for automated cost savings of at least **25%**.
- Engineered a **Prophet-based predictive maintenance** tool using sensor data, delivering **4-hour advance warnings** to prevent equipment slippage and minimize plant downtime.
- Played a pivotal role in delivering business solutions by collaborating with cross-functional teams, demonstrating innovation, technical proficiency, and a results-oriented approach.

Process Engineer - Data Analytics **Dec 2019 – Sep 2020**
Vedanta Resources *Rajasthan*

- Boosted product recovery from **88.4% to 91%**, resulting in an approximate **3% increase** in overall yield and significant operational gains.
- Performed hypothesis testing on production variables, identified inefficiencies that contributed to a **2.6% loss** in output.
- Identified suboptimal mold speed and tapping as root causes, leading to targeted adjustments that **enhanced process efficiency by 5%**.
- Wrote and executed 50+ SQL queries to extract, clean, and **validate datasets with 5,00,000 rows**, ensuring good data accuracy.

Personal Projects

LLAMA3.2:3b LLM based Agentic RAG for financial documents | *Ollama, Docling, LangChain, LangGraph, LangSmith*

- Constructed a self-directed **agent graph** system utilizing the **LLAMA3.2:3b** model, featuring tool-based retrieval, dynamic query rewriting, and conditional response generation delivering high-quality answers.
- Processed and chunked documents into 464 segments using **Docling** and **MarkdownTextSplitter**, embedded with **nomic-embed-text**, and indexed in **FAISS** to enable high-performance semantic retrieval.
- Engineered a **custom retriever** that **auto-filters relevant documents** based on query intent, eliminating the need for manual filters and improving retrieval precision across diverse query types.

Qwen2.5:7b LLM based SQL Agent with LangChain & LangGraph | *Ollama, LangChain, LangGraph, LangSmith*

- Mapped a **modular LangGraph** with three specialized nodes (write query, execute, generate answer) and implemented state management with prompt templates from **LangChain Hub** and output parsing for effective LLM interaction.
- Utilized **LangSmith** to trace and debug LLM calls, leveraging the **Qwen2.5:7b model** from Ollama, and integrated the **Chinook database** to enable accurate text-to-SQL querying within a custom-built agent.
- Enhanced agent robustness and query accuracy by replacing the **linear graph structure** with LangChain SQL tools integrated directly with LLM and the database, resulting in improved performance and reliability.

PEFT Finetuning Phi2:1.5b model on Custom Dataset using HuggingFace | *Hugging Face*

- **Fine-tuned** the **Phi2 base** language model using **QLoRA 8-bit**, reducing trainable parameters to **26M (1.69% of total)** while maintaining model effectiveness, enabling resource-efficient large language model adaptation.
- **Optimized data preprocessing** by determining an ideal max **token length of 500** for efficient **tokenization** to ensure high-quality input representation for fine-tuning on a **custom Amazon product dataset.**
- **Merged fine-tuned parameters** with the **Phi2 base model** to successfully generate tailored product names and descriptions from Amazon product categories, improving automated content generation accuracy.

Fine-Tuning BERT for Emotion Analysis of Textual Data | *Hugging Face*

- Preprocessed textual data for transformer fine-tuning by performing **tokenization** to generate **input ids, attention mask, and token type ids**, ensuring compatibility with BERT's encoder architecture.
- Fine-tuned a pre-trained BERT model for 6-class emotion classification by adding a classification head, defining label2id and id2label, and configuring **Hugging Face Trainer** with Training arguments and custom compute metrics.
- Evaluated model performance on test data, achieving **89% F1-score**, and performed inference using **Hugging Face Inference Pipeline** and the finetuned BERT model.

Multiclass News Classification using LSTM and BiLSTM | *Tensorflow, Keras, Matplotlib*

- Preprocessed news text data with thorough cleaning and **tokenization** using a custom vocabulary size; determined **optimal max token length** to pad/truncate sequences for consistent model input.
- Built and trained a deep learning model using an Embedding layer followed by LSTM layers with **dropout** regularization and a Dense **softmax output layer**; compiled with categorical crossentropy loss and **Adam optimizer**.
- Designed and evaluated a **Bidirectional LSTM** architecture, achieving a **90% validation accuracy** by leveraging context from both past and future tokens in sequences.

Education

Indian Institute of Technology, Madras **Jul 2021 – May 2023**
M.Tech in Chemical Engineering

Anna University, Chennai **Aug 2015 – Apr 2019**
B.Tech in Chemical Engineering