U S Sooraj

ussooraj99@gmail.com +91 9207214691 Thrissur, Kerala, India github.com/ussooraj

PROFILE

An AI/ML Engineer and 2026 graduate specializing in Document AI and LLM fine-tuning. Proven track record of building automated dataset tools, complex OCR extraction pipelines and domain-specific AI models from the ground up.

PROFESSIONAL EXPERIENCE

AI/ML Intern, *ICFOSS* 12/2025 – 04/2026
Thiruvananthapuram Kerala

- Engineered **DhritiOCR**, a robust document-level OCR system optimized for Malayalam-English scripts, featuring advanced extraction pipelines for complex layouts and data tables from unstructured PDFs.

ML Intern, *ICFOSS* 05/2025 – 06/2025
Thiruvananthapuram Kerala

- Developed **ADAPT**, an automated audio dataset generator utilizing diarization and VAD to curate high-quality training data for Malayalam ASR/TTS projects.

Open IoT Student Ambassador, *ICFOSS* 09/2024 – 08/2025
Thiruvananthapuram Kerala

- Monitored live data from district wide IoT weather systems in Pathanamthitta, gaining hands on experience with LoRaWAN gateways.

EDUCATION

Bachelor of Technology Computer Science (Cyber Security), 10/2022 – Present
College of Engineering Kallooppara Thiruvalla, Kerala

SKILLS

Programming & Frameworks — Python, PaddlePaddle, PyTorch, Transformers, FastAPI, OpenCV

AI & Inference — CUDA, ROCm, ONNX, vLLM, llama.cpp

Tools & Infrastructure — Git, Linux, Docker

PROJECTS

DhritiOCR, *SOTA Multilingual OCR Engine*
An end-to-end document parsing system engineered to extract complex layouts and data tables from unstructured PDFs, achieving State-of-the-Art performance in Malayalam recognition for its weight class.

M-Synth, *Multilingual Synthetic OCR Data Generator*
A high quality OCR dataset generation toolkit designed to accelerate vision-model training. It automatically curates character-level balanced datasets with highly adjustable visual augmentations, utilizing a custom rendering pipeline that solves complex font-breakage issues across multi-language scripts.

ADAPT , *Audio Data Annotation and Preprocessing Tool*
Engineered an end-to-end dataset curation tool featuring native CUDA and ROCm compatibility for cross-platform GPU acceleration. The pipeline automates complex ML workflows including vocal isolation, speaker diarization, VAD based slicing and phonetic transcription.

Clara , *Cybersecurity based LLM for Anomaly and Risk Assessor*
A specialized Large Language Model fine-tuned on the Llama 3.1 architecture using PEFT, specifically post-trained to analyze codebases and detect complex security vulnerabilities using custom PrimeVul Dataset.