Pavan Kumar Gaggera

Chicago, IL | (773) 543-1890 | [pavank.gaggera@gmail.com](mailto:pavank.gaggera@gmail.com) | [LinkedIn](https://www.linkedin.com/in/pavan-gaggera/)| [GitHub](https://github.com/PavanKumarGaggera) | [LeetCode](https://leetcode.com/u/PavanKumarGaggera/)

## ****AI/ML Engineer | Generative AI | Deep Learning | MLOps****

## AI/ML Engineer with **2+ years** of experience in designing, developing, and deploying **scalable AI solutions**. Expertise in **Generative AI, LLMs, Multi-Modal Models, and Retrieval-Augmented Generation (RAG) systems**. Proven track record of optimizing **AI inference pipelines**, improving model performance, and deploying **production-ready AI systems**. Passionate about **cutting-edge AI research, automation, and model optimization**.

## ****Core Competencies****

## **Generative AI & LLMs** | Fine-Tuning & Prompt Engineering |Multi- Modal Models | Agentic AI Systems| LangChain &

## RAG (Retrieval-Augmented Generation) | **Machine Learning & Deep Learning** | PyTorch | TensorFlow | Model

## Optimization & Deployment | Genrative Models (GANs, VAEs) | Transformer-Based Architectures **Data Engineering**

## **& Vector Databases** | GPU Performane Tuning (CUDA) | Algorithm Development | FAISS, Pinecone, Weaviate | ETL

## Pipelines | AWS (SageMaker, EC2, S3) | Google Cloud (Vertex AI) | Docker | Kubernetes | FastAPI & Flask | AI-Powered

## Automation & Integration |**Big Data Processing** (Apache Spark, Hadoop)

## Technical Skills

## **Programming Languages**: Python, SQL, Java

## **Machine Learning Libraries**: Scikit-learn, TensorFlow, Keras, PyTorch, XGBoost

## Generative AI: GPT-4, Llama 3, Mistral, OpenAI API, LangChain, CrewAI, Gemini AI

## Vector Databases: FAISS, Pinecone, Weaviate, ChromaDB

## Automation & APIs: FastAPI, OpenAI Function Calling, Zapier, Notion API, Slack API

## **Deep Learning**: CNN, RNN, Generative Adversarial Networks (GANs), Transformers (BERT, GPT).

## **Natural Language Processing (NLP)**: Text Classification, Named Entity Recognition (NER), Sentiment Analysis, Word Embeddings

## Data Engineering & ETL: SQL Query Optimization, Data Pipeline Automation, AI-driven ETL Workflows

## **Data Visualization**: Matplotlib, Seaborn, Tableau

## **Cloud Platforms**: AWS (SageMaker, EC2, S3), Google (Vertex AI), Docker, Kubernetes

## **Version Control**: Git, GitHub

## PROFESSIONAL EXPERIENCE

## ****AI/ML Engineer, RIME SOFT**** | Jan 2022 – Nov 2022

* Developed an **AI-powered knowledge retrieval system** using **LangChain + Pinecone**, enhancing enterprise search accuracy.
* Improved **generation quality by 20%** through **hyperparameter tuning & transfer learning**.
* Reduced **deployment time by 30%** by integrating AI models into production pipelines.
* Designed **scalable AI automation workflows** integrating **RAG-based AI pipelines** with cloud infrastructure.

**Machine Learning Engineer Intern, RIME SOFT** | Aug 2021 – Dec 2021

* Built **predictive ML models** improving **data-driven decision-making** across business units.
* Designed **automated data pipelines**, reducing **data processing time by 40%**.
* Deployed ML models on **AWS SageMaker**, improving deployment efficiency by **30%**.

**EDUCATION**

**Master of Science (MS) in Data Science** – Lewis University

**Bachelor of Technology in Computer Science and Engineering** – Jawaharlal Nehru Technological University.

# KEY PROJECTS

### **AI Chatbot for Business Automation**

### Developed a **chatbot using LangChain & OpenAI API** to automate customer service workflows.

### Integrated **business logic and data retrieval** for a seamless user experience.

### Researched and implemented best practices in **NLP, chatbot design, and automation**.

### **LangChain-Powered Chatbot for Multiple PDFs**

* Built a LangChain-based chatbot capable of interacting with multiple PDFs, leveraging OpenAI's ChatGPT API and Hugging Face models.
* Implemented text extraction from PDFs, text chunking, and embedding generation using OpenAI Embeddings and Instructor Embeddings.
* Created a persistent conversation chain to maintain chat history and improve user interaction.

### **AI-Powered No-Code Web Scraping Agent**

* Built a**no-code AI agent**capable of scraping data from various sources, including**reviews, blogs, news articles, social media (LinkedIn, X, Instagram, Facebook), and YouTube.**
* Utilized**Relevance AI**and**Make.com**to create a multi-agent system, including a**Manager Agent, Web Scraper Agent,** and**Social Media Scraper Agent.**
* Integrated**Slack triggers**to automate scraping workflows and deliver results directly to Slack channels.

**Advanced Multi-Agent AI Application**

* Built an advanced multi-agent AI application using **Python, Langflow, Astra DB, and Streamlit**.
* Integrated multiple **LLMs** to handle different tasks, with a dynamic task-routing system for efficient task delegation.
* Developed a **full frontend interface** using Streamlit, enabling users to interact with the AI application seamlessly.
* Leveraged **Retrieval-Augmented Generation (RAG)** to enhance the accuracy and relevance of AI-generated responses.

## Leveraging Deep Learning for Pneumonia Detection in Lung Imaging

* Developed and trained a deep learning model using CNNs, ResNet, and VGG16/VGG19 on a dataset of 5,856 pediatric chest X-ray images to accurately classify pneumonia types (normal, bacterial, viral), utilizing Voxel51 for data management and advanced visualization, improving diagnostic efficiency.

## Predicting the Sale Price of Bulldozers Using Machine Learning

* Accomplished a 25% increase in model accuracy for predicting bulldozer sale prices, as measured by achieving an RMSLE score of 0.24, by using Random Forest model along with hyperparameter tuning.

## Heart Disease Prediction Using Machine Learning

* Accomplished a 30% improvement in prediction accuracy, as measured by achieving a 90% accuracy rate, by using Python and scikit-learn to build and tune the predictive model.

## Mental-Illness Prediction

## Accomplished a 90% prediction accuracy, as measured by identifying individuals at risk of mental illness, by applying machine learning algorithms on patient data.

## Fatigue Detection

## Accomplished 95% detection accuracy, as measured by improved road safety, by developing a real-time fatigue detection system using machine vision techniques and OpenCV.

## Web Scraping Job Requirements

* Accomplished a 50% reduction in manual data collection time, as measured by increased efficiency, by creating a Python-based web scraper to collect job posting data from various websites.

## Web Scraping 2-Star Books with Python

* Accomplished valuable insights into user sentiment, as measured by analyzing low-rated book reviews, by creating a robust web scraper using Python and BeautifulSoup.