Sri Hari Karthick N.

sriharikarthicknarayanasamy@gmail.com | linkedin.com/in/sriharikarthick | github.com/codebykarthick

Summary

- Professional with **4+ years of experience** in Software Engineering building resilient microservice modules. Proficient in intra and inter-team collaboration for feature delivery.
- Led decomposition of monolithic service into microservice modules rewritten using **Helm, Openshift, Kubernetes, Docker with CI/CD pipelines**, through test-driven development.
- Specialising in advanced Natural Language Processing techniques such as **parameter-efficient fine-tuning (LoRA), prompt engineering, knowledge distillation** and worked on **RAG** systems using **LangChain**.

Key Skills

- **Technical Skills:** Python, PyTorch, Hugging Face, Spring Boot, React Native, FastAPI, Docker, Kubernetes, Helm, LoRA, LangChain, FAISS, Git, CI/CD.
- **MLE/MLOps:** Model fine-tuning, prompt engineering, on-device inference optimisation, RAG pipeline design, containerisation, scalable deployments, logging & monitoring.

Experience

Software Engineer, Citibank Sep 2020 – Sep 2024

- Engineered backend services for payments, user management, partner API integrations, and asynchronous notifications, ensuring reliable, time-sensitive alert delivery for users using **Spring Boot Framework**.
- Migrated application from server-based infra to cloud-native solution through **Kubernetes**, **Docker** and **OpenShift (Equivalent to Azure)**.
- Designed and implemented data residency solutions using **Oracle Database** to meet country-specific compliance requirements, ensuring a successful infrastructure migration.
- Scaled and made production resilient by decomposing monolith, adding **Couchbase** pre-caching for state coordination.
- Led production issue diagnosis by enhancing log metadata in the Kibana for faster querying and **collaborating with support teams and customers**, reducing issue turnaround time.
- Built internal tooling for handling translation and automated file processing using **Pandas** and **Python**.
- Contributed to feature screens for Mobile app in **React Native**.

AI Projects

- **LoRA-T5 Summariser – Parameter-efficient Fine-tuning [GitHub]**
  - Fine-tuned T5-small with LoRA on CNN/DailyMail for sample customisation, improving BERTScore F1 in one epoch. Used HuggingFace for training, packaged and served locally using Docker and LitServe.
- **RAG Chatbot – Answer queries from documents [GitHub]**
  - Developed a Retrieval-Augmented Generation (RAG) chatbot to answer user queries from PDFs and DOCX files using **FAISS and LangChain** in Python.
  - Implemented a **Streamlit dashboard** for interactive querying, with automatic **entity detection, query expansion**, and retrieval of top relevant documents through **OpenAI API** with chat history.
- **DeepTest – On-device Confidence Scoring for Lateral Flow Tests (MSc Dissertation) [GitHub]**
  - Designed lightweight CNN architectures for local mobile inference (React Native) to evaluate fungal keratitis lateral flow test results, built using PyTorch.
  - Compared fine-tuning, few-shot learning, knowledge distillation, ROI cropping, and class rebalancing to achieve F1 0.80 with 400ms latency. **Currently in clinical testing in India** for real-world validation.

Education

University of Edinburgh, MSc Artificial Intelligence (Distinction) Sep 2024 – Oct 2025

- Focus areas: Natural Language Processing, Computer Vision, MLOps, Reinforcement Learning.
- Dissertation: Designed and optimised lightweight CNNs for on-device confidence scoring of Fungal Keratitis lateral flow test images, evaluating various techniques.