Samarasimha Reddy Punnam

psamarasimha.reddy06@gmail.com | linkedin.com/in/spunnam | GitHub | (618) 528 9477

SUMMARY:

ML & GenAI Engineer with extensive experience in developing machine learning models, generative AI applications, and data-driven solutions. Specializing in NLP, LLMs (GPT, BERT), RAG, and fraud detection. Proven expertise in building AI-powered chatbots, optimizing data pipelines, and leveraging cloud technologies (AWS, Azure) for scalable deployments. Experienced in fine-tuning deep learning models with TensorFlow and PyTorch for industry-specific use cases.

TECHNICAL SKILLS:

Deep Learning & Generative AI: LLMs (GPT, BERT, LLaMA), RAG, TensorFlow, PyTorch, Hugging Face, Transformers, CNN, RNN, LSTM, ResNet, BedRock, SageMaker, Langchain

Data Visualization: Tableau (Desktop/Server), Power BI, PyLDAvis, Matplotlib, Seaborn.

Programming: Python, C, C++, Java, Spring Boot, JavaScript, JSON, React-Native, NodeJS, REST API, PHP, Git, Postman

Data Engineering & Databases: MySQL, Oracle, PostgreSQL, SQL Server, MongoDB, DynamoDB, Aurora, Vector databases

Cloud & Distributed Systems: AWS (S3, EC2, SageMaker), Azure (DevOps, Cognitive Services), Hadoop, Zookeeper, Spark, Kafka

PROFESSIONAL EXPERIENCE

AI & Machine Learning Application Developer, Xebia IT Architects, Gurgaon, India **June 2021 – Dec 2022**

- Developed and deployed an AI chatbot using LLMs and Retrieval-Augmented Generation (RAG), automating customer interactions for 8M+ users, increasing satisfaction by 30%. Optimized inference efficiency with quantization techniques, reducing API costs by 25%. Used React Native for a seamless front-end experience.
- Optimized LLM responses using advanced prompt engineering and few-shot learning, integrating RAG with Elasticsearch to improve factual accuracy and reduce hallucinations by over 30%.
- Developed an ML-based fraud detection system achieving 88% accuracy, leveraging machine learning models for real-time anomaly detection and distributed computing frameworks for scalable data processing, improving risk mitigation by 24%.
- Collaborated within an Agile team of 50+ members, optimizing CI/CD pipelines using DevOps practices, containerization, and orchestration to accelerate deployments, reducing model rollout times by 45%.

Graduate Assistant/Teaching Assistant, University of Houston **August 2024– December 2024**

- Assisted in teaching mathematics and statistics, providing academic support to enhance student understanding and performance.

KEY PROJECTS

LangChain Documentation Helper chat Bot **Sept 2024 - Dec 2024**

- Developed an AI-powered chatbot using **LangChain**, **LangSmith**, and **Streamlit** for efficient contextual question-answering.
- Built an ingestion pipeline with **LangChain loaders**, text splitters, and LangChain Expression Language (**LCEL**) to process documentation, generate embeddings with **OpenAI models**, and store them in a **PineCone** VectorStore**.**
- Designed **RAG-based** workflows using LCEL for document retrieval, real-time response generation, and session traceability.

Profile Probe: Personalized Insights Generator with LangChain and Generative AI **July 2024 - Aug 2024**

- Developed a LangChain-based app to fetch LinkedIn data via the Proxycurl API, leveraging Prompt Engineering techniques like Chain of Thought, ReAct, and Few Shot prompting.
- Integrated LangChain Chains, Agents, and the Tavily API for username search, enhancing networking and talent discovery through generative AI.**.**

Sentiment Analysis on Twitter Data **Jan 2025 - Feb 2025**

- Developed a Bidirectional LSTM model with Attention Mechanism using GloVe Embeddings for classifying tweets into positive, neutral, and negative sentiments.
- Achieved 69% validation accuracy, outperforming baseline models with enhanced generalization and minimal overfitting.
- Implemented advanced NLP techniques with L2 regularization, dropout, and early stopping for optimized model performance.

Optimizing News Media Analysis through Advanced Topic Modeling Techniques **Jan 2024 – May 2024**

- Processed 1,648 newspaper articles with data cleaning, **tokenization**, and feature extraction, implementing **LDA** via **Gensim** to identify 5 key themes with a **coherence** score of 0.46 and **perplexity** of -8.37.

EDUCATION:

University of Houston, Houston, TX, Master of Science in Engineering Data Science, **GPA 3.73**
Key courses: Machine learning, **Deep learning,** Applied statistics , Text mining, **Natural Language processing**, Large language models, Probability and statistics,Cloud computing, Information visualization, Tableau, Database management tools.

Indian institute of Technology(BHU), Varanasi, India, Bachelor of Technology

AWARDS AND CERTIFICATIONS:

- **AWS certified: Solutions Architect Associate (SAA-C03)**
- **GEM Award** for October 2022 in Xebia.