HAMZA MUNIR

SENIOR AI ENGINEER | LLM & AGENTIC SYSTEMS

Lahore, Pakistan (Open to Relocation & Remote) | +92 334 967 0802

hammuneer@gmail.com | linkedin.com/in/hammuneer123 | github.com/hammuneer | kaggle.com/hammuneer

PROFESSIONAL SUMMARY

Senior AI Engineer with 7+ years turning AI systems into production across healthcare, compliance, marketing, and finance, with a deep focus on LLM applications, retrieval-augmented generation, multi-agent orchestration, and automated evaluation, underpinned by production data engineering and backend development. Rather than starting from the technology, begins with the business problem, working directly with stakeholders to shape solutions that balance technical feasibility, measurable business value, and user experience. Has led engineering teams and owned client-facing delivery across four continents.

TECHNICAL SKILLS

Languages	Python · SQL · Scala · Bash
LLM & Agents	LangGraph · LangChain · RAG · Multi-Agent Orchestration · OpenAI · Anthropic · OpenRouter · Groq · Prompt Engineering & Caching · DeepEval · Langfuse · n8n · Dify
Frameworks & Libraries	PyTorch · TensorFlow · Keras · Hugging Face · scikit-learn · PyCaret · Matplotlib · Seaborn · Selenium · BeautifulSoup · Gradio · Streamlit
Data & Storage	PostgreSQL/pgvector · MongoDB · Azure Cosmos DB · MySQL · Qdrant · FAISS · ChromaDB · Neo4j · Apache Spark · AWS Glue
Backend & MLOps	FastAPI · Docker · Kubernetes · MLflow · vLLM · Ollama · Helm · Jenkins · Ansible · Terraform · Git/GitHub Actions · pytest · k6 Load Testing · CI/CD
Cloud	AWS (EMR · Glue · S3 · EC2 · SageMaker) · GCP (GCS · Cloud Run · Vertex AI) · Azure ML Studio
Tools & Platforms	Cursor AI · Claude Code · JIRA · Microsoft 365 · Google Analytics · Facebook Ads Manager

PROFESSIONAL EXPERIENCE

QPharma, Inc. — **Nov 2025 – Present**

Senior Data Science Developer — New Jersey, USA (Remote)

- Built a natural-language Question Answering system framed as a **text-to-MongoDB** problem, orchestrating multiple Anthropic models across the pipeline to balance answer quality, latency, and cost; it enables stakeholders to ask analytical questions, generate daily and monthly reports, and get quick answers on liaison progress and doctor-level activity.
- Architected a key opinion leader intelligence platform that synthesizes 24 months of press releases, publications, and social activity per clinician from MongoDB using **GPT-5.2,** cutting profile research time for therapeutic liaisons from **48 hrs** to **5 mins**.
- Re-engineered a legacy, file-based batch pipeline, orchestrating MongoDB reads and multi-service calls, into a fully automated, fault-tolerant Jenkins job with checkpointed recovery and completion alerts, built under **spec-driven development** with **95% test coverage**, cutting a multi-day, dependency-blocked process down to a single automated run.
- Reduced LLM inference cost by **50%** through prompt caching, cross-provider model routing, and **Langfuse** tracing to pinpoint cost drivers, with no measurable loss in answer quality.
- Designed AWS Glue ETL pipelines and optimized Apache Spark (Scala) jobs on EMR over S3 data lakes, processing **50 GB/day** of healthcare data while enforcing schema and HIPAA-aligned validation.
- Documented system and integration architecture for each major feature using the **C4 model** (Context, Container, Component, Code), aligning engineering and stakeholders on scope, dependencies, and design decisions before implementation.

[RE]DESIGN Group — **Nov 2023 – Aug 2025**

Senior Data Scientist & LLM Engineer — California, USA (Remote)

- Led a **4-engineer** team building Bookstore.ai, a hybrid recommendation engine combining content-based and collaborative filtering over PostgreSQL and Qdrant, increasing click-through on recommended titles by **60%**.
- Designed the ingestion layer: an Alembic-managed schema and ETL pipeline normalizing 10M+ book records from Amazon, Ingram, and Goodreads, automatically resolving **50%** of duplicate and conflicting metadata.
- Shipped FastAPI services at **95% test** coverage, validated to 50 RPS under k6 load testing, and deployed via Docker to private cloud with consistent low-latency response times.
- Built **Nora AI**, a conversational assistant on self-hosted Llama 3 (8B) via Ollama on A6000 GPUs with Qdrant retrieval, serving goal-based book recommendations at near-zero marginal inference cost versus **$10000/mo** on commercial APIs.
- Built Redesign RAG, unifying Slack and Jira into a single RAG knowledge assistant and cutting internal time-to-answer on operational questions from **2 days** to a **few seconds**.

Odyssey Analytics **Oct 2022 – Oct 2023**
Senior Data Scientist Lahore, Pakistan

- Directed a Saudi Hajj transport optimization program, processing **5 GB/day** of vehicle telemetry through anomaly detection and traffic-trend models to improve fleet operational efficiency by **50%.**
- Owned AI delivery for **3** concurrent client engagements across MENA, Russia, Australia, and the USA, converting ambiguous business goals into scoped, shipped systems.
- Built an Arabic-language extractive Q&A system over religious texts using LangChain RAG and open-source LLMs, reaching approximately **90%** answer accuracy despite sparse Arabic NLP tooling.
- Delivered a natural-language query system over a commodities knowledge graph, letting traders analyze **20 years** of historical pricing conversationally instead of via analyst tickets.
- Modeled agricultural supplier churn driven by price sensitivity (XGBoost, SVM, logistic regression) at **75%** AUC, enabling retention campaigns that recovered **20%** of at-risk accounts.

Soliton Technologies (Persivia) **Apr 2021 – Oct 2022**
Data Scientist Lahore, Pakistan

- Built ML and deep learning models to forecast medical claim costs (80% accuracy for small-bill customers, 65% for high-bill customers) and 30-day heart-failure readmissions **(70% AUC)**, flagging high-risk patients for proactive intervention.
- Developed clinical-note segmentation and disease-recognition NLP (spaCy, NER, rule-based extraction) across 500K unstructured patient records and labeled diseases with **ICD-10 and HCC Codes** using approximate matching techniques.
- Engineered ingestion pipelines unifying JSON, SQL, and unstructured clinical repositories into one analytics layer; reported outcomes and KPIs directly to the CEO and CTO.

DICE Analytics **Feb 2020 – Dec 2022**
Instructor, Data Science & Deep Learning (Part-Time) Islamabad, Pakistan (Remote)

- Trained **300+** professionals in Python, NLP, and deep learning (CNNs, RNNs, GANs), including custom corporate programs for Zong CMPak; designed all assessments, capstones, and evaluation rubrics.

Techlets Pvt. Ltd. (WALEE) **Nov 2019 – Oct 2020**
Data Associate Islamabad, Pakistan

- Deployed a CNN-based ID card classification model for profile verification, alongside a separate sentiment analysis pipeline, across **50,000+** social media influencer profiles.
- Built distributed crawlers across social media platforms (Facebook, Instagram, TikTok, Twitter), search engines, and e-commerce sources, and owned the MongoDB data quality, cleansing, and aggregation pipelines powering executive BI dashboards.
- Planned and executed paid ad campaigns across Google, YouTube, Instagram, and Facebook for two client brands, a smart-eyewear brand and a cosmetics brand, driving improved engagement and ROI.

Snapthat Technologies **Apr 2019 – Sep 2019**
Data Engineer Intern Islamabad, Pakistan

- Designed scalable web crawlers that automated raw data acquisition for an internal AI product.

EDUCATION

M.S. Data Science — Information Technology University, Lahore **2020 – 2023**
B.S. Computer Science — COMSATS University, Islamabad **2015 – 2019**

PUBLICATION

Khan, U., Khan, H. U., Iqbal, S., & **Munir, H.** (2022). "Four Decades of Image Processing: A Bibliometric Analysis." *Library Hi Tech*, Emerald Publishing. doi.org/10.1108/LHT-10-2021-0351

CERTIFICATIONS — 30+ FROM DEEPLEARNING.AI, ANTHROPIC, COURSERA, IBM & MORE

· Complete Agentic AI Engineering — Ed Donner, Udemy
· Claude Academy: Claude 101 — Anthropic
· Machine Learning — Andrew Ng, Coursera
· Data Science Professional Certificate — IBM
· Deep Learning for AI — Dice Analytics
· DevOps: CI/CD, Docker, Kubernetes & Helm — Dice Analytics
· HIPAA Compliance Training — hipaatraining.us