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 designing and deploying AI and Generative AI solutions for large-scale healthcare workflows.
- Strong hands-on expertise in Python for model development, data processing, and end-to-end AI pipeline automation.
- Extensive experience building and optimizing deep-learning models using TensorFlow and PyTorch.
- Developed and customized LLM-based solutions for clinical summarization, document understanding, and healthcare analytics.
- Designed retrieval-augmented generation (RAG) pipelines integrating embeddings, vector search, and medical context retrieval.
- Worked with Azure AI services for hosting LLM inference, orchestrating workflows, and deploying ML components at scale.
- Built embedding pipelines using transformer encoders to support RAG, search, and retrieval-focused applications.
- Strong foundation in NLP covering text classification, summarization, semantic search, and transformer-based modeling.
- Applied fine-tuning and optimization strategies to improve model accuracy for healthcare-specific tasks.
- Experienced converting unstructured clinical and payer data into usable features for downstream ML and GenAI systems.
- Implemented evaluation strategies including model validation, grounding checks, and performance benchmarking.
- Supported healthcare decision-making by developing AI components that improved automation and reduced Manual review.
- Worked end-to-end across data ingestion, model development, deployment, monitoring, and continuous optimization on Azure.

CERTIFICATIONS

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

TECHNICAL SKILLS

Generative AI	LLMs, Generative AI Pipelines, RAG (Retrieval-Augmented Generation), Fine-Tuning, Prompt Engineering, Model Guardrails, Structured Outputs
Deep Learning & ML	TensorFlow, PyTorch, Scikit-Learn, Neural Networks, Embeddings, Text Classification, Summarization, Transformer Models
Azure AI Ecosystem	Azure OpenAI, Azure Cognitive Search, Azure Functions, Azure Kubernetes Service (AKS), Azure Container Registry, Azure Storage, Azure SQL
RAG & Retrieval Systems	Embedding Generation, Vector Search, FAISS-style Indexing, Document Chunking, Context Augmentation, Retrieval Orchestration
NLP	Tokenization, Named Entity Extraction, Medical Text Processing, Semantic Search, Domain-Specific Text Cleaning
Model Development & Optimization	Model Training, Fine-Tuning, Evaluation Metrics, Hyperparameter Tuning, Drift Detection, Output Validation
Python Engineering	Python, FastAPI, Async Workflows, REST APIs, JSON/Parsers, Modular AI Components, Pipeline Automation
Healthcare Data	Claims Data, Clinical Documents, EHR/Provider Notes, HIPAA-Compliant Data Workflows
Programming & Frameworks	Python, FastAPI, Async I/O, PyTorch, TensorFlow, Scikit-Learn, REST APIs, JSON/Parsers
Tools & Platforms	Git, CI/CD, Jupyter, VS Code, Postman, MLflow (tracking), Application Insights

PROFESSIONAL EXPERIENCE

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

Project Scope:
Built Azure-based GenAI solutions using LLMs and RAG pipelines, and I developed models in Python using frameworks like TensorFlow and PyTorch to support healthcare data workflows. My focus was on turning unstructured clinical documents into accurate, retrieval-enhanced outputs for care management and analytics.

- Developed AI and ML models using Python, TensorFlow, and PyTorch to support clinical analytics, risk prediction, and operational workflows.
- Designed and implemented LLM-based features on Azure, including medical summarization, eligibility interpretation, and document understanding.
- Built end-to-end RAG pipelines using embedding generation, vector indexing, and retrieval layers optimized for healthcare content.
- Preprocessed and transformed unstructured clinical documents using tokenization, normalization, and embedding-based text pipelines.
- Fine-tuned LLMs for domain-specific tasks involving claims narratives, care management notes, and provider documentation.
- Developed scalable inference services on Azure using containerized deployments and model-serving best practices.
- Integrated LLM outputs into clinical decision workflows using REST APIs and structured response templates.
- Automated feature extraction from lab results, progress notes, and medical histories for downstream ML applications.
- Engineered embeddings for clinical terminology, ICD/CPT codes, and payer rules to improve retrieval and contextual reasoning.
- Optimized model performance through hyperparameter tuning, evaluation metrics, and error analysis.
- Implemented data pipelines for ingesting, cleaning, and validating healthcare datasets from claims, EHR, and provider systems.
- Developed RAG orchestrations that combined retrieval, model reasoning, and prompt routing for high-accuracy responses.
- Built monitoring checks for drift, latency, and inference reliability across Azure-hosted GenAI services.
- Collaborated with clinical SMEs to align model outputs with medical guidelines and payer policies.
- Integrated Azure Cognitive Search and vector-based retrieval to enhance LLM contextual grounding.
- Worked with secure healthcare data under HIPAA guidelines, ensuring compliance in all AI workflows.
- Created reusable AI components and Python utilities that standardized training, inference, and evaluation across teams.
- Implemented prompt-engineering strategies, including few-shot templates and medical context prompts, to improve LLM accuracy on clinical and payer tasks.

Environment: Azure, Python, TensorFlow, PyTorch, Azure OpenAI, Azure Cognitive Search, Hugging Face, RAG pipelines, vector embeddings, REST APIs, Git, CI/CD, healthcare claims and clinical documents.

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.
- Optimized pipeline performance by migrating heavy queries to Glue Spark jobs for large historical claims processing.
- 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