Shivam Johri

+91-9108404965 • shivamjohri247@gmail.com • linkedin.com/in/shivam-johri

Lead Machine Learning Engineer with 9+ years of experience overall and 5+ years of experience leading ML teams and delivering high-impact solutions. Skilled in data engineering, model development, and deployment. Proficient in Python, C++, SQL, TensorFlow, PyTorch, and LLM techniques. Certified Google Cloud Associate Cloud Engineer. Adept in statistical modeling, dimensionality reduction, and software engineering principles. Proven ability to mentor junior engineers and drive innovation.

WORK EXPERIENCE

Epam Systems
02/2024 - Present

Senior Machine Learning Engineer • Full-time

- **Project 1: LLM-Based Query Rewriting for Indexing Services**
  Developed an innovative LLM-based query rewriting system to enhance the accuracy and relevance of search results.
  Utilized advanced language models to generate semantically similar and contextually relevant query variations.
  Integrated the query rewriting system into the company's indexing pipeline, significantly improving search engine performance.
  Deployed the system as a microservice on a cloud platform, ensuring scalability and reliability.
  Skills: Python, NLP, Machine Learning, Deep Learning, LLMs, Microservices, Cloud Platforms
- **Project 2: MLops Pipeline for Custom Named Entity Recognition**
  Designed and implemented an **end-to-end MLops pipeline** for a custom Named Entity Recognition (NER) model tailored to financial search queries.
  Trained a state-of-the-art NER model using deep learning techniques like BERT and Transformers.
  Developed a robust data pipeline to preprocess and clean financial text data.
  Deployed the model as a real-time inference service on Google Kubernetes Engine, ensuring high availability and scalability.
  Monitored model performance and re-trained as needed to maintain precision and recall.
  Skills: Python, NLP, Machine Learning, Deep Learning, TensorFlow/PyTorch, Kubernetes, MLops, Cloud Platforms (GCP)
- **Project 3: Recommendation Engine for Stock Name Prediction**
  Developed a robust recommendation engine using advanced NLP techniques to accurately predict stock names based on user search queries.
  Implemented a hybrid approach combining rule-based and machine learning models to improve accuracy and efficiency.
  Utilized techniques such as **TF-IDF, word embeddings, and cosine similarity** to measure semantic similarity between search queries and stock names.
  Deployed the model as a microservice using **Flask** and integrated it into the company's search platform as a feature flag.
  Skills: Python, NLP, Machine Learning, Flask, Microservices

- **Project 4: LLM-Based Query Expansion Service for Financial Queries**
  Developed an LLM-based query expansion service to provide better context for smaller queries in financial search.
  Utilized advanced language models to generate semantically related and contextually relevant query variations, including synonyms, related terms, and long-tail keywords.
  Integrated the query expansion service into the company's financial search platform, significantly improving search results.
  Deployed the service as a microservice on a cloud platform, ensuring scalability and reliability.
  Skills: Python, NLP, Machine Learning, Deep Learning, LLMs, Microservices, Cloud Platforms
  Key Metrics and Results:
  Increased search query coverage by 25%, leading to a significant improvement in search result relevance.
  Reduced query ambiguity by 30%, resulting in more accurate and precise search results.
  Improved user satisfaction by 15%, as measured by user surveys and feedback.

Accenture **09/2021 - 02/2024**

ML Engineering Senior Analyst • Full-time

- **Project 1: Personalized Fashion Recommendations with Recommendations AI**
  Developed and deployed 5 personalized recommendation engines using Google Cloud's Recommendations AI.
  Leveraged advanced machine learning techniques to analyze user behavior and product catalogs, resulting in a 15% increase in user engagement for fashion discovery.
  Optimized recommendation models for real-time performance and scalability.
  Skills: Python, GCP Vertex AI, Recommendations AI, Retail AI, Google Cloud Platform
- **Project 2: Automated Regulatory and Compliance Report Generation**
  Built a Python-based automation solution to streamline the generation of complex **regulatory and compliance reports**.
  Integrated with diverse data sources, including **databases**, **spreadsheets**, and **APIs**, to extract and transform the necessary information.
  Developed robust data processing pipelines to **clean, validate, and standardize data** from various sources.
  Automated report generation, formatting, and distribution, significantly reducing manual effort and improving efficiency.
  Ensured compliance with industry standards and regulatory requirements by incorporating relevant checks and validations.
  Implemented a user-friendly interface to allow for easy customization and configuration of reports**.**
  Skills: Python, Automation, Data Processing, Regulatory Compliance, GCP BigQuery

EDUCATION

BTECH in Electronics and Communication Engineering

APJ Abdul Kalam University • GPA: 6.24 Lucknow • 08/2011 - 06/2015

Shri Ramswaroop Memorial College of Engg and Management, Lucknow

CERTIFICATIONS

Associate Cloud Engineer 11/2023 - 11/2025
Google

SKILLS

- Algorithm Optimization
- Azure
- Big Data Technologies (Hadoop
- Cloud Computing (AWS
- Computer Vision
- Data Preprocessing
- Data Visualization
- Deep Learning
- Distributed Computing
- GCP)
- Git
- Keras
- Leadership and Team Management.
- MLOps
- Model Deployment
- Model Evaluation
- Natural Language Processing (NLP)
- Python
- PyTorch
- Scikit-learn
- Spark)
- SQL
- TensorFlow