

Harmanan Gurvinder Kohli

✉ hkohli40@gmail.com | 📞 +91 98195 27168 | github.com/Harmanankohli | linkedin.com/in/harmanankohli/

Professional Summary

LLM/AI Engineer with experience building agentic workflows, RAG systems, Text-to-SQL tools, embeddings pipelines, and multimodal LLM applications, deployed in cloud environments. Strong background across ML, NLP, Explainable AI, Deep Learning, and enterprise analytics, with additional exposure to XAI research and open-source ML contributions.

Skills

Programming: Python, SQL

GenAI & LLM Systems: LangChain, LangGraph, Bedrock Agents, RAG, VectorDB, Embeddings, Tool Calling, Prompting, Docling, OCR

ML & DL: scikit-learn, TensorFlow, Keras, XGBoost, Pandas, Polars, Numpy, SHAP/LIME, EBM

Cloud & Deployment: AWS (Lambda, Bedrock, DynamoDB, CloudWatch), Azure, GCP (Cloud Run), Streamlit

Analytics & Tools: Tableau, Selenium, SAP, Excel

Work Experience

Capgemini, Mumbai

Aug 2025 - Present

Consultant (GenAI)

- Built a financial document intelligence system for Master Facility Agreements (MFA) and Sanction Letters using Docling extraction + semantic chunking (LangChain) to enable long-context retrieval for 100+ page documents.
- Designed a 2x2 VectorDB retrieval architecture (Financial Domain: Working Capital vs Long-Term Finance x Document Type: MFA vs Sanction Letter) to isolate semantic neighborhoods and reduce embedding interference.
- Implemented agentic orchestration using LangGraph to route multi-step workflows classification → retrieval → synthesis) with deterministic control flow and error handling.
- Integrated Text-to-SQL using Bedrock Agents to support hybrid structured + unstructured reasoning (contract clauses + financial tables).
- Deployed on AWS (Lambda + Bedrock LLMs) with latency, cost, and output formatting optimizations.
- Implemented MCP server tools in Databricks and deployed as an internal app, enabling secure tool integration for LLM-driven workflows.
- Developed an AI-powered accelerator that analyzes product images and descriptions to deliver pricing intelligence, competitor benchmarking, product discovery, go-to-market recommendations, IP risk assessment, and AI-generated marketing content through a multi-agent architecture.

Sterlite Technologies, Mumbai

Oct 2024 - Apr 2025

Data Scientist

- Designed a safety stock optimization model (Z-score, AMC, lead time, SD), reducing inventory holding costs and improving reorder accuracy.
- Applied Counterfactual Explanations (XAI) to RFQ classifications, identifying actionable parameter shifts that improved deal conversion rates.
- Built a proof-of-concept with Gemini Pro Vision to extract structured data from scanned vendor invoices, streamlining accounts payable workflows.

Tata Consultancy Services, Mumbai

Jan 2020 - Jan 2024

Systems Engineer

- Delivered analytics and planning solutions within SAP ecosystem, enhancing forecasting, reporting, and decision workflows.
- Reduced processing time from days → minutes by optimizing data pipelines in SAP BPC.
- Integrated spatial + business + financial data across SAP S4HANA, GIS, and Real Estate modules for geospatial analytics.

Research Experience

Master’s Dissertation: Explainable AI for Predicting Malware Infection

Nov 2023 - Apr 2024

Graduate Researcher

- Trained ML model on Microsoft Malware dataset (8.9M devices) achieving ~66.7% accuracy and ~66.4% recall, aiding proactive mitigation of data breaches and financial risks.
- Applied EBM, SHAP, LIME for interpretability, producing local + global explanations with uncertainty for regulatory + trust alignment (GDPR-compatible design).
- Engineered features from 4.1GB (83-field) dataset, optimizing memory usage by ~50% via columnar transformations in Pandas/Polars.

Launchpad.ai (via Fellowship.ai), Remote

Jan 2024 - Mar 2024

Data Science Fellow

- Built multimodal price discovery agent using Gemini Pro Vision, achieving ~97% accuracy on internal test set.
- Designed scalable RAG pipeline + VectorDB + tool-calling agent, reducing manual lookup effort 50%.
- Implemented agent that queries VectorDB for known items and dynamically falls back to search APIs for unknown products.
- Developed Selenium scrapers for product data acquisition and validation.

OPEN-SOURCE & VOLUNTEERING

Scikit-Learn Contributor

- Added examples for `f_regression()` and `silhouette_score()` to improve documentation usability.
- Contributions PR: `f_regression()` - #28104 and `silhouette_score()` - #28125

Education

BITS Pilani WILP
M.Tech. in Data Science and Engineering

Apr 2022 - May 2024
CGPA: 8.14/10

SRM University, Haryana
B.Tech. in Computer Science and Engineering

Jul 2014 - May 2018
CGPA: 8.14/10

CERTIFICATIONS & AWARDS

- | | |
|--|-------------|
| • Microsoft — Azure AI Engineer Associate | June 2025 |
| • Acknowledged Contributor — The Hundred-Page Language Models Book | 2025 |
| • TCS — Best Team (2x), On-the-Spot (5x), Special Initiative Award | 2021 - 2023 |
| • Semi-Finalist — TechGig Code Gladiators | 2018 |