Dev Garg

+1(979)326-3907 | devgargd7@gmail.com | devgargd7.com | /devgargd7 | /devgargd7

Education

Texas A&M University Jan 2024 - May 2025
*Master of Computer Science, (GPA: **4.0**/4.0)* *College Station, TX*
- Courses: Large Language Models, Software Engineering, Deep Learing, Info Storage and Retrieval

Indian Institute of Technology, BHU 2017 - 2021
Bachelor of Technology - Electronics Engineering (GPA: 8.78/10.0) *Varanasi, India*

Technical Skills

Languages: Python, C++, Scala, Java, Javascript, SQL
Frameworks: Pytorch, Tensorflow, React, Spring Boot, Apache Spark, CUDA, Scikit-learn, Transformers, Pandas
Tools/Platforms: AWS, Azure Datalake, Docker, Kubernetes, Git, GCP
Interests and Experience: Machine Learning, LLMs, Recommender Systems, Scalable Data Systems
Certifications: Google Cloud Professional MLE , Machine learning Spec., MLOps Spec., GenAI with LLMs

Experience

Texas A&M Engineering Experiment Station (TEES) Nov 2024 – Present
Student Machine Learning Engineer *College Station, TX*
- Architected an AI-driven e-learning platform using **AWS (DynamoDB, EC2, S3), React/NextJS, OpenAI APIs,** and **Pinecone**, delivering personalized learning to 400+ students via real-time ML-powered features.
- Developed RAG-based TA chatbot and automated question generation pipelines with **LangChain, vector stores**, and **OpenAI Whisper/GPT-4o**, improving question relevance by **25%**.
- Engineered scalable data and AI pipelines using **AWS**, integrating transcript extraction and performance analytics

Societe Generale GSC June 2021 - Dec 2023
Data Engineer *Bengaluru, India*
- Led development of scalable data processing systems for credit risk exposure analysis, leveraging **Big Data technologies** to enable high-volume data ingestion and transformation for downstream analytics.
- Accelerated system validation by 50% through automated regression testing and reduced compute costs by **€20,000 per quarter** by designing auto-scaling data pipelines, enhancing efficiency for real-time risk monitoring. Mentored junior engineers and integrated critical alerting mechanisms to ensure robust system performance.
- Built and optimized distributed data workflows using **Java, Spring Boot, Apache Spark, Scala, Kafka, ElasticSearch, SQL,** and **Azure Datalake**, supporting analytics capabilities for risk management applications.

Societe Generale GSC May 2020 - June 2020
Data Science Intern *Remote*
- Designed and implemented an **ML-driven incident resolution recommendation system**, leveraging natural language processing to reduce operational risks and improve response times by **38%**. Collaborated with cross-functional teams to deliver a high-quality MVP under aggressive timelines.
- Enhanced data processing and feature extraction pipelines using **Python, Scikit-learn, NLTK, Pandas, spaCy**, and **NetworkX**, enabling efficient analysis of unstructured data and actionable insights for incident management.

Publications

Rithik Kapoor, ***Dev Garg****, Ruihong Huang.* PaperFormer: A Citation-Graph Enhanced Language Model for Scientific Applications. [Under Review at **Association for Computational Linguistics (ACL 2025)**]

Projects

Attention-based Model Architecture for Citation Graph
- Co-developed a citation-aware LLM based on **LLaMA** 3.2-1B, integrating full-text citation contexts via **LoRA fine-tuning** and custom weights, achieving a **51% perplexity** reduction in causal language modeling and **SOTA ROUGE-1 (47.85)** in paper summarization on the SSN dataset.
- Engineered a novel dataset with millions of citation relationships enabling citation-aware review generation with optimized embedding pipelines to enhance context integration, supporting scalable scientific NLP applications.

News Aggregation and Recommendation System
- Developed a scalable event-driven microservices architecture for a real-time news aggregation and recommendation platform, using **FastAPI**, **Kubernetes**, and **Kafka** for high availability and fault tolerance.
- Designed an automated **MLOps** pipeline using **Kubeflow** & **MLflow**, orchestrating feature extraction, model training, deployment, and evaluation, reducing model deployment time with a model monitoring system to track drift and bias, and enabling automatic retraining triggers for degraded models.