Thrangborwell Sohlang

Shillong, Meghalaya | **thrang-portfolio.vercel.app/** | thrangsohlang@gmail.com | +91 6009453430

Summary

AI/ML Engineer designing and deploying production Large Language Model (LLM) and computer-vision systems. Outcomes include over 60% faster analytics cycles, 22% word-error rate on a low-resource ASR, 92% document-extraction accuracy on a 20-document pilot, 500 ms median RAG query latency over a 1,000-document corpus, and 20 ms time-to-answer for tariff look-ups. Proficient with FastAPI, Docker, vector search, Retrieval-Augmented Generation (RAG), and agentic workflows (LangChain/LangGraph/CrewAI).

Skills

- **Programming:** Python, SQL
- **ML/AI:** PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers, YOLO (v3–v8), DeepSORT, VGG
- **LLMs and Agents:** OpenAI API, RAG, LangChain, LangGraph, CrewAI, AutoGen, prompt engineering
- **APIs and Tooling:** FastAPI, Pydantic, OpenAPI/Swagger, Postman
- **Data and Infra:** BigQuery, Scrapy, Selenium, Git, Docker, REST APIs
- **Vector Databases:** Pinecone, ChromaDB, FAISS, Qdrant
- **Cloud:** AWS (S3, CloudWatch)
- **Deployment and UI:** FastAPI, React, Streamlit, VLLM, Ollama
- **Optimisation:** ONNX, TensorRT, OpenVINO
- **Testing:** PyTest, unit testing

Experience

AI Engineer - Sustainability Economics Bangalore
Mar 2025 – May 2025

- Appraised 3 document-AI platforms (Google Document AI, AWS Data Automation, Azure Document Intelligence) for structured extraction from PDFs and spreadsheets.
- Established 1 AI-driven ETL prototype on AWS with CloudWatch monitoring and lineage logging, covering ingest, extract and load stages end-to-end.
- Realised 92% field-level extraction accuracy on a labelled pilot of 20 documents; compiled an error taxonomy and remediation playbooks.
- Co-ordinated multi-agent workflows using 3 frameworks (LangGraph, CrewAI, AutoGen) to automate multi-step data operations.
- **Tech:** AWS (S3, CloudWatch), Google Document AI, Azure Document Intelligence, LangGraph, CrewAI, AutoGen, Python, Docker

AI/ML Developer - Sarjen Systems (Client of DataLabs) Ahmedabad
Aug 2024 – Feb 2025

- Upgraded 1 enterprise RAG system for document search across PDFs and flowcharts: delivered 500 ms median query latency and improved search accuracy by 60%.
- Indexed 1,000 documents with custom parsers to normalise graph and flow artefacts for structured retrieval and ranking.
- Engineered 1 real-time golf-ball tracking proof-of-concept (YOLO + DeepSORT) with 60 s end-to-end latency on 30 s, 120 FPS footage.
- **Tech:** LangChain, custom parsers, YOLO, DeepSORT, Python

Data Scientist Associate - DataLabs AI Hyderabad
Jul 2024 – Feb 2025

- Constructed 1 Scrapy + Selenium pipeline that processed 20 GB of marketing data with resilient crawl and retry logic.
- Cut manual cleaning time by 45% through normalisation and outlier filtering across the processed dataset.
- Introduced 1 prototype GPT assistant for structured sales-data querying to accelerate ad-hoc analysis.
- **Tech:** Scrapy, Selenium, Python, SQL

Data Science Intern - Response Informatics Ltd. Hyderabad
Oct 2023 – Jun 2024

- Co-developed (2-person team) 1 NLP dashboard that generated SQL from natural language and rendered live visualisations in Streamlit.
- **Tech:** Streamlit, SQL, Python, NLP

Projects (Top 3)

HTS Agent (Harmonised Tariff Schedule) - Hybrid Search & Duty Calculator
AI Agent 2025

- Engineered hybrid retrieval (BM25 + embeddings) over HS codes, chapter and section notes, and rulings with transparent citations; achieved 20 ms average time-to-answer.
- Devised a rule-aware duty calculator applying chapter and section notes plus tariff formulae; exposed FastAPI endpoints with unit tests and guardrails.
- Reduced policy look-up errors by 80% and consolidated multi-step look-ups into 1 query-and-compute flow.
- **Tech:** LangChain, ChromaDB, FastAPI, Python, Docker, embeddings, BM25, OpenAPI/Swagger

Research Assistant Agent - Ingestion & RAG for Scientific Content
AI Agent 2025

- Developed a RAG pipeline with deterministic chunking and vector search; FastAPI backend (OpenAPI/Swagger) and 1 React UI for interactive queries.
- Integrated JWT-based authentication and role-based access control with 3 scoped permissions (ingest, search, retrieve).
- Attained ingest throughput of 1 document per minute, 100 ms response latency, and 85% citation precision on manual spot-checks with source-grounded answers.
- **Tech:** Python, FastAPI, React, Pydantic, LangChain, ChromaDB/Pinecone, JWT, RBAC, Postman, OpenAPI/Swagger

GPT-Powered Data Visualisation Tool Internal Tool
2024

- Enabled natural-language SQL and chart generation (GPT-4 + LangChain) across a schema with over 20 tables for business stakeholders.
- Enhanced quality controls (evaluator checks and retry/error handling) and shortened the query-to-report cycle by over 60%.
- **Tech:** OpenAI API, LangChain, SQL, Streamlit

Education

M.Sc., Physics - Indian Institute of Technology Hyderabad (IIT Hyderabad)
Aug 2020 – Aug 2022

- Compared 2 afterglow models using AIC/BIC; broken power-law outperformed single power-law on the evaluated dataset.
- Coursework included 2 advanced modules: Computational Physics; Data Science for Astronomy.

Awards and Languages

- **Scholarship:** INSPIRE Scholarship for Higher Education (2017–2022).
- **Languages:** Khasi (native), English (fluent).