Shreyas Puttaraju

Chicago
(224)-389-0281
shreyas2797@gmail.com
Portfolio | LinkedIn

AI Engineer | Data Scientist

SUMMARY

AI Engineer with a nearly 3-year track record of designing and applying sophisticated AI/Data Science methodologies to deliver impactful solutions. Demonstrated proficiency in conceptualizing and deploying scalable AI solutions, including multi-agent frameworks and RAG integrations. Proven track record in developing innovative tools like candidate scoring systems and advanced document management using AI. Skilled in navigating the evolving landscape of artificial intelligence, I combine technical expertise with forward-thinking strategies to refine decision-making protocols and streamline operational procedures. My approach is marked by a unique fusion of technical excellence and innovative problem-solving.

SKILLS

Generative AI:	AI Agents, Retrieval Augmented Generation(RAG), Conversational AI, Prompt Engineering, Langchain, Large Language Models, HuggingFace, VAEs, GANs
Programming Languages:	Python, R
Data Processing & ETL:	Data Cleaning, Data Wrangling, Data Preprocessing, ETL(Extract Load Transform)
Data Analysis and Visualization Tools:	Statistical analysis, Hypothesis Testing, Scikit-Learn, XGBoost, TensorFlow, PyTorch Matplotlib, Seaborn, Tableau, Power BI, A/B Testing
Machine Learning:	Regression Analysis, Clustering, Decision Trees, Random Forest, Supervised and Unsupervised Learning, Ensemble Methods, Neural Networks, CNN, RNN, LSTM
Cloud Platforms and CI/CD:	AWS, Microsoft Azure, Git, GitHub Actions, Confluence, Jira
Other Relevant Skills:	Business Requirements Gathering, Team Collaboration, Problem-solving, Critical thinking, Leadership, Technical Writing, Strong Communication

PROFESSIONAL EXPERIENCE

AI ENGINEER (Contract) — Austin, TX

PeritusHub — Oct 2024 - Present

- Conceptualized and directed the development of a scalable **multi-agent architecture**, pioneering an **agentic design** approach utilizing **langchain** to solve complex, multi-dimensional tasks.
- Integrated a supervisory mechanism that prompts Open AI and Google LLMs to adopt various personas, ensuring the optimal agent is engaged for each task segment.
- Engineered an advanced Document Management System incorporating **RAG** and **OpenAI LLMs** for enhanced information retrieval and processing.
- Pioneered a Candidate Scoring System leveraging AWS Lambda for serverless computing and LLMs.
- Demonstrated expertise in integrating **NVIDIA's AI Enterprise** and **AWS** platforms to develop conversational AI systems.

DATA SCIENTIST INTERN — San Antonio, TX

H-E-B Groceries — June 2023 - Sep 2023

- Spearheaded the creation of a comprehensive system for extracting product entities and attributes from images, utilizing **Azure Open AI LLMs** combined with advanced computer vision techniques for precision and accuracy.
- Achieved heightened attribute detection and extraction capabilities by integrating **Paddle OCR** and **GPT-3.5-Turbo** within the **LangChain** framework, resulting in significantly enhanced performance.
- This innovation boosted product attribute management efficiency and customer engagement by up to 8%.
- Worked with MLOps and Software Engineering teams to seamlessly integrate the new extraction system into the existing infrastructure, exemplifying strong teamwork and technical coordination skills.

DATA SCIENTIST — India

Intex Technologies — July 2020 – July 2021

- Utilized **SQL** for precise data collection and preprocessing, querying over 4TB of historical data, resulting in improved data integrity.
- Employed Python and the caret package equivalent to deploy diverse machine learning algorithms, achieving 90% model accuracy based on rigorous evaluation.

- Conducted cost-benefit analyses to identify optimal maintenance thresholds, resulting in a 15% reduction in costs and maintaining a 99% equipment uptime rate.
- Collaborated cross-functionally to integrate predictive maintenance models into existing systems, leveraging **Flask** for API development and **Apache Kafka** for real-time data streaming.
- Established a robust real-time monitoring framework using powerful visualization tools like Tableau, resulting in a 20% improvement in predictive accuracy and continuous model refinement.

SOFTWARE DEVELOPER

India
Mphasis
Jan 2020 – June 2020

- Spearheaded the development of an **Intelligent Document Processing (IDP)** system using machine learning and **NLP** techniques.
- Implemented custom deep learning models for document classification, entity recognition, and sentiment analysis, improving accuracy by 23%.
- Integrated third-party APIs for OCR and language translation, enhancing language support and document understanding capabilities.
- Developed a user-friendly web interface using Flask and React.js for real-time interaction with the IDP system.
- Conducted performance optimizations resulting in a 50% reduction in processing time and resource utilization.

EDUCATION

Master’s In Data Science 2023
DePaul University, Chicago, IL

Bachelor’s in Computer Science and Engineering 2020
P.E.S College Of Engineering, Mandya, India

PROJECTS

LLM OPS - LARGE LANGUAGE MODELS IN PRODUCTION

- Spearheaded a pioneering research project on LLMOps, establishing innovative protocols for the scalable and efficient deployment of Large Language Models.
- Authored a comprehensive set of best practices for LLM operational management, enhancing performance and reducing computational costs in AI-driven systems.

RETRIEVAL AUGMENTED GENERATION (RAG) IN LARGE LANGUAGE MODELS

- Evaluated the impact and challenges of RAG, highlighting its potential in advancing AI applications across various sectors.
- Explored and analyzed Retrieval-Augmented Generation in large language models, enhancing AI application understanding.

INTERACTIVE PORTFOLIO WEBSITE USING OPENAI LLM

- Employed OpenAI's GPT-4 and prompt engineering, eliminating the need for manual coding in the creation of my portfolio
- Constructed an interactive three-page portfolio website, demonstrating the transformative application potential of LLMs

STYLE TRANSFER USING GENERATIVE ADVERSARIAL NETWORKS

- Showcased the use of GAN's architecture in style transfer, highlighting the generalizability and adaptability of CycleGAN
- Converted real-world photos to artistic paintings using GAN with TensorFlow, highlighting advanced generative techniques

VISUALIZING HR DATA

- Collaborated with a team to conduct exploratory data analysis on the HR dataset, creating insightful visualizations
- Explored workplace demographics using Tableau and R, analyzing the interplay between employees' age, gender, and department

TRAFFIC VOLUME PREDICTION

- Analyzed Minnesota's traffic volume using deep learning algorithms like CNN, RNN, and LSTM, achieving detailed traffic insights
- Employed Keras to evaluate deep neural networks, delivering highly accurate time series predictions