JEETEN KAPOOR JAIN

jeetenj2411@gmail.com | +1 (782) 882-1431 | linkedin.com/in/jeeten-jain/ | Halifax, NS, Canada

EDUCATION

Saint Mary's University, Halifax, Canada December 2024
Master of Science in Computing and Data Analytics
Relevant Courses: Natural Language Processing, Data & Text Mining, Statistics, Database Management, Data Visualization, DevOps

BITS Pilani, Pilani Campus, India July 2023
Integrated Master in Physics with Bachelor of Engineering in Mechanical Engineering
Relevant Courses: Data Structures & Algorithms, Object Oriented Programming, Linear Algebra, Probability & Statistics, Calculus

TECHNICAL SKILLS

- **LLMs & Applied NLP:** Fine-tuning LLMs (Hugging Face Transformers, PEFT with LoRA, QLoRA on Mistral, Phi-3.5), Retrieval-Augmented Generation (LangChain, LlamaIndex), Prompt Engineering (zero-shot, few-shot, CoT), Function/Tool Calling, structured text generation (Outlines), NLP pipelines (named entity recognition, summarization, sentiment analysis)
- **Model Development & Experimentation:** Classical Machine Learning (scikit-learn, XGBoost, CatBoost), Experiment Tracking (MLflow, Weights & Biases), Deep learning (PyTorch), hyperparameter tuning (Optuna)
- **LLM Application Engineering:** LLM app development (FastAPI, Streamlit, Gradio), LLM APIs (OpenAI GPT-4, Claude, Gemini), Vector Databases (Chroma, Pinecone; OpenAI embeddings), AWS services (SageMaker, Lambda, S3, RDS, Elasticsearch)
- **Data Processing & Visualization:** ETL pipelines, document parsing (PyMuPDF, AWS Textract), data wrangling (Pandas, NumPy), web scraping (Scrapy, BeautifulSoup), EDA and visualization (Plotly, Matplotlib, Seaborn)
- **Programming & Dev Tools:** Python, SQL, Bash, JavaScript/TypeScrip, Git/GitHub, Jupyter, Docker, Postman

PROFESSIONAL EXPERIENCE

SiftMed | St. John's, NL, Canada March 2024 – January 2025
Machine Learning Engineer

- Engineered **author identification system**, first using **Phi-3.5 LLM** for structured entity extraction via **prompt engineering**, then applying **CatBoost** classifier with feature engineering on extracted attributes to achieve **96.2%** accuracy in production.
- Deployed Phi-3.5 LLM on **AWS SageMaker** using **Text Generation Inference (TGI)** backend with custom integration of the **outlines library** for enforcing **structured JSON** responses.
- Led development of **document classification system** achieving **50% improvement** in accuracy over previous production model by implementing **LLMLingua**-based token compression (**20x reduction**) and **RoBERTa** for source/content classification.
- Reduced computational overhead by **75%** in document classification by identifying **4 key pages** critical for analysis. Deployed an optimized pipeline with a compression endpoint on **AWS SageMaker**.
- Developed **ETL pipeline** migrating **2M+ medical records** from **RDS & Elasticsearch** to **AWS Athena** using Python, enabling simplified company-wide data access.
- Built a scalable **document summarization pipeline** recursively using **LangChain MapReduce, OpenChat 3.5**, and AWS Textract to reduce document length by up to 90% while preserving essential content.

CEERI Pilani | Pilani, Rajasthan, India January 2023 – September 2023
Research Assistant

- **Fine-tuned Mask R-CNN** on MetaGraspNet dataset (**mAP = 93.58%**) using **Detectron2** with **Albumentations** for advanced **image augmentations**; designed a **graph-based algorithm** leveraging edge and depth data to infer occlusions and generate optimal grasp sequences among detected objects. *(Best Paper, ICMLDE 2024 – peer-reviewed)*
- Engineered a miniaturized **VGG-16 CNN** achieving **97% accuracy** on Φ-Net dataset in **semantic segmentation** of structural damage (e.g., cracks, spalling) using **transfer learning**. *(Paper, CSCT 2022 – peer-reviewed)*

PwC US Advisory - TMT |Mumbai, India July 2022 – December 2022
Data Scientist Associate

- Implemented customer segmentation using **PySpark** and **K-means clustering** on **PCA**-reduced variables (90% variance retained), identifying key profiles to inform a $10M migration strategy.
- Migrated **entity resolution** system to **AWS Lambda**, reducing monthly infrastructure costs by around 20% through efficient serverless design for a dataset of 1M+ records.

PROJECTS

Transformer-Based Sequence Reversal Model

- Implemented core transformer architecture components, including **token embeddings, positional encodings, multihead self-attention,** and **feed-forward networks**, creating a custom **encoder-decoder model** using **PyTorch**.
- Engineered a synthetic dataset for integer sequence reversal, **fine-tuned the transformer encoder** with an additional linear layer, and **optimized hyperparameters** to achieve a **96.8% token-wise accuracy** on evaluation.