Raju Gopi

+91 9353973699 rajugraj29@gmail.com www.linkedin.com/in/raju-gopi-contact

SUMMARY

ML/Gen AI Engineer| 2+ Years of Experience

Results-driven **ML/Gen AI Engineer** with **2+ years of experience** in developing AI-driven solutions for business automation. Skilled in **LangChain, OpenAI, Google Gemini**, and **Hugging Face Transformers**, with expertise in **Retrieval-Augmented Generation (RAG), code modernization, and predictive analytics**. Proven ability to integrate scalable AI models, optimize performance, and streamline workflows through intelligent automation. Passionate about delivering high-performance solutions that enhance business efficiency and innovation.

EDUCATION

RNS Institute of Technology **08/2018 – 07/2022**
*Bachelor of Engineering – **8.1** CGPA* ***Bengaluru, India***

TECHNICAL SKILLS

Programming Languages: Python (Primary), Java, SQL
Generative AI & LLMs: Paid and Open-Source LLMs (OpenAI, Google Gemini Pro, Llama2 and Hugging Face Transformers)
AI/ML Technologies: Fine-tuning with custom data, Vector Embedding, NLP, Prompt Engineering, Apache Spark (PySpark), QLoRA (Quantized Low-Rank Adapters), LoRA (Low-Rank Adaptation), RAG (Retrieval-Augmented Generation)
Frameworks & Tools: LangChain, FastAPI, GitHub, VS code, Jira, DataBricks
Deployment & Platforms: GitHub Action, AWS, Docker, Kubernetes.
Visualization & Documentation: PowerBI, Microsoft Office
Data Science & Miscellaneous: Pandas, NumPy, Scikit-learn, Traditional ML Algorithms, MLFlow, Pyspark , pytorch
Frontend Development: HTML, CSS, Bootstrap

EXPERIENCE

Infosys

08/2022 – Present
Bengaluru, India

- **AI & NLP Model Integration**: Developed and integrated AI models like Google GenAI, OpenAI, and Hugging Face Transformers for automating business processes, including natural language understanding, task generation, and data categorization
- **Deep Learning Framework Implementation**: Developed and optimized neural networks using PyTorch for NLP tasks, implementing custom architectures and distributed training workflows that improved model accuracy by 15%
- **Model Optimization & Fine-tuning**: Fine-tuned Machine Learning Models like BART and GPT using domain-specific datasets, applied performance metrics such as sacrebleu for evaluation, and implemented LoRA/QLoRA techniques for efficient model adaptation
- **LangChain Framework Development**: Built advanced **LangChain** applications for document processing and conversational AI, integrating multiple **LLM providers** and implementing custom **prompt templates** for business-specific use cases
- **Traditional ML Algorithm Implementation & Optimization**: Developed and deployed classification and regression models using **Scikit-learn**, including **Random Forest**, **SVM**, and **Gradient Boosting** algorithms for predictive analytics, implemented comprehensive model evaluation pipelines using **GridSearchCV** and **cross-validation** techniques for hyperparameter tuning, achieving 85% accuracy on business forecasting tasks
- **Containerized ML Deployment & Orchestration**: Designed and deployed scalable ML models using Docker containerization and Kubernetes orchestration, implementing auto-scaling policies and CI/CD pipelines that reduced deployment time by 60%
- **End-to-End MLOps Pipeline Management**: Built comprehensive MLOps workflows using MLflow, automating model training, testing, deployment, monitoring, and drift detection with integrated A/B testing frameworks, achieving 95% pipeline success rate
- **Cloud-Native ML Infrastructure**: Architected production-ready ML solutions on AWS using SageMaker, implementing Infrastructure as Code with Terraform for reproducible deployments across multiple environments
- **Python Backend Development**: Developed scalable Python backends using FastAPI for web applications, creating high-performance APIs and robust integrations that improved data accessibility and processing efficiency by 40%
- **RAG & Vector Database Systems**: Created scalable VectorDB integrations for retrieval-augmented generation (RAG) applications, enabling advanced search and knowledge retrieval capabilities
- **Apache Spark Integration**: Designed and implemented large-scale data processing workflows using **PySpark** and **Apache Spark**, handling petabyte-scale datasets for ML model training and feature extraction with distributed computing capabilities
- **Data Pipeline Automation**: Designed comprehensive data workflows using SQL for extraction and manipulation across relational databases, with optimized queries supporting ML pipeline integration
- **Exploratory Data Analysis (EDA):** Conducted comprehensive data exploration, handling structured and unstructured datasets to identify trends, correlations, and anomalies, ensuring high-quality input for ML models.
- **Feature Engineering:** Applied feature engineering techniques to optimize data for **predictive modeling** and performance improvements.
- **Version Control & Collaboration**: Utilized **Git** and **GitHub** for version control and collaboration across teams, ensuring smooth development workflows and effective codebase management.
- **FMECA Analysis**: Led **Failure Modes and Effects Criticality Analysis (FMECA)** for gas turbines using **Excel Power Query** and **Power BI**. Automated data extraction with **Python**, reducing processing time by over 90%. Developed dashboards to highlight critical failure modes.

ACCOMPLISHMENTS

- **Insta-Award- Client Appreciation:** Designed and developed intuitive and visually compelling Power BI dashboards to present complex, large-scale data in a clear, accessible, and actionable format, enabling stakeholders to easily interpret and make informed decisions.