Vinay Babu Gorantla

Email: Vinayc.gorantla@gmail.com Phone: +44-7780310625

Portfolio: Vinay Babu Gorantla - Portfolio

Professional Summary:

Machine Learning & MLOps Engineer with hands-on expertise in building and deploying scalable AI systems across NLP, Deep Learning, and cutting-edge Generative AI domains. Specialized in developing end-to-end ML and LLM workflows using tools like LangChain, LangGraph, LangSmith, LangServe, and modern vector store architectures such as ChromaDB and FAISS. Experienced in integrating OpenAI API, Hugging Face models, Gemma, and Groq for production-grade LLM applications.

Proficient in containerized deployments using Docker, experiment tracking with MLflow, and modular, testable pipeline design with Python, FastAPI, and Streamlit. Adept at applying advanced Natural Language Processing techniques and Retrieval-Augmented Generation (RAG) pipelines to solve real-world problems. Strong UK-based experience delivering cloud-ready, data-driven applications with real business value. Open to hybrid or remote roles across the UK.

Experience Summary:

- Built end-to-end Generative AI applications using OpenAI API, Hugging Face models, and Google's Gemini/Gemma models for LLM-based inference. Implemented RAG (Retrieval-Augmented Generation) systems with vector databases like ChromaDB and FAISS.
- Hands-on with LangChain components including Chains, Agents, Retrievers, Memory, Toolkits, and Document Loaders; proficient in LangGraph for stateful workflows, LangServe for deploying LLM apps as APIs, and LangSmith for prompt debugging, tracing, and evaluation.
- Integrated multiple LLM backends like Groq, OpenAI, Anthropic Claude, and Hugging Face to build chatbots, document assistants, and LLM-powered automation tools.
- Tuned LLM performance for latency, token efficiency, and stability, applying prompt compression, few-shot examples, and tracing using LangSmith to enhance user experience and reduce hallucinations.
- Designed reusable utility layers and backend services using FastAPI, enabling real-time interaction between frontend UIs and ML/LLM APIs.
- Skilled in building AI-powered web apps with persistent memory and conversational context using LangChain + Vector Stores, deployed via FastAPI, Streamlit, and Flask with ChromaDB or FAISS integration.
- Delivered scalable, production-ready MLOps solutions with MLflow for experiment tracking, DVC for versioning, Docker for containerization, and CI/CD pipelines via GitHub Actions.
- Developed and deployed supervised and unsupervised learning models for anomaly detection, risk profiling, and time series forecasting using Scikit-learn, XGBoost, and LightGBM.
- Proficient in traditional machine learning techniques including regression, decision trees, random forests, SVMs, clustering, and dimensionality reduction, with a focus on feature engineering and hyperparameter tuning using GridSearchCV and RandomizedSearchCV.
- Developed deep learning models (RNNs, LSTM, GRU, and transformer-based architectures) for sequence modelling, image classification, and NLP tasks using TensorFlow, Keras, and Hugging Face Transformers.
- Applied advanced Natural Language Processing (NLP) techniques such as entity recognition, sentiment classification, topic modelling, vector embeddings, and conversational search using libraries like spaCy, NLTK, and Sentence Transformers.
- Designed and deployed full-cycle ML pipelines from data ingestion and transformation to model training, validation, and deployment using Flask APIs, Streamlit dashboards, and modular Python codebases.

Previous Employment History:

Machine Learning Engineer (Contract)| VM2R Services LTD, City of Bristol, England United Kingdom

| Jan 2022 to Present

(Generative AI Focus – Mid 2023 to Present)

- Designed and deployed LLM-powered document QA and assistant applications using LangChain, integrating components like Chains, Memory, and Prompt Templates.
- Built Retrieval-Augmented Generation (RAG) pipelines using ChromaDB for vector storage and Sentence Transformers for dense embeddings.
- Integrated OpenAI API, Groq, and Hugging Face models (Flan-T5, Falcon, etc.) into chatbot and document summarization use cases.
- Utilized LangGraph for managing multi-step, stateful workflows and LangSmith for tracing, prompt testing, and performance evaluation.
- Deployed LangChain apps as APIs using LangServe and connected them to responsive frontend UIs for real-time LLM interaction.
- Conducted prompt engineering and evaluation to optimize LLM output relevance, response quality, and hallucination reduction.

(Machine Learning and Data Science – Feb 2022 to Mid 2023)

- Developed and deployed supervised and unsupervised learning models for anomaly detection, risk profiling, and time series forecasting using Scikit-learn, XGBoost, and LightGBM.
- Designed modular Python pipelines covering data ingestion, feature engineering, model training, evaluation, and scoring.
- Applied ensemble learning methods (Random Forests, Gradient Boosting) with hyperparameter tuning via GridSearchCV and RandomizedSearchCV.
- Leveraged NLP techniques (NER, sentiment analysis, topic modelling) to extract insights from text data using spaCy and NLTK.
- Operationalized ML workflows with MLflow for experiment tracking and DVC for data/model versioning.
- Containerized models using Docker and deployed RESTful APIs via FastAPI for real-time scoring.
- Built CI/CD pipelines using GitHub Actions to automate model testing, integration, and deployment.
- Created interactive dashboards to visualize model predictions and KPIs using Power BI.
- Collaborated with Data Engineering teams to scale solutions across Azure Synapse, Data Factory, and Databricks.

Data Scientist | WIPRO Technologies ltd, Bangalore, India | Nov 2018 to Jan 2022

- Led infrastructure alert analysis for a major Middle Eastern retail bank using anomaly detection techniques and statistical modelling, reducing false alerts from 80,000 to under 1,000 per month — resulting in significant operational cost savings and improved signal-to-noise ratio.
- Spearheaded data-driven transformation programs to streamline offshore banking operations, leveraging process mining and workflow automation techniques to deliver measurable gains in efficiency and service delivery.
- Collaborated cross-functionally to identify and resolve process bottlenecks through root cause analysis, enhancing overall productivity and enabling automation through Python-based scripting and dashboards.
- Conducted impact assessments and implemented data science-backed optimization strategies to improve operational KPIs across key banking departments, contributing to improved service levels and turnaround times.

Software Engineer | HCL Technologies, Chennai, India | Jun 2016 to Nov 2018

- Managed IT Service Management processes, including Incident, Change, Problem, and Service Disruption Management using BMC Remedy.
- Analysed and resolved recurring incidents, implementing root cause fixes to enhance application stability and reduce issue frequency.
- Supported deployment and release cycles by executing change management activities across test and

production environments.
- Wrote and optimized complex SQL queries for issue diagnostics and reporting, improving resolution turnaround time.
- Conducted knowledge transfer sessions and documentation for smooth onboarding of new team members.

Educational Qualifications:

B. Tech Computer Science and Engineering | VIGNAN's University | Aug 2010 to Apr 2014

Technical Skills Summary:

Generative AI & LLMs
- LangChain (Chains, Agents, Retrievers, Memory, Toolkits, Document Loaders)
- LangGraph (stateful LLM workflows), LangSmith (prompt debugging, tracing, evaluation), LangServe (LLM app deployment as APIs)
- OpenAI API, Groq, Gemini/Gemma, Hugging Face Transformers, Anthropic Claude
- Prompt Engineering, Retrieval-Augmented Generation (RAG) pipelines
- Vector Stores: ChromaDB, FAISS
- Embeddings: OpenAI Embeddings, Sentence Transformers
- LLM App Deployment with FastAPI, Streamlit, Flask

Core Machine Learning & NLP
- Scikit-learn, XGBoost, LightGBM, TensorFlow, Keras, Hugging Face Transformers
- RNN, LSTM, GRU, BERT, Flan-T5, Falcon
- Word Embeddings, Named Entity Recognition, Sentiment Analysis, Topic Modeling
- Regression, Classification, Clustering, Dimensionality Reduction, Time-Series Forecasting

MLOps & Deployment
- MLflow, DVC, GitHub Actions, CI/CD, Docker, Kubernetes (introductory)
- Model Versioning, Experiment Tracking, Logging, Monitoring
- API development using FastAPI, Flask, and Streamlit
- Modular, container-ready Python applications

Data Engineering & Wrangling
- Pandas, NumPy, SQL, PySpark, T-SQL
- Feature Engineering, Data Cleaning, Data Pipelines

Data Analysis & Visualization
- Matplotlib, Seaborn, Power BI
- Statistical Analysis, Hypothesis Testing, Exploratory Data Analysis (EDA)

Cloud & Tools
- Azure (ML Studio, Synapse, Data Factory, Data Lake Gen2), Databricks, AWS SageMaker (basic)
- Git, Bash, VS Code, Jupyter, Setup.py, requirements.txt

Programming & Scripting
- Python (Advanced), R (Basic), Shell/Bash (Basic), OOP Concepts

Disclaimer:

"I, Vinay Babu Gorantla, certify the accuracy of the information provided in this resume to the best of my knowledge. I am committed to ongoing learning and professional growth as a data scientist.