

Saksham Parashar

Ghaziabad, Uttar Pradesh | P: +91 9953879866 | sakshamparashar12@gmail.com | [LinkedIn](#) | [GitHub](#)

AI Engineer with 2.5+ years of experience building LLM applications, RAG and agentic pipelines, and machine learning systems for enterprise clients, including an LLM-powered root-cause analysis tool that cut failure diagnosis from ~2 days to under 5 minutes. Skilled in LangChain, LangGraph, Azure OpenAI, and hybrid retrieval with CrossEncoder reranking, with applied ML experience in time-series forecasting and GPU-accelerated AI-RAN simulation using NVIDIA Sionna — backed by backend engineering in Spring Boot and AWS.

WORK EXPERIENCE

CELEBAL TECHNOLOGIES

Data Scientist

July 2026 – Present

- Build **GenAI (RAG, NLP) and classical ML solutions** for enterprise clients, covering predictive maintenance, anomaly detection, and time-series forecasting.
- Develop and evaluate **ML and GenAI pipelines** in Python on Databricks and AWS, integrating enterprise data sources including SAP.
- Integrate deployed models into client business workflows, with a focus on model performance and reliability.

TATA CONSULTANCY SERVICES

Software Engineer (Backend and Data Systems)

April 2025 – July 2026

- Architected an **LLM-powered root-cause analysis tool** that ingests bug telemetry and auto-classifies operational failures with **~80% accuracy**, cutting diagnosis time from **~2 days to under 5 minutes**.
- Delivered a **Human-in-the-Loop (HITL) GenAI pipeline** for deviation tracking that summarizes incident logs, automates compliance checks, and traces data lineage through a live-editing review interface.
- Proposed and built **chunked retrieval for AWS document pipelines**, cutting payload download time from **~1 minute to under 15 seconds** (75% faster).
- Developed and scaled **backend microservices** for enterprise shipment tracking, invoice management, and workflow automation, deployed through Jenkins CI/CD pipelines.

BYTIVE TECHNOLOGIES – NOIDA

AI / Data Scientist

March 2024 – April 2025

AI-RAN & Network Optimization

- Developed a **5G RAN handover simulator** modeling UE mobility, signal propagation (RSRP/SINR), and handover logic (A3 events, TTT, CIO) for 100–500 UEs across 1,000+ timesteps.
- Designed an **adaptive optimization loop** that tunes handover parameters from network KPIs (RLF rate, SINR, early/late/wrong handovers), reducing simulated call drops by **~13%**.
- Led development of a Radio Network Digital Twin to model network behavior and evaluate infrastructure scenarios, enabling data-driven planning that reduced projected CAPEX by ~3%.

Time-Series Demand Forecasting

- Built a **multi-model forecasting system** (AutoGluon, Temporal Fusion Transformer, GNN) on 15K+ records spanning 26 products over 794 days, improving accuracy by **~9%** over baseline.
- Engineered lag, rolling-window, seasonality, and graph features, and built a 6-fold benchmarking framework for model selection (best MAPE: **21.47%**).
- Converted forecasts into inventory decisions with Bayesian threshold optimization, running the end-to-end pipeline on Databricks and Microsoft Fabric through to BI dashboards.

GenAI & RAG

- Developed **RAG and agentic applications** with LangChain, Azure OpenAI, FAISS, and Pinecone for natural-language analytics and semantic search.
- Improved retrieval quality with **hybrid search (BM25 + dense vectors)**, parent-child chunking, CrossEncoder reranking, query routing, and tool calling.

PROJECTS

CDC Metadata Search & Discovery Platform | Azure OpenAI, LangChain, FAISS, Pinecone, Streamlit

- Delivered a **RAG platform** with Sentence Transformer embeddings that lets users search CDC metadata repositories in natural language, reducing search time by **60–80%**.
- Re-architected the LLM integration with **streaming responses** to cut perceived latency, and shipped a Streamlit interface that returns answers grounded in retrieved metadata.

AI-Powered Pull Request Review System | Azure Foundry, Prompt Engineering, Azure DevOps

- Built an **LLM code review tool** that parses PR diffs, builds contextual prompts, and generates feedback on correctness, readability, and best practices.
- Designed a **multi-step evaluation workflow** with cost optimizations that reduced API calls by **57%**, and integrated it into the PR workflow for automatic reviews.

5G Digital Twin | Python, NVIDIA Sionna RT, AWS L40S GPU

- Built a **10-stage, cache-backed propagation pipeline** computing per-link radio channels (gains, delays, angles); traced **11,163 links in under 5 minutes** on a 5,847-object San Francisco scene.
- Designed a **trace-once, query-many cache**: applying antenna patterns, tilt, and azimuth after tracing matched full re-traces to within **0.001 dB**, turning each antenna what-if from a 5-minute GPU re-trace into ~1 second.
- Diagnosed **7 silent ray-tracer failure modes**, including duplicate paths inflating received power by up to **43 dB** and a buffer default discarding **31% of valid paths**; the fix matches closed-form physics to 0.00005 dB.
- Classified all **85 configuration settings** by re-computation cost and wrote **330 automated checks** (0 failures), with planning-time guards that reject invalid runs in 0.2 s instead of after a 5-minute GPU trace.
- Built an evaluation framework that converts **ray-traced channel impulse responses into OFDM link simulations**, reporting per-user SINR, MCS, BLER, and throughput against 3GPP TDL baselines.
- Verified the link chain against theory (**simulated BER 0.0786 vs. 0.0787 theoretical**) and showed that ray-traced channels expose NLOS coverage outages that statistical models cannot represent, at **96% GPU utilization** on an NVIDIA L40S.
- Wrote a **dependency-free pre-flight validator** for MIMO sweeps (up to 64T64R, RZF, LMMSE/K-Best) that blocks rank, pilot, cyclic-prefix, and GPU-memory errors before GPU time is spent; caught **all 7 seeded bugs**.

SKILLS

- **Languages:** Python, SQL
- **GenAI & LLMs:** LangChain, LangGraph, LlamaIndex, AutoGen, CrewAI, Azure OpenAI, GPT-4, LLaMA, Ollama, Whisper, Prompt Engineering, Guardrails & Output Evaluation
- **Retrieval:** FAISS, Pinecone, Sentence Transformers, Hybrid Search (BM25 + dense), CrossEncoder
- **Machine Learning:** PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, AutoGluon, LSTM, Prophet
- **AI-RAN & Wireless:** NVIDIA Sionna (RT, PHY), Ray Tracing, 5G NR PHY (OFDM, MIMO, LDPC), Handover Optimization (A3, TTT, CIO), RSRP/SINR Analysis
- **Data & MLOps:** Databricks, Microsoft Fabric, PySpark, Pandas, NumPy, MLflow, ETL, Azure Lakehouse
- **Backend & Cloud:** Spring Boot, FastAPI, Flask, Django, REST APIs, AWS, Azure, Docker, Jenkins, Git
- **AI Coding Tools:** Claude Code, GitHub Copilot, Codex

CERTIFICATIONS & TRAINING

- AWS Certified AI Practitioner
- AWS Certified Machine Learning Engineer – Associate

EDUCATION

Bachelors in Technology, Computer Science and Engineering (Data Science)
CGPA: 8.2

2020-2024