Srutej Kodiganti

Dublin, CA · srutej.kodiganti@gmail.com · +1 (770) 282-9950 · https://www.linkedin.com/in/srutejkodiganti/

Work Experience

Talent Screen — Dublin, CA
Machine Learning Engineer — Sept 2022 - present

- Developed chatbots which use Generative AI (GenAI) and Large Language Models (LLMs) to deliver personalized and context-aware responses, enhancing user engagement and satisfaction.
- Applied NLP techniques using transformer-based models (e.g., BERT, GPT) for natural language understanding and generation, enhancing the chatbot's ability to process and generate contextually accurate responses based on user input and domain-specific language.
- Developed and implemented custom web scraping pipelines using Python libraries (e.g., Scrapy, BeautifulSoup) to gather domain-specific data from websites and APIs, ensuring the chatbot model is continuously fed with high-quality, relevant data to enhance its understanding and response capabilities.
- Implemented a Retrieval-Augmented Generation (RAG) framework to improve the chatbot's response generation by integrating external knowledge sources, thereby enhancing the richness of the interaction.
- Evaluated the RAG framework by assessing retrieval accuracy and response quality, leveraging precision-recall metrics for retrieval and BLEU/ROUGE scores to measure the relevance quality of generated responses.
- Utilized Vector Databases to store and retrieve embeddings efficiently, facilitating quick and accurate similarity searches for improved user query handling.
- Utilized LangChain for advanced conversational capabilities and workflow automation, enhancing the chatbot's ability to manage complex dialogues and tasks.
- Fine-tuned LLMs using domain-specific datasets to optimize the chatbot's performance, ensuring high accuracy and relevance in conversations.
- Integrated Hugging Face Transformers to leverage pre-trained LLMs, streamlining fine-tuning and deployment processes for domain-specific applications, enhancing response quality and model adaptability.
- Implemented Agentic AI to enable autonomous decision-making and dynamic task management, enhancing the chatbot's adaptability in real-time interaction.
- Deployed the chatbot on AWS using Bedrock for scalable model training and inference, ensuring reliable and efficient performance under varying load conditions.
- Developed and deployed scalable RESTful APIs with FastAPI, enabling real-time communication between the chatbot and external systems for data retrieval and model inference.
- Designed and managed scalable infrastructure solutions using Amazon EKS for Kubernetes-based container orchestration, ensuring high availability and efficient resource utilization and Git for CI/CD pipeline management.
- Implemented LLMOps pipelines using Docker and custom orchestration scripts, streamlining versioning, deployment, and monitoring workflows specifically tailored for large language models (LLMs), ensuring efficient and reliable production performance.
- Coordinated cross-functional efforts to integrate domain-specific knowledge into Generative AI models, ensuring the system met both technical and business requirements.

Innovapath Inc — Dublin, CA
Machine Learning Engineer — Oct 2020 - May 2022

- Built and deployed a neural network-based fraud detection system using TensorFlow and PyTorch, integrated with Apache Kafka for real-time streaming and PostgreSQL for structured data storage, reducing fraudulent transactions.
- Designed and implemented a Machine learning-based fraud detection model for analyzing tabular transaction data, addressing class imbalances and achieving high precision in identifying fraudulent activities while seamlessly integrating with the SageMaker deployment pipeline.
- Implemented advanced data preprocessing techniques, including outlier detection, data normalization, and feature scaling, to improve the quality and consistency of input data, resulting in better model accuracy and reliability.

- Optimized and deployed fraud detection models on Amazon SageMaker, leveraging Docker for containerized deployment and AWS CloudWatch for performance monitoring, ensuring scalable and reliable operations.
- Developed machine learning pipelines to automate model training, testing, and deployment, making the process more efficient. Tuned model parameters and enhanced feature engineering to improve the performance and scalability of the fraud detection system.
- Worked with cross-functional teams to integrate machine learning models into real-time data streams, enabling faster and more accurate fraud detection.
- Conducted regular model evaluations and retraining to adapt to evolving data patterns, maintaining accuracy and relevance in predictions.
- Created detailed documentation and guidelines for maintaining and updating machine learning models, ensuring smooth transitions between development and production teams.

Skills

Languages:	Python
Frameworks:	TensorFlow, PyTorch, Keras, MLflow
Libraries:	Langchain, Llamaindex, NLTK, Transformers, Scikit-learn, NumPy, Pandas, HuggingFace
LLMs:	LLaMA 2, Anthropic Claude, BERT, LLaMA 3.1, BART
Database:	SQL, PostgreSQL, MongoDB
AWS:	EKS, Bedrock, RDS, EC2, CloudWatch, Lambda, S3 bucket

Education

California State University, Fullerton — Fullerton, CA
Master of Science Information Systems Business Analytics *GPA: 3.87* - May 2024

University College of Engineering, Osmania University — India
Bachelor of Engineering Computer Science and Engineering *GPA: 3.80* - May 2022