Afrin Peshimam

EMAIL: PESHIMAM.AFRIN@GMAIL.COM | LINKEDIN: LINKEDIN.COM/IN/AFRIN-PESHIMAM/
PHONE: +91 8169741805

Professional Summary

- **Generative AI Engineer/Architect** specializing in designing and deploying sophisticated GenAI solutions on **Microsoft Azure** with prior experience as **Senior Data Scientist** with proven success in demanding **startup** environments, recognized with **$62,500 phantom equity** and **two "Employee of the Quarter" awards**.
- Skilled at managing multiple concurrent projects and translating complex requirements into scalable, data-driven solutions.
- Proven success architecting **multi-agent virtual assistants**, advanced **RAG** systems, and leveraging NLP, LLM/embedding fine-tuning for patent analysis, information extraction, and classification, all while ensuring robust **AI Governance**, security, and compliance.
- Eager to apply technical leadership, startup agility, and a commitment to innovation to drive cutting-edge, data-driven solutions in a growth-oriented organization

WORK EXPERIENCE

Generative AI Engineer

Nallas Corporation (Dec 2024 – Present)

1. **Azure Enterprise GenAI Multi-Agent Virtual Assistant**

Summary:

- Architected an end-to-end enterprise-grade virtual assistant for the client, enhancing productivity across Content and Member Engagement teams with a strong focus on **AI governance, security, data compliance, and scalability** through ethical-AI best practices.
- Engineered **a multi-agent AI architecture using Azure AI Agent Service and the AutoGen framework to orchestrate** specialized Azure AI agents hosted on Azure Container Apps (ACA), integrating **continuous agent tracing, evaluation, monitoring, and AI guardrails using Azure AI Content Safety** to optimize performance, ensure reliability, and maintain responsible AI standards.
- Architected the **Retrieval-Augmented Generation (RAG) pipeline** using **Azure AI Search** for grounding responses in internal client resources and implemented **Azure Bing Search grounding** for access to real-time external information.

Technologies Used:

- **AI/ML:** Azure AI Foundry, Azure AI Agent, AutoGen Framework, Azure OpenAI Service, Azure AI Search (RAG), Azure AI Content Safety, Azure AI Evaluation (Safety, Security, Agent Evaluation, RAG Evaluation), Azure AI Document Intelligence, Azure AI Language, Azure Bing Search Grounding, Python, Docker.
- **Azure PaaS/Cloud:** Azure Container Apps (ACA), Azure API Management (APIM), Azure Cosmos DB, Azure Blob Storage, Azure Functions, Azure Key Vault, Azure Monitor (Application Insights, Log Analytics for Monitoring & Tracing), OpenTelemetry, Azure VNet, Azure Private Link, Network Security Groups (NSGs), Microsoft Entra ID.
- **Security & Governance:** Microsoft Defender for Cloud, Microsoft Sentinel, Microsoft Purview
- **Development & Operations:** REST APIs, Git, Azure DevOps.

2. **AI-Assisted SDLC**

Summary:

- Leveraging CrewAI Multi-Agent technology to revolutionize the software development lifecycle (SDLC).
- Designed and implemented AI-driven solutions to optimize collaboration among software agents, enhancing efficiency and minimizing manual intervention.
- Integrated generative AI capabilities to automate critical SDLC tasks, including documentation creation, requirement analysis, code generation, and code reviews.

- Enabled dynamic generation of system design artifacts, including Low-Level Design (LLD) and High-Level Design (HLD) documents.
- Introduced a feature allowing users to input minimal information such as project name, project goals, and a list of product features to automatically generate a comprehensive requirements document, followed by end-to-end SDLC automation.
- Automated task generation and assignment directly to Jira for streamlined project management and tracking.

Technologies Used: CrewAI, Flask, Azure DevOPs, Jira, Gemini

Senior Data Scientist

XLSCOUT (3 YEARS)

- **Phantom Equity:** Awarded **$62,500** in phantom equity for exceptional performance and contribution to company growth.
- **Employee of the Quarter (Q3, 2022 & Q1, 2023):** Recognized for outstanding project leadership, innovative AI solutions, and consistent performance

1. Patent Drafting LLM with Chatbot

Summary:

- Developed a groundbreaking solution that helps law firm, streamline the patent drafting process for both United States and Japan jurisdictions.
- **Azure OpenAI GPT-4** was used to draft English and Japanese patents, ensuring accuracy and compliance with intellectual property laws.
- **Azure GPT-4 Vision** was used to convert patent drawings into concise brief descriptions of drawings.
- In addition, the process was streamlined by converting brief descriptions of drawings into understandable diagrams such as flow charts and block diagrams.

Technologies Used: FastAPI, Azure OpenAI GPT-4, Azure GPT-4 Vision, AWS Bedrock, Claude, MongoDB, Mermaid, Celery, Redis

2. Hybrid Search Tool for Patent CPC Code Prediction

Summary:

- Built a **hybrid search** tool to predict the top 20 relevant CPC codes using patent text and CPC hierarchy data.
- **Fine-tuned BGE embeddings** on patent data, improving accuracy and relevance in CPC code predictions and deployed the model on AWS with an inference endpoint.
- Enhanced search efficiency by integrating traditional and vector-based search methods.

Technologies Used: LlamaIndex, Vector Database Weaviate, AWS Sagemaker, BGE embeddings, FastAPI

3. Information Extraction

Summary: Developed an **Information Extraction** system for English and Japanese patents, which extracts raw materials along with its quantities, ionic conductivity, electric conductivity, and end product outcomes. Completed this client project, securing an annual budget of $100k.

Technologies Used: LangChain , FAISS Vector Store, Azure OpenAI GPT-4

4. Patent Summarization with LLM Mistral-7b

Summary: Developed tools for **Patent Summarization** using instruction **fine-tuned LLM Mistral-7b**. Utilized AZURE GPT4 for building training data and optimized finetuning efficiency with QLoRA and PEFT. Model was deployed using **Hugging Face Text Generation Inference** on Azure.

Technologies Used: AZURE GPT4, Hugging Face, QLoRA, PEFT, TGI

5. English to Japanese Translation results

Summary: Fine-tune SLM for English-to-Japanese translation model for patent documents by fine-tuning on a bilingual dataset of 150,000 entries, ensuring precise conversion of complex legal terminology essential for international patent applications.

Technologies Used: FastApi, HuggingFace transformers, GCP, Celery, Redis.

6. **Patent Diagram Figure Extraction Using YOLOv10 (Finetuned Model)**
Summary: Developed a **Fine-tuned YOLOv10** model that extracts patent diagram images from a given patent number and its associated PDF, achieving **98.5% mAP@50** for high accuracy and efficiency, significantly reducing manual effort in patent analysis.
Technologies Used: YOLOV10, ROBOFLOW.

7. **Patent Classification**
Summary: Developed a NLP system to classify patents, into 66 categories. results
Technologies Used: Combined Supervised and Unsupervised approach to solve this problem. Bert Topic for topic modelling, dimensionality reduction using Truncated SVD, clustering using HDBSCAN, Support Vector Machine SVM, Word Embedding using Universal Sentence Encoder.

8. **Non-Patent Literature Extraction and Ranking**
Summary: Developed a FastAPI which takes a query and extracts NPL data from various journals, ranking it based on semantic similarity.
Technologies Used: Data extraction using google api, urllib, pdfminer, beautiful soup, crossref api. Word embeddings were made using Sentence Transformer & ranking by calculating cosine score. The entire task was done parallel & distributed using python multiprocessing & ray.

9. **SDI – A Patent Noise Filtering Tool**
Summary: Developed a patent noise filtering tool. Trained multiple ML models using a set of patent documents that are labelled as relevant and not relevant (noise) for a particular domain. Compared the models using different performance metrics such as confusion matrix, Recall, Precision, F1-score etc.
Technologies Used: scikit-learn, **Sentence Transformers** , Universal Sentence Encoder, nltk, spacy, tensorflow, SVM, Random Forest, KNN, Logistic Regression.

Data Science Internship iNeuron (1 Year) *1stRank* *Certificate*

TECHNICAL SKILLS

Programming Language & Database :
- Python, SQL (MySQL, PostgreSQL), NoSQL (MongoDB, Redis, Cassandra), CosmosDB

AI Tools & Technologies:
- Azure OpenAI, Microsoft Azure AI Foundry, AWS Bedrock, GPT, Claude Models, LangChain, LlamaIndex, CrewAI, AutoGen, Semantic Kernel

Machine Learning & NLP:
- Traditional ML: Linear Regression, Logistic Regression, Random Forest, Gradient Boosting Trees, SVM
- Libraries/Frameworks: Apache Spark, Pandas, NumPy, Scikit-Learn
- Deep Learning: Transformers, Sentence-BERT, Gensim, Spark MLlib, Keras, TensorFlow, PyTorch, LSTM
- NLP Tools: NLTK, spaCy, Spark NLP, Word2Vec

Data Visualization:
- Tableau, Power BI, Matplotlib, Seaborn, Plotly, Folium & Excel

Web Scraping:
- Scrapy, Beautiful Soup, Playwright, Requests & Urllib

MLOps & Model Deployment:
- MLflow, CI/CD, Git, Docker, Jenkins, Celery, FastAPI, Databricks, AWS, Google Cloud, Azure ML, TGI (Text Generation Inference), vLLM

Front-End Development:
- HTML, CSS, Bootstrap, Streamlit, Kivy, Flutter

Parallel & Distributed Processing:
- Ray, Dask, PySpark & Multiprocessing

PROJECTS

Multi-Source Hybrid RAG Q&A Chatbot

- Developed a **Retrieval-Augmented Generation (RAG)** chatbot capable of answering questions from a multi-format knowledge base, including **PDFs, DOCX, text, web pages, and YouTube links**.
- Designed and implemented a **hybrid search architecture** leveraging **LlamaIndex** for query generation and **Weaviate** vector database for semantic and keyword-based search, balancing relevance and specificity with an 80% vector-based and 20% keyword-based hybrid search.
- Integrated **Gemini 1.5 Pro** model for response generation and embeddings, optimizing context-aware responses with custom prompts and **contextual memory** for natural follow-up interactions.
- Built robust **indexing functionality** for dynamic knowledge base updates, supporting efficient data addition, removal, and retrieval from diverse sources.

Therapy Progress Tracking System for Psychotherapists

- Developed a prototype for tracking client therapy progress by analyzing session data to assess changes in symptom frequency and severity as indicators of positive or negative progress.
- Designed and implemented a user interface enabling therapists to submit session data for automated progress estimation.
- Utilized **LlamaIndex** with **Weaviate Vector Database** for hybrid search and **ReAct Agent** for reasoning and comparison of progress across client sessions.
- Delivered a proof-of-concept tool to support therapists in making data-driven evaluations of therapy effectiveness.

Amazon Recommender System link

- A Movie Recommender System using Item-based Collaborative Filtering
- Data description: 10 Million ratings from 480189 users for 17770 movies.

Quora Duplicate Question Prediction link

- Identify which question asked on Quora are duplicates of question that have already been answered. The system will help users to instantly get answers, if the question has been answered before.

Credit Card Fraud Prediction

- Did data wrangling & analysis, on large amount of highly imbalanced credit card transaction data.
- Solved the classification problem by creating a Majority vote classifier class which combined XGBoost & LightGBM model to get AUC of 0.93.

Customer Attrition Model link

- The goal of this use case was to predict possible default customers based on the customers historical data.
- Built customer attrition model using Machine Learning algorithms Logistic Regression, to create cost effective marketing campaigns focusing on valued, high-risk customer.

Stock Sentiment Analysis link

- Generated investing insight's by applying sentiment analysis on financial news headlines from finviz.com
- Tools used: BeautifulSoup for Web Scraping, Sentiment Intensity Analyzer from NLTK for sentiment score & matplotlib for data visualization of sentiment's for different stocks on daily basis.

Web Scraping Projects

- Created Web Scraper using Scrapy, for Real Estate website zillow.com, E-commerce website amazon.com which extract entire product details based on search query, tripadvisor.com & airbnb.com to extract restaurant details based on the search results.
- Created web scraper app using flask and BeautifulSoup, for UK Real Estate website RightMove which lets user select city and download the scraped data. link
- Created a Web Scraper App using Scrapy, ScrapyRT & flask for a gaming website steam.com. link

EDUCATION & CERTIFICATIONS

Bachelor of Commerce Management & Marketing University of Mumbai (2015)

Practical Multi AI Agents and Advanced Use Cases with crewAI (2025) link

Create agentic AI solutions with Azure AI Foundry (2025) link
Architect AI-Powered Applications (2025) link
Generative AI Databricks (2025) link
Full Stack Data Science With 1 Year Internship iNeuron.ai (2021-2022) link
Applied Data Science Specialization IBM (2019) link
Applied Machine Learning in Python University of Michigan (2019) link
Machine Learning with PySpark Data Camp (2019) link
Object Detection with Amazon Sage maker Cousera Project Network (2020) link
Excel to MySQL Analytic Technique for Business Duke University (2019) link
Business Analytics Specialization University of Pennsylvania (2019) link
Excel Skills for Business Specialization Macquarie University (2019) link
Advanced Business Strategy University of Virginia (2019) link

Brand & Product Management IE Business School (2019) link

SOFT SKILLS

- Problem Solving
- Curiosity
- Communication
- Creative Thinking
- Adaptability
- Empathy
- Critical Thinking
- Business Acumen
- Continuous Learner