Rahul Pandey
Data Scientist
Well versed in Data Science, Artificial Intelligence, Gen-AI.
Contact: +91 9473961497 | **Email:** pyrahulpndt@gmail.com
Social: Website 💼 | LinkedIn 🔗 | GitHub 🐙

Professional Summary

Data Scientist with around 4 years of working experience on Machine Learning, Deep Learning, Advance NLP, Generative-AI, Agentic-AI, LLM, Pytorch, TensorFlow, AWS, Sagemaker, Azure, MLOps. Winner of the Nvidia AI NIM Hackathon at the Nvidia AI Summit 2024, showcasing innovative problem-solving and technical expertise.

Work Experience

- **Accenture**
  AI/ML Computational Science Analyst   May, 2024 – Present
- **TO THE NEW**
  Software Engineer - Data Scientist   September, 2021 – April, 2024

DevOps Self-Healing AI System	
Product Overview	**Agentic DevOps automation framework** that enables autonomous incident detection, root cause analysis (RCA), and intelligent remediation through a layered, multi-agent architecture.
Responsibilities	• Developed **LLM-driven incident classifiers** to parse real-time logs and metrics from Splunk and Prometheus, enabling intelligent alert generation. • Engineered **graph-based RCA agents** using service dependency mapping and historical anomaly trends for proactive fault prediction. • Built **multi-agent orchestration** pipelines using LangGraph and Google ADK to coordinate alert parsing, severity analysis, and remediation via Ansible. • Implemented **human-in-the-loop (HITL) workflows** with Slack and ServiceNow integrations to validate and approve sensitive infrastructure changes. • Designed a **Retrieval-Augmented Generation (RAG) system** that leverages internal knowledge bases and FAISS-vectorized incident history to suggest resolution steps based on past occurrences. • **Created a system-level brainmap** capturing end-to-end agent workflows, signal routing, inter-agent dependencies, and HITL checkpoints to guide scalable implementation.
Avatar Based Conversational Advisor	
Product Overview	Development of an Avatar based Advisor with real-time response to queries
Responsibilities	• Developed a 95% accurate avatar-based chatbot using Nvidia Omniverse, ASR, and TTS for real-time responses. • Enhanced conversation flow by 30% through RAG and MemGPT integration. • Deployed Nvidia ACE, RIVA ASR, and TTS for lifelike voice outputs, increasing user satisfaction by 30%. • Improved security and conversation accuracy using Nvidia guardrails, reducing error rates by 15%. • Leveraged Nvidia NeMo and NIM for fine-tuning Llama, achieving 98% language generation accuracy for targeted cases. • Optimized VectorDB for 10x faster data retrieval in dynamic chats. • Reduced latency by 20% using FastAPI for backend communication. • Dockerized and deployed on Azure App Services, boosting scalability by 40%.
Automatic Filling of Applications and SmartOffice	
Product Overview	The objective of the project was to leverage AI for automatically filling data into PDFs for 150+ insurance carriers, ensuring maximum accuracy in filling the form fields.
Responsibilities	• Built AI solution for auto-filling PDFs for 150+ insurance carriers with 95% accuracy. • Deployed custom DistilBert model for label prediction, achieving 98% accuracy. • Applied rule-based Section Mapping to improve data accuracy in PDFs. • Mapped Webform data to PDF fields with 95% accuracy. • Integrated AWS Sagemaker for real-time deployment and auto-retraining of models. • Dockerized and deployed the solution on ECS for scalable operations.

Video Ready Revamp	
Product Overview	Video ready is an internal product of To The New for OTT platforms.
Responsibilities	- Worked on recommendation engine building for VideoReady tool. - Utilized multiple algorithms like LightFM, SVD, TF-IDF, Scikit-surprise for building content based, collaborative and hybrid filtering models. - Built Dashboards for accuracy Multiple metrics utilization for evaluation of models. - Curation of data from multiple sources and feature engineering for input to models. - Providing on-demand multiple rails.
Tata Play OTT Support POD	
Product Overview	Tata Play is an internal account at TTN, aimed at resolving user queries through a support system built on Freshservice, where agents address tickets as they arise.
Responsibilities	- Custom Bert model and Few-shot learning for Classification of Tickets achieving accuracy of 95%. - Creation of generic script irrespective of data changes, containerized the whole solution using docker and FastAPI.

Technical Skills

Generative-AI	Retrieval Augmented Generation (RAG), LangChain, LlamaIndex, LangGraph Fine-Tuning, LORA, QLORA, Parameter Efficient Training(PEFT), SFT
LLM Models	Llama, Open-AI, ChatGPT, Gemini, Mistral AI, LLM Models, Hugging-Face
Agentic-AI	LangGraph, Autogen, Google ADK, PydanticAI, CrewAI
Machine learning	Regression, Classification, Clustering, Ensemble learning, XGBoost, Random forest, Data Modelling, Forecasting, Predictive Analysis, Statistical Analysis, Probability Theory, Hypothesis Testing, Anova Testing
Deep Learning	Natural language processing, Recurrent Neural Network, LSTM, Transformers, BERT, Convolutional Neural Network, Image Processing, Sentiment Analysis, Active Speech Recognition (ASR), Text to Speech (TTS),
Nvidia Tech	Omniverse, Meta-Human, Avatar, NIMs, Nemo, RIVA, ACE Agent, Guardrails
ML & DL Tools	Numpy, Tensorflow, Pytorch, Scikit-learn, Pandas, PySpark, NLTK, MLRun
IDE/ Build Tools	Git, Docker, CI/CD, Kubernetes, Jupyter Notebook, Jupyter lab, VS-Code
Cloud Services	AWS, Sagemaker, Databricks, Azure, GCP
Database	Vector DB, GraphDB, Milvus, Qdrant, Pinecone, MySQL, MongoDB
Programming Languages/Frameworks	Python, FastAPI, Flask, RestAPI

Academic Qualification

- **Bachelor's Degree (B.Tech. – CSE)**
  IMS ENGINEERING COLLEGE   August, 2017 – July, 2021

Achievements & Certification

- **Winner** of Nvidia AI NIM Hackathon at Nvidia AI Summit 2024.
- Databricks Certified Machine Learning Associate
- Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning
- Convolutional Neural Networks in TensorFlow
- Statistics & Mathematics for Data Science & Data Analytics
- NLP-Natural Language Processing with Python