GIRISH GUPTA

(4084716287) | girishrg1249@gmail.com

# PROFESSIONAL SUMMARY:

* **AI/ML Engineer with 8+ years of industry experience specializing in designing, implementing, optimizing, and deploying machine learning models and LLM-powered applications, including RAG systems and intelligent agents.**
* **Strong knowledge of Machine Learning frameworks, including Scikit-learn, TensorFlow, PyTorch, and Keras,Numpy,Pandas for model building and evaluation.**
* Hands-on expertise in **Generative AI**, with extensive experience working with **Large Language Models (LLMs)** such as OpenAI GPT using Langchain.
* Proficient in **developing agentic frameworks** **LangChain, LangGraph**, and prompt engineering techniques
* Skilled in **fine-tuning and optimizing LLMs** for specific use cases, implementing best practices in model evaluation, performance tuning, and retrieval-augmented generation (RAG) workflows.
* Strong background in **AWS cloud services**, including **Lambda, S3, SageMaker, Bedrock, EC2**, and **API Gateway**, for building, deploying, and managing secure, scalable, and cost-effective AI solutions.
* **Excellent understanding of cloud-native services for ML workloads, including S3, Lambda, EC2,Redshift.**
* Proven expertise in deploying LLM-powered applications and agentic frameworks on AWS services, including Lambda, ECS, S3, ECR, and Bedrock, ensuring efficient model management, automation, and orchestration.
* Experienced in building traditional ML pipelines for classification, regression, and clustering tasks using Scikit-learn, XGBoost, and LightGBM, with robust data preprocessing and feature engineering.
* Skilled in designing SQL-based data pipelines for model training and evaluation, integrating MySQL, and Amazon Redshift for feature extraction, aggregation, and analytics
* Experienced in **Prompt Engineering**, specializing in designing, refining, and optimizing prompts for A**nthropic Claude** and **OpenAI GPT**, to solve complex software engineering and business problems.
* Proficient in **Python programming**, with hands-on experience in AI/ML-focused libraries such as **TensorFlow, PyTorch, Transformers**, and extensive knowledge of prompt engineering, RAG techniques, and model fine-tuning also expertise in Langchain

# TECHNICAL SKILLS:

| **Category** | **Skills / Tools** |
| --- | --- |
| **Programming Languages** | Python,SQL |
| **Generative AI & LLMs** | GPT, OpenAI APIs, Hugging Face Models, Bedrock APIs, Prompt Engineering, RAG Pipelines, LLM Fine-Tuning, LangChain (Agents, Tools, Memory |
| **Deep Learning Frameworks** | NumPy, Pandas, TensorFlow, PyTorch, Keras |
| **Neural Networks** | CNN, RNN, LSTM,Transformers, Activation Functions (ReLU, Leaky ReLU, Softmax), Optimizers (Adam, SGD). |
| **Data Processing & Analysis** | Pandas, NumPy, Matplotlib, EDA, Feature Engineering, Feature Scaling (Min-Max, Z-Score), PCA, SMOTE, ADASYN |
| **Data Science** | Hypothesis Testing, A/B Testing, Regression Analysis, Time Series Forecasting (ARIMA, LSTM), Clustering (K-Means, DBSCAN), Classification Metrics (F1, ROC-AUC), Model Interpretability (SHAP, LIME), Cross-Validation |
| **NLP & Text Analytics** | NLTK, spaCy, Word2Vec, Tokenization, Lemmatization, Stemming, N-grams, Named Entity Recognition, Sentiment Analysis |
| **MLOps** | Docker, Kubernetes, GitHub Actions, CI/CD,MLFlow,Databricks Workflows,Unity Catalog |
| **Vector DBs & LLMs** | Pinecone, GPT-4, LangChain. |
| **Cloud Platforms** | AWS (Lambda, S3, EC2, ECS, EKS, Glue, CloudWatch, Redshift, ECR, Bedrock) |
| **Big Data** | Apache Spark,Spark Sql |
| **Tools & Libraries** | BeautifulSoup, Streamlit, OpenAI API, Jira |
| **Optimization Techniques** | Hyperparameter Tuning, Grid Search, Dropout, L2 Regularization, Backpropagation, Batch Normalization, Ensemble Methods |
| **Deployment** | Bedrock, Lambda, API Gateway, Streamlit, GitHub Actions, AWS ECS/EKS,Datab |

**WORK EXPERIENCE:**

## Data Scientist Sept 2023 – Present

Fizer,NewYork,NY.

* Integrated LangChain Agents and Tools to enhance LLM capabilities with task-specific reasoning, retrieval, and tool-calling across GPT.
* Integrated AWS Bedrock to deploy foundation models like GPT enabling secure, scalable LLM solutions without infrastructure complexity.
* **Integrated LangChain framework to orchestrate the RAG pipeline**, utilizing Pinecone for vector storage and retrieval, and OpenAI’s GPT for answer generation.
* Utilized VS Code for developing, debugging, and deploying LLM and RAG pipelines with extensions like LangChain, Python, and Docker.
* Developed and fine-tuned Large Language Models (LLMs) using techniques like Low-Rank Adaptation (LoRA), and Parameter-Efficient Fine-Tuning (PEFT).
* **Implemented Few-Shot, One-Shot,Zero-Shot,Persona Prompting techniques**, optimizing LLM behavior through prompt engineering strategies.
* Designed multi-step AI workflows using LangChain Agents and Tools, coordinating task execution across LLMs like GPT to enhance contextual reasoning and retrieval in RAG pipelines.
* C**onducted structured data modeling and implemented feature engineering techniques to optimize model performance across various business use cases.**
* **Preprocessed textual data by applying data cleaning, tokenization, normalization, and storage techniques, ensuring high-quality inputs for embeddings and retrieval tasks.**
* **Developed end-to-end chatbot pipelines using LangChain’s ConversationalRetrievalChain and other modular chains (e.g., RetrievalQAChain, LLMChain) to enable context-aware, accurate responses.**
* **Built intuitive front-end interfaces using Streamlit to facilitate seamless interaction with AI-powered applications, supporting dynamic user queries** with backend services deployed on **AWS EC2**/**EKS**.
* Applied **parameter-efficient fine-tuning** (PEFT) techniques such as **LoRA, Adapters** on **LLMs** (GPT, Claude) using **Bedrock APIs.**
* Deployed LangChain Agents on scalable AWS ECS/Fargate using Docker containers, with orchestration managed through AWS EKS for high availability and performance.
* Built LLMOps workflows with SageMaker Pipelines, S3, and Lambda for automated RAG LLM retraining and deployment using LangChain
* Developed and deployed scalable machine learning models using **Python** within **Databricks Notebooks**.

## ML Engineer Mar 2020 – Aug 2023

State Farm,North Calorina

* Developed LSTM models for text generation, translation, and time-series prediction using PyTorch and TensorFlow Keras.
* Implemented time series models (LSTM, SARIMAX) to optimize inventory management **Pandas**, **SQL**, and **NumPy** for historical trend analysis.
* Built and trained deep learning models using TensorFlow, Keras, and PyTorch, including ANN,CNN,RNN.with hidden layers and activation functions like ParametricReLU, Sigmoid, Softmax , leveraging **Python** and **Pandas** for AIML preprocessing.
* Tuned hyperparameters such as number of neurons, learning rate, batch size, and epochs using techniques like grid search to achieve optimal performance supported by **NumPy AIML** computations.
* Leveraged **SQL** queries on **Amazon Redshift** to retrieve and join multilingual datasets, followed by text normalization, tokenization, and lemmatization
* Queried multilingual corpora from **Amazon Redshift** using **SQL**, performing tokenization and lemmatization with NLTK for AIML applications.
* Applied regularization techniques like Dropout and L2 Regularization to prevent overfitting and improve generalization.
* Built **data pipelines** on **Apache Spark** for large-scale data ingestion, transformation, and aggregation.
* Utilized **SparkSQL** for querying and analyzing structured data efficiently within distributed environments. Developed and fine-tuned deep learning models like LSTM,Transformers,BERT for NLP tasks using PyTorch and TensorFlow.
* Performed **data cleansing, validation, and transformation** to ensure quality and consistency in model training.
* Automated end-to-end workflows in **Databricks Workflows** to support scheduled training, testing, and deployment.
* Automated ETL workflows for text preprocessing, including tokenization, lemmatization, and vectorization, using AWS Lambda and S3 triggers to preprocess unstructured data with SQL Queries.
* Implemented CI/CD pipelines for Transformers, LSTM models using PyTorch and TensorFlow, with automated retraining while retaining data from S3.
* Utilized TensorFlow and PyTorch to develop and optimize NLP models for tasks like text classification and named entity recognition.
* **Utilized Python-based development in VS Code, automating data preprocessing (AWS Lambda, S3), model training (SageMaker), and evaluation with Pandas, Scikit-learn,Numpy,SQL.**

## Data Scientist Jan 2017 - Feb 2020

## Chewy,South Calorina.

* Achieved reduction in customer churn rates by developing and implementing ML models (random forest,XGBoost) using Sikitlearn and preprocessing using **Python** and **NumPy**, identifying key attrition drivers to design SQL-backed retention strategies stored in **Redshift**.
* Explored datasets using **Python** in VS Code, performed EDA to uncover actionable insights, and cleaned data by handling missing values and outliers on datasets stored in **AWS S3** and queried via **SQL**.
* Applied normalization and scaling techniques (Min-Max, Z-score) using **NumPy** and **Scikit-learn** to standardize input features for neural network models.
* Leveraged **Pandas**, **NumPy**, **Matplotlib**, and Seaborn in VS Code to generate histograms, scatter plots, box plots, and heatmaps; integrated into **MLOps** reporting pipelines on **EC2** and **ECS**.
* **Reduced dimensionality and improved computational efficiency using PCA, LDA, and t-SNE techniques in Python (Scikit-learn, TensorFlow), deploying optimized datasets into AWS S3 .**
* Developed and optimized supervised learning models (Decision Trees, Random Forests, XGBoost, LightGBM) in PyTorch and TensorFlow within VS Code, deploying trained models.
* Engineered features from raw datasets using PySpark and Python to improve model performance and business outcomes.
* Integrated **MLflow** for model tracking, experiment management, and lifecycle control.
* Implemented **Unity Catalog** to govern and manage access to data assets across workspaces and teams.
* Designed and deployed end-to-end ML models for classification/regression using **Python**, deployed via **MLOps** pipelines on **SageMaker** and **EKS**, and logged results to **Redshift** for long-term analysis.
* Created histograms, scatter plots, and heatmaps using **Pandas** and **NumPy**, visualization purposes.
* Conducted EDA on large-scale datasets with **Pandas** and **NumPy** to identify trends and inconsistencies, with findings logged into **Redshift** tables for analytics through **SQL** dashboards.
* Performed data preprocessing, including missing value imputation (KNN, MICE), outlier detection (IQR/Z-score), and type corrections, storing clean data in **DynamoDB** and **S3**.
* Tuned model hyperparameters using Grid Search and Random Search with **NumPy**, automating training and evaluation in **MLOps** pipelines integrated with **SQL**-based feature stores.
* Designed feature engineering pipelines (encoding, scaling, transformation) with **Python**, **Pandas**, and **NumPy** to enhance model accuracy, integrated into **SageMaker Pipelines** for **MLOps** automation.
* Designed and deployed end-to-end ML models for classification and regression, improving business outcomes with scalable deployment on **Sa**and **EKS**.
* Developed a hybrid approach combining filter, wrapper, and embedded methods to improve model robustness and reduce noise in structured datasets.
* Tuned model hyperparameters using Grid Search and Random Search, with automated training and evaluation
* Implemented L1 (Lasso) and L2 (Ridge) regularization-based feature selection to handle high-dimensional datasets in regression tasks.

**EDUCATION-**

Master’s of Engineering in Artificial Intelligence,University of Cincinnati.