Talal Khan

+92 304 2901824 • talalkhan213@gmail.com • talalkhan47 • tk-474

Academic History

2021 – 2025: BSCS, Institute of Business Administration

2019 – 2021: A-Levels, Nixor College

2016 – 2019: O-Levels, Beaconhouse School System, Jubilee Campus

Experience

Aga Khan University - Clinical Systems

ICT EHR Trainee *July 2025 – Present*

- Providing support for current applications of AKU.
- Building new applications for new requirements on .NET framework.
- Working on an EHR system, Meditech Expanse to move and consolidate data from legacy systems to Meditech.
- Integrating AI into current applications of AKU and also building new AI applications to improve the expereience of the users.

CodeXcue

Machine Learning Intern *May 2024 – July 2024*

- Worked on making regression models for predicting tips in a restaurent.
- Made classification model to identify spam emails.
- Explored and understood Hyperparameter Tuning and how much difference it makes.
- As a golden project made a model to predict laptop prices based on the specifications with an impressive R2 score of 0.9985.

Major Reports/ Research Work/Projects

Urdu Transcription & Diarization App - FYP:

- Designed and developed an end-to-end Urdu audio processing flutter application for automatic speaker diarization, transcription, and summarization called Harf ba Harf.
- Utilized FastAPI for backend services, integrated with Ngrok for secure link with Firebase our database.
- Implemented Whisper large-v3-turbo for accurate Urdu speech-to-text transcription with a WER of 27%.
- Used pyannote-audio for speaker diarization, enabling segment-wise differentiation of speakers in multi-speaker audio files.
- Integrated a Fine-Tuned MBart model for generating concise Urdu summaries of transcribed content.

Using LoRA to Fine-Tune TinyLLaMA (SFT + PFT):

- Fine-tuned TinyLLaMA-1.1B on Databricks Dolly-15k using five LoRA configurations with varying rank, alpha, dropout, and target modules.
- Conducted supervised fine-tuning (SFT) followed by preference fine-tuning using DPO on Argilla UltraFeedback dataset.
- Integrated Hugging Face Hub for model storage and automatic pushing post-training across all trial runs.
- Evaluated models using BLEU score and perplexity, achieving 14% perplexity reduction with best configuration (Trial 5).
- Applied gradient checkpointing and accumulation for compute-efficient training on constrained T4 GPU resources.

RAG for Design Analysis & Algorithms textbook:

- Developed a Retrieval-Augmented Generation (RAG) pipeline to answer questions from a Design and Analysis of Algorithms (DAA) textbook PDF.
- Implemented hybrid retrieval using FAISS (dense embeddings) and BM25 (sparse retrieval) with reciprocal rank fusion for improved relevance.
- Integrated LLMs (Qwen, Phi-2, LLaMA-3) to generate context-aware answers by synthesizing retrieved chunks.
- Designed an evaluation framework to assess retrieval relevance, answer faithfulness, and LLM response quality.

Skills/Special Courses

Programming Languages: Python, Java, JavaScript, C, C++, C#, Rust, GoLang

Libraries/Framework: React.js, Node.js, .NET, FastAPI, Django, PostgreSQL, Express.js, TensorFlow, PyTorch, scikit-learn

Add ons: PowerBI, Tableau, Excel, Adobe Photoshop, Figma, Canva, Advanced Latex, Adobe Premier, Machine Learning

Interests and Extracurricular Activities

Gaming (Online & Offline)

Cooking / Baking