

Simran Kharbanda

simrank3574@gmail.com | linkedin.com/in/simran-kharbanda | github.com/Sim2200 | sim2200.github.io/Portfolio

EDUCATION

University of Maryland, College Park — *M.S. Applied Machine Learning*

Aug 2025 – Present

Manipal Institute of Technology — *B.Tech Electronics & Communication, Minor in Data Science*

Oct 2021 – May 2025

EXPERIENCE

Observe.AI — *Forward Deployed Engineer*

May 2026 – Present

- Build and deploy voice agents with Vapi, n8n, and OpenAI APIs for claims support and customer service: prompt workflows, tool calling, authentication, and real-time claim status retrieval via REST APIs and webhooks.
- Technical owner of a Companion AI agent build for a major enterprise client (New Hire and Tenured configurations) from development through go-live, including n8n workflows with custom JavaScript nodes for post-call nudges.
- Diagnose production issues across multi-agent systems (LLM reliability failures, telephony latency, data sync, fallback logic, session state) and built reusable evaluation and logging workflows for conversation quality.
- Write ClickHouse SQL for Superset dashboards on agent behavior and nudge effectiveness; authored migration docs for a platform-wide agent architecture overhaul across six live configurations.

Flipkart — *Data Intern*, Bangalore

Feb 2025 – Jun 2025

- Built end-to-end ML pipelines (PySpark, SQL, Pandas) over 120M+ e-commerce transactions, improving recommendation CTR by 14%.
- Developed LightGBM and Prophet demand-forecasting models (MAE < 5%) to optimize inventory across 12 warehouses.
- Deployed Dockerized FastAPI inference services (20–40 ms latency in staging); automated workflows with Airflow and PostgreSQL, cutting manual analytics by 90% and runtime by 60%.

Manipal Institute of Technology — *Research Assistant*

Jan 2024 – May 2025

- Designed TabTransformer, ClinicalBERT, ResNet-18, and 3D U-Net models to predict myeloma treatment response from multimodal clinical data (BTD, FCM, PET/CT, bone marrow).
- Identified IBV as a novel biomarker through SHAP attribution, improving AUC by 12%; accelerated GPU training with CUDA, PyTorch, Spark, and MLflow.

Ernst & Young — *Generative AI Intern*, Noida

Jun 2024 – Aug 2024

- Designed a RAG chatbot with Flan-T5 and Mistral-7B for Excel workflow automation over 10K+ corporate files.
- Engineered hybrid retrieval with LangChain and FAISS (<400 ms latency, 99.3% precision); A/B tested 400+ prompts and embedding setups, improving query accuracy by 25%.

Earlier internships

Apr 2024 – Jul 2024

- Cisco** (AICTE Cybersecurity): intrusion detection pipelines with Snort and TF-IDF, 93% F1 and 28% fewer false alerts. **Akzo Nobel** (IT): Isolation Forest/ARIMA anomaly detection, 12% fewer defects. **Hero MotoCorp** (Data Science): Random Forest/Prophet sales forecasting on 5M+ records, 18% better MAE.

PUBLICATION

Digital Sentiment and the Retail Crowd: How Finfluencers Shape IPO Valuations

2025

- Co-author, *Journal of Behavioral Finance* (Taylor & Francis), doi.org/10.1080/15427560.2025.2566736. Sentiment modeling (VADER, TextBlob, BERT) across 395 Indian IPOs; endorsements linked to 7.2% more underpricing, 6.8% higher initial returns ($p < 0.05$), and $10.9\times$ retail demand ($p < 0.01$).

PROJECTS

SLA-Aware Multi-Variant ML Serving Platform | *Python, FastAPI, ONNX, Kubernetes, AWS EKS*

2026

- Platform on AWS EKS routing requests between a high-accuracy and a quantized ONNX model; implemented the models, FastAPI services and SLA-aware router, cutting p95 latency from 30+ s to 5 s under load.

Scientific Research Analyzer | *LangChain, Hugging Face TGI, Dask, LangFuse*

2025

- Built multi-agent workflows for scientific literature synthesis with LangChain Graphs; tuned coordination with TGI, Dask, and LangFuse, improving retrieval performance by 24%.

TECHNICAL SKILLS

LLMs & Agents: OpenAI APIs, LangChain, Hugging Face, TGI, RAG, FAISS, Vapi, n8n, LangFuse, LLM evaluation

ML: PyTorch, TensorFlow, scikit-learn, LightGBM, XGBoost, Prophet, SHAP

Data & Infra: PySpark, Airflow, Kafka, MLflow, Docker, Kubernetes, FastAPI, AWS, PostgreSQL, ClickHouse, MongoDB

Languages: Python, SQL, C++, R, JavaScript, Go