Prashanth Gajula

linkedin GitHub Medium

Email: prashanthkgajula@gmail.com
Mobile: 346-481-1993

Summary

Results-driven AI Engineer with **5+ years of experience** designing and deploying production-grade **ML/AI** and data-driven solutions across cloud-native environments. Skilled in building complex products from scratch, including **LLM-powered applications**, **generative AI models**, and **data pipelines** handling large-scale datasets. Proficient in cloud-native architectures, **containerized deployments (*Docker, Kubernetes*)**, and infrastructure-as-code for scalable and reliable AI solutions. Experienced in mentoring engineers, implementing **CI/CD workflows**, **MLOps & LLMOps** practices, and delivering high-quality, test-driven software. Adept at integrating vector databases, **RAG architectures**, and applying **A/B testing**, LLM evaluation frameworks, and bias mitigation techniques to deliver business value through innovative AI-driven solutions.

Skills

- **Programming languages**: Python, R, C#, C,Java
- **Vector Databases & Embeddings**: Pinecone, ChromaDB, Qdrant, DataStax Astra DB
- **Generative AI Expertise**: LLM fine-tuning, RAG (Retrieval Augmented Generation), Conversational AI, Prompt Engineering, AI Agents, VAEs, GANs, LoRA, PEFT, Quantization, Pruning, Guardrails (Nemo, Guardrails AI), LLM Evaluation & Safety Frameworks
- **Machine Learning & AI Frameworks**: TensorFlow, Keras, PyTorch, scikit-learn, LangChain,LangGraph, LangSmith, LlamaIndex, Hugging Face Transformers, OpenAI APIs, Microsoft AutoGen, Google ADK,Crew AI, Computer Vision, Deep Learning,Speech-to-Text (Whisper), Text-to-Speech (TTS models),Kubeflow, Weights & Biases (W&B),Time Series Models
- **Cloud & Data Tools**: AWS Bedrock, Azure Data Factory, Databricks, Azure Data Lake, Azure Blob Storage, AWS SageMaker
- **Data Analytics & Visualization**: Power BI, Tableau, Cognos, Matplotlib, Seaborn
- **Databases**: SQL Server, MySQL, MongoDB, Couchbase, KDB.AI
- **Containers & DevOps:** Docker, Kubernetes, Terraform, GitHub Actions, Azure DevOps (CI/CD)

Experience

- **CyberSoft Technologies** Texas
  AI Engineer *Feb 2023 – Sep 2025*

- Utilized advanced vector databases such as **ChromaDB** and **Pinecone**, enhancing AI application data retrieval speeds by 60%, substantially improving overall performance and efficiency.
- Tailored and refined **open-source and proprietary LLM models**, achieving a 50% boost in accuracy rates and a 40% reduction in inference time, enhancing overall model performance significantly.
- Led deployment of GenAI models on **AWS Bedrock**, enabling a 70% improvement in application scalability and reliability, while leveraging cloud services to reduce operational costs by 30%.
- Developed and deployed Generative AI-powered Q&A chatbots using **LangChain** and vector databases (Pinecone, ChromaDB), reducing human support interventions by 40% and improving customer response accuracy by 60%.
- Implemented **RAG architecture** for enterprise knowledge retrieval, enabling context-aware and scalable conversational AI solutions.
- Implemented **multimodal embeddings (text and image)** to optimize semantic search and vector retrieval, improving response accuracy and reducing query latency.
- Implemented **CI/CD pipelines** and test-driven deployment of LLM applications, ensuring scalable and reliable production releases.
- Mentored junior engineers and collaborated with cross-functional teams for AI solution delivery.
- Designed and implemented **Agentic AI workflows** leveraging LangChain and Crew AI, enabling autonomous task orchestration with human-in-the-loop feedback, which improved solution adaptability and reduced manual intervention by 50%.
- Developed intelligent agents capable of dynamically selecting tools and APIs, enhancing complex decision-making processes and reducing time-to-insight for enterprise clients by 35%.
- Designed and implemented a comprehensive LLM performance evaluation framework leveraging key metrics such as **latency**, **throughput**, **token usage**, and **error rates**, enabling data-driven optimization of LLM-powered enterprise solutions.
- Designed and managed scalable **Azure Data Factory pipelines** for end-to-end **data migration** from legacy to modern systems.
- Transitioned all source database SQL scripts to **DACPAC** deployments, streamlining migration and reducing pipeline **execution time by approximately 90%** from 40–50 minutes to just 4–5 minutes.
- **Collaborated with product owners** to analyze functional differences between PrimeroEdge and SchoolCafé platforms, refining migration logic and modifying SQL scripts to ensure accurate data mapping aligned with evolving business needs.
- Authored 10+ interactive dashboards in **Power BI Report Builder** by integrating SQL stored procedures, which improved performance visibility across **30+ school districts**. Empowered stakeholders to monitor **KPIs in real time**, identify delays in import processes, and enhance decision-making for operational planning.
- Developed a **live analytics dashboard** displaying key performance indicators such as total revenue, student enrollment, and food wastage. Implemented trend analysis to highlight underperforming school districts based on predefined benchmarks. Utilized **SQL stored procedures** to retrieve backend data and integrated **C# APIs** to dynamically populate the dashboard.

- **University of Houston** Texas
  Data Scientist -Research Assistant *Dec 2021 – Dec 2022*

- Developed an automated grading system using machine learning to evaluate student responses, improving grading efficiency and consistency and reducing manual grading time by 60%.
- Designed and implemented end-to-end ML pipelines using **Python** and **Apache Airflow** for data preprocessing, feature extraction, model training, and evaluation.
- Applied NLP techniques **(TF-IDF, BERT embeddings)** to assess text-based responses, increasing grading accuracy by 25% compared to manual evaluation.
- Collaborated with faculty and graduate researchers to publish findings and present results at academic seminars and departmental research meetings.
- Utilized research tools and frameworks including **PyTorch**, **TensorFlow**, and **scikit-learn** for model development and experimentation.
- Leveraged **MLflow** for experiment tracking and **hyperparameter tuning**, reducing model iteration cycles by 25% .

- **Reliance JIO** India
  Data Scientist *Jan 2020 – Jul 2021*

- Expertly employed Python for data manipulation, **statistical analysis**, and **predictive modeling**, vital in fostering innovative AI-driven solutions, driving advancements in technology and problem-solving within the organization.
- Utilized **SQL** and **MySQL** databases proficiently to query and manage extensive datasets, guaranteeing data integrity and accessibility for analysis and modeling purposes, enhancing efficiency and accuracy in decision-making processes.
- Conducted comprehensive exploratory data analysis (EDA) on datasets including sales data (318,672 rows, 8 columns, no null values).
- Implemented various machine learning algorithms such as **Regression Analysis**, **Clustering**, **Decision Trees** in data analytics projects to extract actionable insights from complex datasets, driving data-informed decision-making processes.
- Collaborated with cross-functional teams to build predictive models for **sales forecasting** and pricing optimization, increasing revenue predictability by 25%.
- Designed and developed interactive data visualizations using tools like **Matplotlib**, **Seaborn**, **Tableau**, and **Power BI**, facilitating the communication of key findings and insights to stakeholders.
- Built a **collaborative filtering-based recommendation engine** for personalized content delivery, increasing user engagement significantly.
- Automated ML model deployment using GitHub Actions and **CI/CD pipelines**, reducing release cycle times by 40% and increasing model reliability.
- Designed end-to-end ETL pipelines using **Apache Airflow** and **SQL** for integrating structured and unstructured data, enabling faster analytics and improved decision-making.

Projects

- **MCP-Google ADK Integration**: Developed an **LLM-powered agent system** that allows AI models to **interact with real user data** by securely connecting to **Gmail, Google Drive, and Calendar**. Built **custom MCP servers** exposing these services as standardized tools, with a **modular architecture** separating authentication, tool logic, and server registration. Implemented a full **OAuth 2.0 flow** with automatic token refresh using **credentials.json** and **token.json**, ensuring secure and persistent access to Google services. Integrated the MCP server with **Google ADK**, enabling an **LLM agent (Gemini)** to dynamically **discover, select, and execute tools** using **async event-driven orchestration**, transforming static LLMs into **actionable agents** capable of performing real tasks.
  Tech Stack: **Python**, **MCP (Model Context Protocol)**, **Google ADK**, **Gmail API**, **Google Drive API**, **Google Calendar API**, **OAuth 2.0**, **Async I/O**, **Gemini (LLM)**
- **TalkToPDF**: A **voice-enabled** AI companion that lets users **talk to PDFs** to ask questions, get **summaries**, and hear responses aloud. Built with **Python** and **Streamlit**, it integrates **OpenAI Whisper (whisper-1)** for **speech-to-text**, a **Retriever Chain** powered by **FAISS** and **o3-mini** for **context-aware** lookups, and **GPT-4o-mini-tts** for **text-to-speech**—creating a seamless, **hands-free** study/research experience. Designed with a **modular architecture** for clarity and extensibility, improving **accessibility** and **productivity** for students, researchers, and professionals.
  Tech Stack: **Python**, **Streamlit**, **OpenAI Whisper (whisper-1)**, **FAISS**, **o3-mini**, **GPT-4o-mini-tts**
- **Fitness and Wellness ChatBot**: The Fitness and Wellness ChatBot is an **AI-powered conversational assistant** designed to provide **personalized fitness and wellness guidance**. Built with **Python** and **LangChain**, it integrates **LLM capabilities** to answer user queries, suggest **diet plans**, and recommend **workout routines** tailored to individual needs. The chatbot leverages **OpenAI's API** for intelligent responses and **Pinecone** for efficient vector-based search to enhance context retention. The project follows a **modular architecture** for scalability and maintainability and is **deployed on AWS** using a **Dockerized CI/CD pipeline** for automated builds and deployments. This tool empowers individuals to make informed health and lifestyle choices with ease and reliability.
  Tech Stack: **Python, LangChain, OpenAI API, Pinecone, AWS, Docker, GitHub Actions**

Education

- **University of Houston** Houston, Texas
  Master of Science - Engineering Data Science; GPA: 3.9/4 Aug 2021 – Dec 2022

Honors and Awards

- University of Houston – Engr Dean's Scholarship (Aug 2021 – Dec 2022)