Dheeraj Etlam

Senior Data Scientist

Brief Work Summary

Data Science professional with 5+ years of hands-on experience in building and deploying cutting-edge Generative AI solutions, focusing on predictive modeling, time series forecasting, and optimization. Expertise in developing scalable AI/ML systems leveraging Python, AWS, and Azure, with a strong focus on MLOps, Large Language Models (LLMs), and AI-driven applications. Proficient in designing real-time and batch inference pipelines, fine-tuning models for GenAI use cases, and applying prompt engineering techniques to enhance model performance. Proficient in prompt engineering, retrieval-augmented generation (RAG), and orchestrating GenAI workflows using Python, Langchain, and OpenAI/Azure OpenAI APIs. Skilled in integrating LLMs into enterprise-level applications and driving innovation through data-driven solutions. A proven track record of collaborating with cross-functional teams to deliver impactful business insights and automation. Passionate about advancing the field of Generative AI and transforming complex data challenges into actionable business solutions.

Education

Bachelor of Technology in Computer Science and Engineering.

Technical Competencies

Category	Skills / Tools
Languages & Frameworks	Python, SQL (MSSQL, PostgreSQL), FastAPI, LangChain, Streamlit, Gradio.
Generative AI	Prompt Engineering, Chain of Thought, Zero/Few-Shot Learning, Agentic AI, RAG.
LLM APIs	OpenAI, Azure OpenAI, Hugging Face Transformers, Ollama, Vector Stores (FAISS, Chroma).
Cloud Technologies – AWS	AWS Bedrock, AWS Textract, Amazon Sagemaker, Amazon Forecast Service, Athena, S3, Glue, Lambda, Amazon EventBridge, CloudWatch, Step Function.
Cloud Technologies – Azure	Azure AI Foundry, Azure OpenAI, Azure Functions, Azure Container Registry, Azure App Service, Azure Storage Account, Application Insights, APIM.
Data & Visualization	Pandas, NumPy, Matplotlib, etc.
Version Control	Git

Functional Skills

- Proficient in understanding business requirements and leading teams to develop applications that meet those requirements efficiently and effectively.
- Skilled in overseeing data collection from various sources, guiding the team in pre-processing, exploratory data analysis (EDA), and visualizing data to derive actionable insights.
- Experienced in developing applications that leverage deep learning, natural language processing and Generative AI techniques to solve complex problems.
- Led the development of applications using Generative AI, overseeing the deployment, hosting, and inferencing from in-house LLMs to deliver high-quality solutions.
- Strong ability to mentor and guide team members in quickly learning new technologies, fostering a culture of continuous improvement, and delivering dynamic and robust solutions.
- Adept at collaborating with cross-functional teams, ensuring alignment between technical execution and business goals throughout the project lifecycle.

Prior Project Experience

Retail Product Recommendation System: (Jan 2024 – Oct 2024)

The client, an e-commerce platform, wanted to improve how users discover products through conversational queries. The solution provides an advanced query system that enables users to ask natural language questions and receive tailored product suggestions for easy purchasing.

- Designed and implemented a system to improve the user shopping experience by suggesting products based on user queries.
- Utilized AWS SageMaker for model deployment and ML-Ops practices for end-to-end automation, the system was designed to handle real-time data ingestion, model inference, and operational deployment at scale.
- Implemented ML-Ops practices to streamline the model training, deployment, and monitoring process.
- Automated model retraining pipelines to keep the recommendation system up to date with new data.
- Implemented model versioning and rollback mechanisms to maintain stability in production environments.

Chemical Domain Query System (Mar 2023 – Dec 2023)

The Client focused on developing an advanced query system tailored to the chemical domain, designed to generate accurate responses based on complex, domain-specific information.

- Designed and implemented a RAG system to efficiently answer user queries by retrieving relevant information from a variety of data sources.
- Utilized the LLAMA model in combination with the LangChain framework to fine-tune the model for improved query understanding and accurate responses.

- Integrated multiple PDF extractors to extract and process information from documents, enabling the system to query large volumes of unstructured text data from PDFs, ensuring comprehensive and precise answers.
- Developed and optimized various tasks, such as document parsing, text extraction, information retrieval, and data aggregation, to enhance the system's ability to generate contextually relevant responses based on user queries.

ML-Ops Engineer: Vehicle Damage Detection (Mar 2022 – Feb 2023)

The client, an automotive inspection company, needed a more efficient way to identify vehicle damage according to M22 standards. The problem was that manual assessments were time-consuming and prone to inconsistency. The solution automates and streamlines the damage identification process, ensuring accurate, standardized assessments of vehicle damage based on M22 criteria.

- Guided the implementation of cutting-edge object detection models, specifically YoloV8 and Faster RCNN, to detect various types of vehicle damage, such as dents, scratches, and other surface imperfections.
- Ensured that these models were accurately trained to meet the required M22 standards, optimizing them for efficiency.
- Co-ordinated the creation of a fully integrated ML-Ops pipeline using AWS SageMaker, which facilitated the automation of model deployment, real-time monitoring, and continuous optimization.

Face and License Plate Anonymization (Feb 2021 – Feb 2022)

The client needed a solution to protect sensitive information. The problem was that face and license plate data in images posed privacy risks. The solution anonymizes face and license plates, ensuring privacy compliance while preserving the integrity and usefulness of the data.

- The goal was to create realistic, anonymized versions of sensitive information.
- Co-ordinated the fine-tuning of the GAN model to ensure high-quality results, enabling the model to generate plausible substitutes for both faces and license plates that were unrecognizable yet retained critical features like position, size, and orientation.
- The result was a robust and scalable solution that met both technical requirements and privacy standards, significantly reducing the risk of personal data exposure in sensitive image datasets.

Forklift Free Space Detection (Oct 2019 – Jan 2020)

The client, an autonomous vehicle company, needed an enhanced system for better vehicle perception. The problem was that detecting available space in front of vehicles for safe navigation was challenging. The solution

was a robust image segmentation tool that accurately identified free space, improving vehicle perception and ensuring safer navigation

- Led the development of an image segmentation model for detecting free space in front of vehicles.
- The core of the solution was built using the U-Net architecture, known for its efficiency in segmentation tasks, which was customized to detect relevant environmental features.
- TensorFlow was utilized for model development, with training and fine-tuning performed on AWS to handle the computational demands and ensure scalability
- Additionally, incorporated OpenCV for post-processing tasks, refining the segmentation output for real-time application.
- This solution significantly enhanced the vehicle's perception system, enabling safer and more efficient navigation.

Data Mining and Sentiment Analysis (NLP) (Sep 2018 – Sep 2019)

The primary goal of the project was to collect data from online reviews through web scraping, with the aim of analysing and visualizing customer sentiment to gain valuable insights into user feedback and overall satisfaction.

- NLP techniques were utilized to analyse customer sentiments, identify key issues, and uncover emerging trends from the data, providing actionable insights for product improvement.
- The BERT algorithm was employed to develop a robust model for accurate sentiment classification and text analysis.
- Data visualization was seamlessly integrated using Kibana dashboards, enabling real-time insights and interactive exploration of customer feedback and trends.
- The findings were shared with various R&D teams, facilitating data-driven decision-making to enhance the product based on customer satisfaction and pain points.
- The solution was developed using a combination of Python, HTML, CSS, SQL, and Docker, ensuring scalability, performance, and smooth deployment across different environments.