

REHAAN SHAIKH

Pune, India

+91-7385599133 88rehaan88@gmail.com LinkedIn GitHub

PROFILE

AI engineer and NLP postgraduate at Cardiff University with hands-on experience in fine-tuning transformer models, deploying ML systems, and building RAG-based applications. Focused on real-world impact through LLMs and end-to-end AI workflows.

SKILLS

Languages & Tools: Python | R | SQL | Git | Docker | FastAPI
AI & ML: Neural Networks | CNNs | RNNs | Scikit-learn | Model Optimization
LLMs & Techniques: BERT | DistilBERT | GPT | LLaMA | Gemini | RAG | LoRA | PEFT | Prompt Engineering
Frameworks & Libraries: PyTorch | TensorFlow | Hugging Face | LangChain | FAISS
Deployment & Cloud: AWS | Azure | Hugging Face Spaces

EXPERIENCE

AI Engineer (Dissertation Project) | SimplyDo, Cardiff, United Kingdom June 2025 – Sept 2025

- Built a semantic idea retrieval system using **Gemini embeddings + FAISS**, reducing query time by 85%.
- Added **filtering and fallback mechanisms** to handle vague/metadata-only queries.
- Automated **data pipelines** with client-specific privacy controls.
- Achieved **100/100 SUS usability score** via client testing across diverse queries.
- Collaborated with CTO to align tech with product goals for production readiness.

Data Science Intern | Unified Mentor, India (Remote) May 2024 – July 2024

- Performed EDA on sales data, engineering features that improved forecasting accuracy by **15%**.
- Built ML models for **classification and time-series forecasting** in retail use cases.
- Pre-processed raw data using **Pandas/NumPy**, reducing model error.

Data Analyst Intern | SkillReigns, Pune, India Oct 2023 – Mar 2024

- Built a **churn prediction model** using scikit-learn on 30k+ e-commerce records; improved retention by **12%**.
- Delivered insights via **SQL + Power BI**, reducing reporting time by **40%**.
- Automated SQL workflows to speed up business reporting.

PROJECTS

Fine-Tuning BERTweet for Sentiment Analysis ↗

- Fine-tuned **BERTweet** on multiclass Twitter data using Hugging Face; achieved **63% accuracy, 58% F1** under class imbalance.
- Benchmarked against CNN + Word2Vec and DistilBERT, highlighting transformer models' edge for short-text sentiment.

CRAG: Corrective RAG Chatbot ↗ | Code

- Built a RAG chatbot for cardiology using **Chroma + FLAN-T5 + BGE-small**; improved response accuracy by **40%**.
- Deployed with **Mistral-7B, TinyLLaMA**, fallback web search (Tavily API) on **FastAPI + Streamlit**.

Parameter-Efficient Fine-Tuning Analysis ↗

- Replicated BitFit paper; compared **BitFit, LoRA, Diff Pruning** on BERT, GPT-2, and T5.
- Found LoRA offered best **accuracy–efficiency** trade-off on small datasets; BitFit was simplest.

EDUCATION

Cardiff University
Masters in Natural Language Processing

Cardiff, United Kingdom

Modern College
Bachelors in Statistics

Pune, India