ABHINAV KM

AI/ML Engineer

mailabhikm@gmail.com

+91 8075708912

Blog
Github
StackOverFlow

SKILLS

Python	Azure	NLP	Machine Learning	Deep Learning
Generative AI	Data Analysis	Visualization	Linux	Documentation

TECHNOLOGIES

Langchain	LlamaIndex	Tensorflow	spaCy	Transformers
llama.cpp	SkLearn	Seaborn	Docker	Svelte

WORK EXPERIENCE

Software Engineer L3 (R&D) May 2022 - Present

Zerone Consulting Ernakulam

- Developed AI powered fault-tolerant multi-tenant services handling large volumes of data (Python, LLM, FastAPI, Middleware)
- Familiar with the Azure eco-system and the deployment of services such as App Services, Azure Functions and Container Apps (Python, Azure, Durable Functions, Docker)
- Designed RAG systems that leveraged RAG techniques and Vector Databases to generate context-aware responses with Text-To-Speech and Speech-To-Text support (Langchain, Whisper, Milvus)
- Trained Machine Learning and Deep Learning models for Classification, Image Recognition and Forecasting (Tensorflow, Pandas, Scikit-Learn)
- Designed NLP pipelines with spaCy for tasks such as NER, Annotation and Summarization (spaCy, Gensim, NLTK)
- Deployment, fine-tuning and inference of Open-Source LLMs on-premise (Ollama, Llama.Cpp, Unsloth, HuggingFace, Wandb)
- Coordinated the development, versioning and deployment of micro-services across different project teams
- Created and maintained technical documentation of projects and internal tools

EDUCATION

B.Tech Computer Science - KTU (2017-2021)

PROJECTS

LLM Graph Builder

- Ingests documents to build a knowledge graph that displays correlation using LLMs and Neo4j - repo

Imgen

- REST API service designed to perform image processing on minimal configuration (fork) - repo

Digital Brain

- Collects the user's digital footprint from various online sources to act as a digital brain - repo

ACHIEVEMENTS

Presented and published IEEE paper in the domain of Deep Learning

- K. Abhinav, R. Aneesh, P. James and A. Varghese, "Attendance Marking System using Periocular Recognition with Temperature Monitoring (ASPR)," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), Noida, India, 2021

Open-Source Contributor

- Langchain - Added support for Model2Vec (transformer embeddings), documentation for using Llama.cpp (LLM inference on GPU)

Active user on StackOverFlow

CERTIFICATIONS

Google Data Analytics Professional Certificate

Serverless Computing using Cloud Functions - IBM

Certified Specialist in Full Stack Development - ICT Academy