Nehul Bhatnagar

+91-8949446740 | nbhatnagar3010@gmail.com | linkedin.com/in/nehulbhatnagar

Education

National Institute of Technology, Jalandhar — Jalandhar, India
B.Tech in Electronics and Communication Engineering — *2019 – 2023*

Work Experience

Machine Learning Engineer II — Nov. 2023 – Present
Revionics *— portfolio of* ***Goldman Sachs*** — *Bengaluru, India*
End-to-end owner of 5 production-grade ML, LLM, and agentic systems for retail price optimization

- Architected and deployed a **client-facing RAG assistant** over product help documentation using **Vector DBs**, letting clients self-serve answers, cutting how-to support tickets by **70%**.
- Designed and built an Agent-to-Agent (**A2A**) protocol service on **Google ADK**, exposing a multi-agent retail analytics and pricing assistant to external AI agents, with **task timeouts**, **cancellation propagation**, and **idempotent** per-agent token metering for usage-based billing.
- Led technical design of a **multi-agent** LLM pricing platform, building guardrails, **evaluations**, and workflow execution engines via **FastAPI** services.
- Proposed and built an **agent orchestrator**, lowering AI-tool spend by **75%** (**$90K** projected annual savings).
- Helped secure **$1.3M ARR** across 3 clients by building a product relationship engine: an **NLP** + numeric **Siamese network** with **FAISS** and **Leiden clustering**, reducing manual product grouping effort by **80%+**.
- Generated **$2M+** in new European revenue by developing VAT-aware **Python forecasting libraries**.
- Saved **$100K+ annually** in compute and cut wall time by **80%+** by building a high-performance pricing job.

Machine Learning Engineering Intern — May 2023 – Aug. 2023
Coinbase — *Remote, India*
Built LLM-powered social intelligence systems and scalable data ingestion pipelines

- Designed and deployed an end-to-end **Social Intelligence pipeline** using **LLM-powered NLP** to extract crypto market signals from high-volume social media streams in near real-time.
- Engineered fault-tolerant automated data pipelines using **Apache Airflow** to ingest and process **15K+ tweets/hour** for downstream model inference.
- Implemented narrative detection pipelines combining **BERTopic** and **LLM vector embeddings**, optimizing topic clustering accuracy on unstructured text.

Summer Analyst — Jun. 2022 – Jul. 2022
Goldman Sachs — *Bengaluru, India*
Engineered high-throughput distributed pipelines and optimized trade data processing

- Accelerated **Apache Kafka** processing **7x (14 hrs to <2 hrs)** by architecting a scalable multiprocessing pipeline handling **6M+ messages (60GB+)** per run.
- Unified 6 globally fragmented regression environments into a single pipeline, adopted by 6 international engineering teams.

Publications

Co-Author & Core ML Contributor — Feb. 2026
SocialPulse: Open-Source Subreddit Sensemaking Toolkit — *ICWSM 2026 / arXiv*

- Built the **NLP + GenAI modules** that extract structured analytics from unstructured subreddit data.

Skills and Expertise

Languages & Frameworks: Python, SQL, FastAPI, TensorFlow, C#, C++, Scikit-learn

Machine Learning & LLMs: Multi-Agent Systems (Google ADK and A2A), RAG, LLM Evaluation, Embeddings and Vector Search (FAISS, Pinecone), Siamese Networks, Clustering, NLP, Fine-tuning

Cloud & Infrastructure: GCP, Docker, Kubernetes (EKS/GKE), Microservices, REST APIs, CI/CD, Observability

Data & Distributed Systems: Apache Airflow, Apache Kafka, BigQuery, Databricks, Snowflake, Spark, MSSQL, MongoDB