Shreyas Puttaraju

Chicago,IL | 224-389-0281 | shreyas2797@gmail.com | Linkedin | Portfolio

EXECUTIVE SUMMARY

- Engineered advanced generative AI applications and proofs-of-concept (POCs), specifically focusing on AI agents and RAG Pipelines. Leveraged various frameworks and libraries to create scalable, Cloud-Agnostic AI solutions.
- Expert at writing high-performing prompts for AI systems including image generation models, ensuring clarity, consistency, and optimized task performance, leading to enhanced LLM response and accuracy.
- Strong foundation in data science and analytics, demonstrated by comprehensive experience in ETL processes, data analysis, visualization, and predictive modeling.

WORK EXPERIENCE

Jewelers Mutual Group **Neenah, WI(Remote)**
AI Applications Engineer *July 2024 - Present*

- Spearheaded the rapid development of four generative AI applications, leveraging LLMs from Azure OpenAI to enhance automation and decision-making workflows.
- Designed and implemented pipelines to convert large volumes of unstructured data from multiple types of documents into structured formats, enabling downstream analytics and reporting of financial data at scale.
- Developed an advanced agentic RAG pipeline built entirely on Azure services, utilizing Azure AI Search for highly efficient indexing and retrieval.
- Integrated a custom document enrichment step to generate and embed metadata, significantly enhancing the context and precision of retrieval.
- Implemented agentic RAG, equipping it with custom-built tools specifically designed to optimize document retrieval logic, resulting in measurable performance improvements over conventional RAG models.
- Engineered both internal and external-facing generative AI tools to optimize business operations, increasing productivity and significantly reducing manual effort across multiple teams.
- Developed robust evaluation frameworks to assess the effectiveness and reliability of generative AI applications, implementing benchmarking strategies to ensure consistent performance.
- Applied advanced prompt engineering techniques to fine-tune LLM outputs, optimizing accuracy, coherence, and consistency in financial document processing tasks such as extraction, summarization, and analysis.
- Developed and deployed several simple, ad-hoc generative AI tools tailored to specific team needs, significantly increasing productivity and streamlining straightforward operational tasks.

PeritusHub (Stealth Start-Up) **Austin, TX(Remote)**
AI Engineer *Jan 2024 - July 2024*

- Designed and implemented a scalable multi-agent architecture using Langchain and AWS Bedrock, enabling coordination of specialized AI agents to handle complex conversational tasks across workflows.
- Built a supervisory control system to dynamically select and route tasks to the most appropriate AI agent, improving efficiency and contextual accuracy of multi-step operations.
- Developed modular system prompts and interaction protocols for over 7 autonomous agents, ensuring consistent decision-making, task execution, and behavioral alignment across personas.
- Engineered six extensible Python-based tools to enhance agent functionality, each exposing APIs that integrated seamlessly with the multi-agent platform.
- Architected a cloud-native ETL framework on AWS, facilitating real-time data ingestion, transformation, and persistence to support scalable LLM-based pipelines.
- Implemented a CI/CD pipeline using GitHub Actions, automating build, testing, and deployment workflows for rapid iteration and reliable system delivery.
- Built a serverless Candidate Scoring API using AWS Lambda, processing resume data via LLMs to generate role-fit assessments and enable automated talent screening.

H-E-B Groceries **San Antonio, TX**

Data Scientist Intern *June 2023 - Sep 2023*

- Engineered a comprehensive system for extracting product entities and attributes from 80,000 product images.
- Integrated Azure OpenAI LLM with PaddleOCR and GPT-3.5-Turbo within the LangChain framework, building a hybrid pipeline that enhanced data extraction precision and robustness.
- Designed and deployed RESTful APIs to expose the extraction functionality, enabling smooth integration with internal tools and downstream data workflows.
- Authored optimized and reusable prompts to improve model understanding of complex product metadata, contributing to a measurable increase in data quality and downstream utility.
- Built an interactive front-end using Flask, JavaScript, and HTML, and connected it with backend services through APIs, ensuring seamless user interaction and real-time data feedback.
- Collaborated closely with MLOps and software engineering teams to implement the system into production infrastructure, ensuring performance, scalability, and operational reliability.

VOLUNTEERING EXPERIENCE

Change The Present Organisation **New York, NY(Remote)**

Volunteer AI Engineer Intern *Sep 2023 - Mar 2024*

- Automated social media content creation for over 1400 gifts using advanced Prompt Engineering with OpenAI's LLMs and Python scripts, enhancing the efficiency and appeal of marketing campaigns.
- Built content creation pipeline for LinkedIn, Pinterest, and Facebook to streamline multi-platform publishing.

DePaul University **Chicago, IL**

Volunteer Research Assistant *Oct 2022 - Mar 2023*

- Analyzed lung nodule images using the LIDC dataset and implemented Cycle-GAN for image reconstruction, significantly improving nodule classification accuracy.
- Performed a comparative analysis of various GAN models and optimized Cycle-GAN's loss function, reducing the FID score from 4100 to 800, and enhancing model performance.

EDUCATION

Master's in Data Science

DePaul University **Chicago, IL**

Bachelor's in Computer Science and Engineering

P.E.S. College Of Engineering **India**

SKILLS

Generative AI:	Large Language Models, Langchain, HuggingFace, GANs, AI Agents, Retrieval Augmented Generation (RAG), Conversational AI, Prompt Engineering
Programming Languages:	Python
Data Processing & ETL:	SQL, Data Cleaning, Data Wrangling, Data Preprocessing, ETL Pipeline Design
Statistical Techniques:	Statistical Analysis, Hypothesis Testing, A/B Testing, Predictive Modeling
Visualization Tools:	Matplotlib, Tableau, and Power BI
Machine Learning:	Supervised and Unsupervised Learning(Regression Analysis, Clustering, Decision Trees, Random Forest), Ensemble Methods, XGBoost, Neural Networks(CNN, RNN, LSTM)
Cloud Platforms and CI/CD:	AWS, Microsoft Azure, Docker, Kubernetes
Other Relevant Skills:	Project Management, Agile Methodologies, Cross-functional Collaboration, Strategic Planning, Effective Communication