# Shivam Johri

+91-9108404965 • shivamjohri247@gmail.com • shivam-johri
Shivamjohri247

## Summary

Results-driven Senior Machine Learning Engineer with 9.5+ years of experience, including 6+ years specializing in architecting and deploying end-to-end MLOps pipelines and delivering impactful AI/ML solutions. Expert in automating model lifecycles, leveraging deep learning, LLMs, and cloud-based technologies. Proven ability to translate complex business needs into scalable, high-performance solutions. **Google Cloud Associate Cloud Engineer certified.**

## Technical Expertise

**MLOps / Data Eng.**: Apache Airflow, MLflow, Kubernetes, Docker, Git/GitHub, CI/CD pipelines, Apache Spark, Apache Solr, ElasticSearch

**AI/ML Dev.**: Deep Learning (BERT, Transformers), LLMs, Recommendation Engines, NLP, TensorFlow, PyTorch, Keras, Scikit-learn, FastAPI, Huggingface

**LLM / Agentic AI**: LangChain, LangGraph, Semantic Kernel, Google Agent Builder (Vertex AI Agents), OpenAI Function Calling, Retrieval-Augmented Generation (RAG), Vector Databases (FAISS, Pinecone, ChromaDB)

**Cloud / Big Data**: Google Cloud Platform (Vertex AI, BigQuery, GKE), AWS SageMaker, Hadoop Ecosystem (Spark, Hive), Distributed Computing

**Programming**: Python, SQL, Go, Shell Scripting, Data Preprocessing, Visualization, Data Governance, Responsible AI

## Professional Experience

**Suzega** **India**
*Senior AI Engineer* *May 2025 – Present*

- Leading development of agentic AI systems capable of autonomous task execution, decision making, and workflow orchestration.
- Building scalable LLM-powered agents with memory, retrieval-augmented generation, and reasoning capabilities.
- Designing agent pipelines for real-world enterprise use cases across domains such as knowledge automation, customer support, and intelligent workflows.

**EPAM Systems** **Remote, Poland**
*Senior Machine Learning Engineer* *Feb 2024 – May 2025*

- Worked on automating end-to-end MLOps pipeline for financial **Named Entity Recognition (NER)** utilizing Transformers fine-tuning, significantly enhancing search relevancy and accuracy by 77% in search platform and deployed in production.
- Developed and deployed an LLM-powered query rewriting system for indexing services, driving substantial improvements in search relevancy and query processing.
- Worked on creation of Query classification pipeline using fine-tuning approach using BERT embeddings and XGBClassifier for search relevancy improvement and enhanced result ranking for search.

**Accenture** **India**
*ML Engineering Senior Analyst* *Sep 2021 – Feb 2024*

- Led the development of an LLM-based query expansion service for financial queries using Gemini models, achieving a 25% increase in query coverage and enhancing overall search effectiveness for a finance domain client.
- Developed high-accuracy signature detection models for fraud prevention in handwritten documents using advanced computer vision techniques like FasterRCNN and YoloV3.
- Designed and deployed 5 personalized fashion recommendation models on GCP's Recommendations AI, boosting user engagement by 15% for a retail client.
- Automated regulatory report generation processes, resulting in a 40% reduction in manual effort and improved compliance efficiency for Google partnered regulatory reporting clientele.

**Tata Consultancy Services** **Bangalore, India**
*Machine Learning Engineer* *Mar 2016 – Aug 2021*

- Developed an end-to-end pipeline for text detection and information extraction from ID cards using CTPN, automating critical data processing workflows using ETL pipelines.
- Built and deployed machine learning models for document classification and data extraction from complex PDF documents (using Tesseract OCR), reducing validation costs by 30%.
- Implemented a sophisticated email indexing and entity recognition system, enabling advanced data retrieval and analysis for clients.

## Education

**Shri Ramswaroop Memorial College of Engg. & Management (APJAKTU)** **Lucknow, India**
*B.Tech, Electronics and Communication Engineering* *2011 – 2015*
GPA: 6.24/10

## Certifications

**Google Cloud**: Associate Cloud Engineer (Valid till 11/2025)

## Projects

**LLM Fine-tuning Pipeline**: Baseline pipeline for fine-tuning SLMs on downstream tasks; extensible for custom needs.

**Emotion Classifier**: Open-source project using Huggingface Transformers for emotion detection in text. GitHub Link

## Volunteering

**Shri Ramswaroop Memorial College Fest "XERO-2K15"** **Lucknow**
*Organizer* *Feb 2015 – Mar 2015*
Led organizing team for two consecutive years; managed cultural + technical events with high participation.