SAI HARSHITHA PADALA

pharshithawork@gmail.com || +1-980-352-6677|| https://www.linkedin.com/in/psharshitha/

PROFESSIONAL SUMMARY

- Data Scientist with 9+ years of experience specializing in LLM development, RAG pipelines, and scalable Gen AI systems using LangChain and LangGraph.
- Strong expertise building AI agents that combine LLM reasoning, tool use, retrieval steps, and workflow orchestration.
- Hands-on experience working with vector databases (FAISS, Chroma, Pinecone-equivalent architectures) for embedding search and retrieval.
- Advanced background in NLP, including text classification, summarization, embeddings, and transformer-based model development.
- Deep proficiency in deep-learning using PyTorch and TensorFlow, including training, optimization, and neural network experimentation.
- Built and productionized pipelines leveraging BERT, sentence encoders, and transformer models for downstream ML and RAG applications.
- Applied LoRA and other parameter-efficient tuning methods to fine-tune LLMs for domain-specific Gen AI tasks.
- Designed robust RAG architectures that improve grounding, retrieval accuracy, and overall LLM response quality.
- Strong understanding of data governance, ensuring metadata quality, lineage tracking, reproducibility, and compliant ML operations.
- Experienced deploying end-to-end LangChain and LangGraph pipelines integrating embeddings, LLMs, AI agents, and retrieval logic into production.
- Skilled in designing retrieval-optimized embedding pipelines using transformer encoders and vector search to support high-accuracy RAG workflows.
- Experienced in building multi-agent Gen AI systems where LLM tools, routing logic, and reasoning chains collaborate to complete complex tasks.
- Strong command of feature engineering, model evaluation, and experimentation, including A/B testing, hyperparameter tuning, and grounding analysis.

CERTIFICATIONS

- Certified Oracle Cloud Infrastructure 2025 Certified Generative AI Professional LINK
- Coursera- Data Science Professional Certificate- IBM LINK

TECHNICAL SKILLS

Generative AI	LLMs, Gen AI Pipelines, RAG (Retrieval-Augmented Generation), Fine-Tuning, LoRA/QLoRA, Prompt Engineering, CoT (Chain-of-Thought), Model Guardrails, Structured Outputs
Agentic AI	LangGraph, LangChain, Multi-Agent Workflows, AI Agents, Tool Calling, Planning & Routing, Memory Components, Model Context Protocol (MCP), Observability Hooks
LLM Ecosystem	Vertex AI LLMs, Gemini Models, Hugging Face Transformers, Embedding Models, Sentence Transformers, BERT, FAISS, Vector Indexing, Semantic Search
RAG Stack	Vertex AI Vector Search, BigQuery Retrieval, Metadata Filtering, Document Chunking, Context Augmentation, Embedding Generation, Ranking Pipelines
LLM Tuning & Training	Vertex AI Training, Fine-Tuning, PEFT (LoRA/QLoRA), Training Pipelines, Evaluation Frameworks, Model Monitoring, Output Validation
NLP & Deep Learning	Text Classification, Summarization, Tokenization, Embeddings, Transformer Models, PyTorch, TensorFlow, Attention Mechanisms
Data Governance & Modeling	Data Quality, Metadata Management, Lineage Tracking, Reproducibility, Compliance-Aware ML Workflows
Cloud Platforms	Google Cloud Platform (GCP), Vertex AI, Cloud Run, Cloud Functions, Cloud Storage (GCS), Pub/Sub, Artifact Registry
Programming & Frameworks	Python, FastAPI, Async I/O, PyTorch, TensorFlow, Scikit-Learn, REST APIs, JSON/Parsers
Orchestration & Integration	Agent Routing, LLM Toolchains, Modular Retrieval Layers, MCP Tools, Event-Driven Pipelines
Evaluation & Optimization	Prompt Testing, LLM Evaluation Metrics, Latency Optimization, Cost Optimization, Hallucination Reduction, Token Efficiency
Tools & Platforms	Jupyter, VS Code, Postman, Application Insights, MLflow (tracking & monitoring)

PROFESSIONAL EXPERIENCE

Client: Optum , Chicago, IL
Role: Data Scientist-Gen AI Engineer
Duration: October 2023 – Present

Project Scope:
Designed and deployed end-to-end Gen AI solutions on GCP using LLMs, RAG pipelines, LangChain, LangGraph, vector embeddings, and AI agent frameworks. Built scalable inference and training workflows integrating prompt optimization, model evaluation, and automated reasoning capabilities.

- Built Gen AI pipelines using LLMs, embeddings, and retrieval components to automate multi-step analytical workflows.
- Designed RAG architectures using vector DBs and hybrid search to ground LLM outputs with factual context.
- Developed AI agents with LangChain and LangGraph, combining tool use, reasoning steps, and workflow orchestration.
- Fine-tuned LLMs using Hugging Face to support domain-specific tasks and improve instruction-following accuracy.
- Implemented prompt engineering patterns (few-shot, system prompts, structured templates) for stable Gen AI performance.
- Built feature extraction, preprocessing, and EDA scripts in Python to align datasets for ML and Gen AI model development.
- Created evaluation pipelines tracking AUC, F1, BLEU, grounding scores, and drift metrics for LLM and RAG components.
- Developed scalable vector-embedding modules integrated with FAISS and BigQuery for fast retrieval inside Gen AI workflows.
- Productionized LLM inference services on GCP using Vertex AI, Docker, and CI/CD for reliable, monitored deployments.
- Implemented guardrails, validation rules, and output checks to ensure responsible and predictable Gen AI behavior.
- Optimized RAG chunking, metadata, and retrieval scoring to reduce hallucinations and improve model grounding.
- Collaborated with engineering teams to integrate LLM agents, RAG retrieval, and ML components into unified Gen AI pipelines.
- Engineered end-to-end data pipelines that feed cleaned, enriched datasets into LLM, RAG, and traditional ML models for unified Gen AI workflows.
- Designed multi-stage retrieval + reasoning flows where RAG handles grounding and LLM agents handle logic, summarization, and decision steps.
- Implemented continuous training and re-ranking strategies to improve vector embeddings and retrieval relevance across Gen AI applications.
- Developed monitoring dashboards combining ML metrics, LLM quality scores, and retrieval diagnostics to proactively maintain Gen AI system performance.

Environment: Python, LLMs, Gen AI, RAG, LangChain, LangGraph, Hugging Face, OpenAI, Vertex AI, BigQuery, FAISS, Vector DBs, Databricks, GCP, Docker, CI/CD, GitHub, Prompt Engineering, AI Agents

Client: Spencer Health Solutions, Morrisville, NC
Role: Data Scientist
Duration: December 2021 – July 2023
Project Scope:
Built ML pipelines on AWS SageMaker to predict member adherence, risk scores, and payer cost drivers using structured claims, enrollment, and pharmacy data.Developed an early RAG-style retrieval workflow using S3 + Athena + embeddings to pull historical claims, formulary rules, and provider notes for analytics use cases.

- Developed ML models for risk scoring, adherence prediction, and member stratification using Python, scikit-learn, and AWS SageMaker distributed training.
- Created ETL pipelines using AWS Glue + Lambda + Athena to standardize claims, pharmacy fills, encounter data, and eligibility files.
- Implemented an early RAG workflow where embeddings stored in S3 retrieved clinical and claims snippets to support analytics interpretation.
- Built feature engineering scripts to derive chronic-condition flags, episode-of-care timelines, utilization frequencies, and medication adherence metrics.
- Designed SageMaker inference endpoints to deploy models for real-time payer analytics dashboards.
- Integrated formulary rules, provider network metadata, and medication-tier information into model inputs for more accurate payer predictions.
- Automated dataset refresh cycles using Step Functions for claims, provider directories, medication lists, and historical outcomes.
- Developed PyTorch-based sequence models to analyze refill patterns, gaps in therapy, and multi-drug compliance behaviors.
- Built Athena queries to process millions of claims records, mapping CPT/HCPCS codes to cost drivers and UM decision variables.
- Implemented explainability using SHAP/LIME for model transparency across UM and care-management teams.
- Prepared model validation reports aligned with payer accuracy, fairness, and audit requirements.
- Collaborated with pharmacists, clinical analysts, and data engineers to test adherence-prediction outputs and ensure trust in the model.
- Created S3-based embedding stores for retrieving prior cases, formulary exceptions, and provider patterns.
- Supported internal analytics teams with Python utilities for data cleaning, ICD/CPT grouping, and time-bound patient history extraction.

Environment: Python, AWS SageMaker, AWS Glue, Athena, S3, PyTorch, XGBoost, LightGBM, Docker, Boto3, ICD-10/CPT/HCPCS, claims & pharmacy datasets.

Client: USCC Chicago , Illinois USA
Role: Data Scientist -Machine Learning Engineer
Duration: December 2019 – November 2021
Project Scope:
Developed machine learning and PySpark workflows to solve key telecom business problems—predicting customer churn, improving retention strategies, and optimizing revenue across subscriber segments.

- Developed end-to-end churn forecasting pipelines using PySpark on AWS EMR, integrating daily subscriber activity, billing records, and call center interactions into ML-ready datasets.
- Built and maintained PySpark-based ETL pipelines to process customer usage, billing, and interaction data for downstream predictive modeling.

- Developed machine learning models for churn prediction, customer segmentation, and ARPU forecasting using Python (scikit-learn, TensorFlow).
- Designed feature stores and transformation logic in SQL to standardize data inputs for model training and validation.
- Automated data extraction, preprocessing, and scoring workflows using Airflow DAGs and shell scripts to ensure repeatable production runs.
- Implemented model monitoring and retraining triggers based on data drift and performance degradation using Python-based automation.
- Collaborated with marketing and operations teams to translate predictive insights into retention strategies and campaign targeting.
- Deployed models as RESTful APIs using Flask and Docker, integrating outputs with internal analytics dashboards
- Performed hyperparameter tuning, cross-validation, and model explainability studies to improve prediction accuracy and transparency.
- Supported migration of analytical workloads from on-prem Hadoop to early Databricks and cloud-based infrastructure for scalability and maintainability.

Environment: Snowflake, PySpark, AWS EMR, SageMaker, Redshift, Lambda, Step Functions, SQL, Python, scikit-learn, XGBoost, TensorFlow, Tableau, QuickSight, GitHub, AWS Glue, Confluence

Client: Cygnet Infotech, Hyderabad , India
Role: Data Analyst
Duration: June 2016 – September 2019
Project Scope:
Built analytics dashboards and automated reporting workflows to solve real customer-service challenges—tracking SLAs, reducing escalations, and improving NPS/CSAT insights for operations teams.

- Analyzed customer-service and product-usage data to identify trends in resolution times, escalation rates, and recurring issue categories for performance optimization
- Created interactive Power BI and QlikView dashboards tracking SLA compliance, customer-satisfaction metrics, and agent-level performance KPIs used by operations leadership
- Developed Excel-based reconciliation reports to track monthly billing discrepancies, refunds, and invoice-level anomalies, ensuring financial accuracy and transparency
- Collaborated with business teams to define KPI logic and automated weekly / monthly reporting using SQL queries and Excel macros, improving report turnaround time
- Built user-friendly Excel dashboards with PivotTables, slicers, and conditional formatting to help non-technical stakeholders filter data by region, agent, or issue type.
- Partnered with QA and product teams to categorize issues by severity and frequency, helping prioritize bug fixes and product enhancements
- Designed scorecards and ranking charts to visualize weekly NPS, CSAT, and agent-level satisfaction metrics for performance reviews
- Supported quarterly business reviews by preparing trend analyses and visual summaries of performance metrics, customer feedback, and SLA attainment

Environment: SQL, Power BI, QlikView, Excel, PivotTables, VLOOKUP, Excel Macros, Slicers, Conditional Formatting, Customer Support KPIs, NPS, CSAT, SLA Metrics

EDUCATION

Bachelor of Technology (B. Tech) in Information Technology

KLUniversity *May 2016*

Vijayawada, Andhra Pradesh, India