SUJIT SHELAR

Pune, India

sujit130193@gmail.com | sujitshelar | sujitshelar | sujitshelar_ | +91-9561100167

Professional Experience

AI/ML Computational Science Specialist (Team Lead)
Nov 2023 – Present

Accenture Solutions Private Limited — *Pune, India*

1. Content Planning Solution on Veeva Vault

- Developed the interactive chatbot capable of updating the content dynamically, retrieve the content
- Automation in the form of agent and necessary tools facilitating multiturn conversation

  Roles and Responsibility: Backend code development and orchestration, Mentoring junior team members, Tech Stack: AWS, Azure OpenAI, Langgraph, Langchain, Veeva Vault

2. Database Querying and Visualizer Chatbot

- Developed an interactive scalable chatbot capable of converting natural language queries into SQL, retrieving data from a Databricks database, and presenting results through dynamic visualizations scaled for several business units.
- Developed the agents to handle complex queries, enhancing the chatbot's capabilities.
- Handled the agent memory and TPM (Tokens per minute) in the session using rolling summarization.
- Handled the RPM (Rate per minute) for the concurrent users using InMemoryRateLimiter.
- Added user preferences memory to the chatbot.

  Roles and Responsibility: Backend code development and orchestration, Mentoring junior team members , Tech Stack: AWS, Azure OpenAI, Databricks, Langgraph, Langchain

3. Knowledge Assistant for Chiller Technicians

- Developed a knowledge bot for chiller technicians using Azure OpenAI, Azure AI Search, and Semantic Kernel.
- Created a search index for number of documents using AI Search. Implemented role-based routing with Semantic Kernel Plugins to predict query intent and direct queries to appropriate knowledge base.
- Ensured accurate responses while maintaining the GPT-3.5 4k token limit for both prompts and responses along with retrieved context using semantic search of Azure AI Search.

  Roles and Responsibility: Backend code development and orchestration, Mentoring junior team members , Tech Stack: Azure, Azure OpenAI, RAG, Semantic Kernel, Langchain

4. Multilingual multi-document type Translation Solution

- Engineered a comprehensive system to translate various document formats—including PDF, HTML, DOCX, PPTX, and IDML—into 20 target languages.
- Implemented a pre-processing step to identify and redact Personally Identifiable Information (PII) before translation, ensuring data privacy and compliance with regulatory standards.
- Leveraged Azure Translation services and Azure OpenAI (GPT-4) to process both primary content and alternative texts. Incorporated human-translated documents into Azure AI Search, providing large language models with contextual references for culturally adapted and stylistically consistent translations.
- Developed APIs for user and file management, utilizing Azure Cosmos DB for high availability and scalability. Implemented dashboards for project managers and linguists, offering real-time insights into translation workflows and facilitating efficient oversight.
- Created complex Logic App workflows for different routes for end-to-end translations.

  Roles and Responsibility: Backend code development and orchestration, Mentoring junior team members , Tech Stack: Azure Logic App, Azure OpenAI, RAG, Azure Translation, Langchain

Finetuning the LLM

- Fine-tuned Meta's LLaMA 3 (8B) model using LoRA and 4-bit quantization on A100 GPU for domain-specific medical Q&A, achieving efficient inference with reduced memory footprint. SujitShelar/llama3-medchat-8b-lora
- Fine-tuned a 4-bit QLoRA-enabled Qwen-1.5B LLM on GSM8K for advanced mathematical reasoning using a A100, SujitShelar/deepseek-gsm8k-lora
- Fine-tuned Meta's V-JEPA 2 ViT-Large video encoder on the 6 766-clip HMDB-51 action-recognition benchmark, reaching 42.9 % top-1 accuracy with a head-only training on a single A100 GPU. SujitShelar/vjepa2-vitl-fpc16-256-hmdb51

Senior Machine Learning Engineer
Aug 2017 – Nov 2023

Tata Motors Ltd — *Pune, India*

1. Conversational AI Assistant Development

- Developed a conversational AI assistant using the RASA framework, implementing NLP components and fine-tuning speech-to-text models for Indian accents to enhance in-cabin human experience.
- Presented the proof of concept to management and the connected vehicle platform team for potential implementation.

2. Sentiment Analysis on Automobile Reviews

- Collected customer review data by web scraping top automotive review sites.

- Utilized Google NLP API for entity extraction and AutoML for sentiment analysis.
- Provided actionable insights to critical component owners regarding customer feedback and areas for improvement.
- Conducted comparative analysis of features with competitor automakers to identify competitive advantages and shortcomings.

3. Range Prediction Polygon on Infotainment

- Developed a feature that displays the remaining driving range as a polygon on Google Maps, offering drivers a visual representation of travel capacity in all directions.
- Implemented the Bellman-Ford algorithm for optimal route prediction and calculated energy consumption using vehicle dynamics.
- Leveraged the Osmnx library for accessing real-world street networks, HERE Isoline Routing for real-time traffic data, and OpenWeatherMap API for ambient temperature and humidity data.

4. Driver Drowsiness and distraction prediction

- Developed a real-time computer vision system to automatically detect driver drowsiness and distraction, triggering alarms when necessary.
- Implemented eye aspect ratio (EAR) analysis over 20 consecutive frames, generating alerts if EAR fell below 0.25.
- Trained and fine-tuned a VGG16 model on the Kaggle State Farm Distracted Driver Detection dataset, achieving a LogLoss of 1.29.
- Enhanced road safety by proactively identifying and alerting against driver fatigue and inattention.

5. Battery Digital Twin Development

- Led the development of a hybrid model integrating physics-based and data-driven approaches to predict battery Remaining Useful Life (RUL), utilizing advanced architectures and deploying scalable solutions on AWS. Mentored a team in implementing anomaly detection algorithms to forecast critical battery events, enhancing system reliability and reducing downtime.

Education

Indian Institute of Technology, Madras — **Aug 2015 – May 2017**
Masters in Technology in Mechanical Engineering, CGPA: 9 / 10 — *Chennai, India*

Savitribai Phule Pune University — **Jul 2011 – Jun 2015**
Bachelors in Engineering in Mechanical Engineering, 67% — *Pune, India*

Technical Skills

- Computer Languages: Python, Matlab, PostGreSQL
- Generative AI: Azure OpenAI, Azure AI Search, AWS Bedrock, Langchain, LlamaIndex, VectorDBs, Prompt Engineering, Huggingface
- Agentic Frameworks: Semantic Kernel, LangGraph, CrewAI
- Large Language Model : Quantization, Finetuning LLM
- GraphDB: Neo4j
- Time Series Forecast: Univariate, Multivariate LSTM model variations.
- Deep Learning: Transformers, BERT.
- Cloud Technologies: Azure, AWS
- Tools : Cursor, Windsurf, VS-Code, Azure ML Notebook, AWS Sagemaker

Certifications

1. Multi Agents Systems (CrewAI)
2. Introduction to LangGraph (Langchain Studio)
3. Azure AI Fundamentals (AI 900)
4. Machine Learning and Deep Learning Specializations (Coursera, Stanford University)
5. Introduction to tensorflow for Artificial Intelligence, Machine Learning and Deep Learning (Coursera)
6. Neo4j Certified Professional
7. Neo4j Graph Data Science Specialization