Zohaib Khan

+92334-2200666 | zohaib.khan3502@gmail.com | Personal Website | LinkedIn | GitHub

Education

Lahore University of Management Sciences Lahore, Pakistan
BS in Computer Science, CGPA: 3.82, SCGPA: 3.94 *Aug. 2020 – May 2024*

Relevant Coursework: Machine Learning[†], Deep Learning[†], Introduction to Artificial Intelligence, Signals and Systems, Data Science, Dynamic Programming and Reinforcement Learning[†], Probability, Statistics, Econometrics, Applied Probability[†], Numerical Analysis

[†] *Graduate Level courses*

Research Projects

Fairness of LLMs across Social Dimensions

- **Objective:** To probe for biases in Large Language Models, including Gemma and Llama-3.2, along the dimensions for Gender, Race, and Religion.
- **Methodology:**
  - We automated the creation of a dataset of prompts to probe LLMs for such biases, including themes of debate generation, story generation, CV creation, generation of career advice etc. This dataset of over 3000 prompts would be the foundation of our study into how LLMs let their biases seep into their generations.
  - For our study, we utilized Gemma-2-2b, Gemma-2-9b, Llama-3.2-1b, and Llama-3.2-3b for a lack of more compute. We performed inference with these models, anonymized the outputs to remove explicit mentions of such identity, and used GPT-4o as a Judge for evaluating whether there was a difference in quality over a pair of generations.
  - Our results showed that all LLMs tended to prefer the generations associated with certain groups over others, regardless of the type of trigger that was involved.
  - We also conducted an analysis, varying the language of the input that was fed to the model - this was done by translating to Arabic and German using GPT-4o-mini. Our results showed that there were some variations in the groups being favored when we analyzed these new results.
- **Status:** Our work has been submitted to NAACL 2025. You can find the preprint here: https://arxiv.org/abs/2410.12499.

LLMs as Factcheckers

- **Objective:** To evaluate Large Language Models on their fact-checking capabilities, for them to serve as tools to verify the veracity of claims. Alongside this, we aimed to explore how well these models generalized to languages beyond English, and if they exhibited any biases, alongside an analysis of why they sometimes succeed and sometimes failed to give a correct response.
- **Methodology:**
  - In order to evaluate our choice of LLMs, we created a large dataset of over 300000 claims by scraping data from the Google Fact Check API which contained reviewed claims from over 50 reputable Fact-Checking organizations including Politifact, Snopes, Alt News, and others all over the globe.
  - To clean the data, we removed instances whose reviews were ambiguous and not objectively True nor False. This allowed for a more standardized evaluation and ensured that there were no biases during the annotation process.
  - To mimic how an average user may interact with an LLM, we devised prompting strategies that were deliberately simplistic and allowed us to gauge whether certain nuances in questioning elicited the model to change it's response for the same claim. Alongside this, more mechanical methods were devised to structure the model's responses and override it's default behavior.
  - To collect the model responses, we created an inference pipeline for the following models: GPT-3.5, GPT-4, LLaMA-2 (7B, 13B, 70B), Mistral-7B, and Mixtral-8x7B. The verbose responses from the models were then mapped back into binary True and False values as to facilitate an evaluation akin to that of a classification model.
  - To deal with non-English claims, the models were prompted in the native language of the claim. To address the issues with annotating non-English model responses, we utilized the Google Translate API in mapping the responses back into English.

    - ∗ To consolidate our findings, we designed an evaluation pipeline involving computing the certainty rates and the stability of model responses, alongside modified versions of accuracy, precision, and recall, to account for the special nature of responses where the model was uncertain.

- **Status:** We have submitted our work to AAAS Science.

Dynamic Group-Query Attention

- **Objective:** To improve upon the original design for Group Query Attention (GQA) for Transformers by performing the grouping for Attention Heads in a more informed fashion, to improve both performance and efficiency of these models.
- **Methodology:**
    - ∗ To move beyond the original and static process of grouping heads, we devised an approach involving looking at the L_2 norms of the Key projection matrices, among other heuristics. These would allow us to perform a more efficient form of grouping in allocating more Queries to ”denser” Keys.
    - ∗ To prototype our models, we evaluated on Vision Transformers that were uptrained post-conversion. Our experiments showed that our new mechanism performs better than the original GQA design.
    - ∗ To test the practicality of our approach, we evaluated the performance of our approach on CIFAR-10, CIFAR-100, Food101, and Tiny-ImageNet, reaching improvements by up to **8%** in ViT-L.
    - ∗ We will be exploring this approach for Language Models including GPT-2, T5, and Pythia. These will be evaluated on datasets for evaluating LLMs including GSM8k, MMLU, ARC, and LAMBADA.
- **Status:** Our work will be submitted to ACL 2025. You can find the preprint for our results on ViT here: https://www.arxiv.org/abs/2408.08454.

Approxify: LLMs for Code Approximations & Large Codebase Rewriting

- **Objective:** Approximate computing is a computational paradigm that aims to improve efficiency of programs by allowing for some degree of inaccuracy or approximation in computation. Our objective was to evaluate the effectiveness of Large Language Models in performing these approximations on large codebases.
- **Methodology:**
    - ∗ To address the challenges of prompting an LLM like GPT-4 with the content of large codebases where the context length is a limitation, and the repository can be separated across multiple files, we utilized a Program Dependency Graph (PDG) to feed in the repository in chunks.
    - ∗ To implement our PDG, we utilized clang that allowed us to trace dependencies of entities like variables, types, and functions in C and C++, across multiple files. The graph underwent a topological sort so that the model would not deal with dependencies that had not been seen before.
    - ∗ To perform the actual approximations to the code, we prompted GPT-4o and GPT-3.5 to modify snippets with techniques including but not limited to loop perforation, loop truncation, approximate memoization etc. Since these techniques involve certain parameters, the models were prompted to add these in as ”knob variables” for the user to tweak, and to suggest ranges of values to test out.
    - ∗ To ensure that the code did not break post-modification, the pipeline involved compiling the code again and pruning the search space of the knob variables before moving on to the next snippet of code.
    - ∗ To find the optimal set of knob values for approximating the repository as a whole, we utilized a Bayesian Optimization approach to aid our search for the lowest error.
    - ∗ To test the efficacy of this approach, we evaluated our pipeline on benchmark repositories like SUSAN, FFT, LQI, and String Search. For a more grounded and practical application, i.e. using edge-maps to detect leopards (from video-streams of wildlife), we found that replacing SUSAN with an approximated version shaved off 8 seconds in our batched inference benchmark, while our classifier’s performance incurred less than a 20% drop in accuracy on our validation set.
- **Status:** We have submitted our work to ACM SenSys 2024.

Urdu Chatbot for Self-Attachment Therapy

- **Objective:** To create a Urdu chatbot assisting users in Pakistan with performing Self-Attachment Therapy (SAT), a self-administered psychological technique. This project seeks to add on to the suite of applications under the same premise, across different languages, as was initiated by researchers at Imperial College.
- **Methodology:**
    - ∗ To design the system and address the sensitive nature of the application, our chatbot uses rule-based and classification-based modules for user comprehension and navigates a dialogue flowchart accordingly, recommending appropriate SAT exercises.

- ∗ To generate our pool of Urdu dialogues, we leveraged previous iterations of the project (in Farsi, Mandarin, and English) to translate the existing sentences in Urdu, and rewriting to introduce more variety.
- ∗ To maintain information regarding the emotional state of the user, we trained an Encoder-based Emotion Classifier in Urdu to predict user emotions based off the conversation history.
- ∗ To simulate the actual conversation, a retrieval-based system was implemented that fetched responses from the pool of Urdu utterances, based off embedding similarity with messages from the conversation. It was important to fetch fixed responses rather than rely on Language Models for generation for safety and control, owing to the sensitive nature of the task.
- ∗ To evaluate our system on its efficacy, we follow the protocol of previous iterations of a non-clinical study encompassing participants of various backgrounds and age brackets who are asked to interact with the system and rank it on the basis of several characteristics and measures.

- **Status:** We are conducting evaluations before intending to wrap up the project, with an intended submission to ACL 2025.

Analyzing User Retention on Information Dissemination Platforms

- **Objective:** To determine whether there is a preferred method of presenting information to users via information dissemination platforms. For this study, we elected to present facts in three modes: as short messages/factoids, stories, and quizzes.
- **Methodology:**
  - ∗ For our study, we utilized an Interactive Voice Response (IVR) platform similar to Polly and Baang that were previously launched in Pakistan. This, combined with a simple push-button interface, allowed us to reach a large portion of the population of Pakistan that were illiterate.
  - ∗ To go about measuring whether a user was able to effectively learn from a certain mode, we utilized a ”Pre-test and Post-test” design, where a user would be asked to answer a question before and after being presented with a treatment (as information presented in one of the three modes).
  - ∗ To gather insights from our dataset of over 600000 interactions collected during the deployment, we performed many iterations of cleaning, aggregation and processing on the raw data to test our hypotheses. This involved analysis of data on different levels all the way from each of the modes across all users, down to individual interactions of unique users.
  - ∗ To discount irrelevant or careless interactions from users, we analyzed the Local Average Treatment Effect (LATE) of our treatments and distinguished between compliers and non-compliers. This allowed us to prune out instances of our data that would otherwise hide the true effect our system had on users.
  - ∗ To consolidate our findings, we ran an Analysis of Variance (ANOVA) to highlight the best mode of communication for information dissemination, and quantified the portions of the population with positive belief updates post-treatment.
- **Status:** We are continuing to test different strategies for pruning irrelevant interactions and iterating on our analysis. A paper publication is intended for COMPASS 2025.

Urdu Text-to-Speech Synthesis

- **Objective:** To design a Text-to-Speech (TTS) system to clone voices for Pakistani figures in Urdu.
- **Methodlogy:**
  - ∗ To address the low-resource nature of the Urdu language, we collected our own dataset for the target figures - the audios were sourced from YouTube and annotated manually before being converted into appropriate formats like LJSpeech for model training.
  - ∗ To create the system, we trained single-speaker and multi-speaker models, such as YourTTS and VITS, for the synthesis of audios in Urdu for the target figures.
  - ∗ To address poor data quality and subpar model performance, we highlighted issues such as repetitions, noisy segments, irrelevant audios, alongside sourcing additional audio data for model training.
- **Status:** The completed systems were utilized in furthering a study on fighting misinformation and the detection of spoofed audios in the context of Pakistan.

OCR for Handwritten Urdu

- **Objective:** To train an Optical Character Recognition (OCR) model for handwritten Urdu.
- **Methodology:**
  - ∗ To address the low-resource nature of the Urdu language, we collected data in the form of scanned images of (pages of) Urdu books, overseeing the annotation process done via crowdsourcing.

  - ∗ To create our system, we trained an OCR model using Google Tesseract for Handwritten Urdu, boasting Character-Error Rate of 0.02, with explicit support for digits. Other models, such as EasyOCR and CRNNs were excluded on the basis of latency and performance.
  - ∗ To improve upon the performance, we utilized a post-processing mechanism with Transformers to make correct minor mistakes with the raw transcript.
- **Status:** This project has been completed; the code can be found here.

Experience

Research Associate Jan. 2022 – Present
CSaLT at LUMS *Lahore, Pakistan*

- Working as a Research Associate for CSaLT under Dr. Agha Ali Raza, contributing and handling projects including but not limited to: Urdu OCR, Urdu TTS, IVR Platform Data Analysis, improving upon the design of Transformers, evaluating LLMs on understanding and usability for downstream tasks.

Research Associate June 2024 – Present
CITY at LUMS *Lahore, Pakistan*

- Working as a Research Associate for the Center for Urban Informatics, Technology and Policy (CITY) Lab under Dr. Muhammad Tahir. My work revolves around making models more efficient and robust to out-of-domain data.

Fatima Fellow June 2024 – Present
Fatima Fellowship *Remote*

- Volunteering as a researcher at the Fatima Fellowship program, working on Multilingual Redteaming for LLMs. This is in collaboration with graduate students at Georgia Tech and Stanford, and undergraduate students around the world.

Head Teaching Assistant: CS6304 - Advanced Topics in ML Sep. 2024 – Present
LUMS *Lahore, Pakistan*

- Currently working as the Head TA for CS6304 at LUMS under Dr. Muhammad Tahir.
- Designed this course as the first of its kind in LUMS, intended to educate students on some advanced topics in Machine Learning, including but not limited to: Neural Network Compression, Out-of-Domain Detection and Generalization, Federated Learning, Deep Unfolding etc.
- Designed assignments, demos, and reading material. Graded course components and held Office Hours for student support.

Lab Lead June 2023 – June 2024
CSaLT at LUMS *Lahore, Pakistan*

- Worked as a lab lead for CSaLT (among 2 others) under Dr. Agha Ali Raza, with responsibilities including but not limited to: handling lab logistics, supervising students, managing research projects, updating lab media.
- Handled the on-boarding and assignment of over 60 students for Senior-level Projects, and Junior-level Directed Research Projects.
- Expanded the scope of the lab with new projects, new collaborations, and better systems for handling regular procedures.

Content Writer May 2024 – Present
Self-employed

- Occasionally write blogs on my personal website on the topics of Machine Learning, NLP, and Generative AI, with a focus on implementing and understanding of architectures and tools on a lower-level.

Head Teaching Assistant: CS5302 - Speech and Language Processing Jan. 2024 – May 2024
LUMS *Lahore, Pakistan*

- Worked as the Head TA for CS5302 at LUMS under Dr. Agha Ali Raza, managing a team of 6 TAs for **a class of over 130 students**.
- Designed this course as the first of its kind in LUMS, intended to educate students on the foundations of Generative Artificial Intelligence covering topics including but not limited to: Transformers, Large Language Models, Training and Finetuning mechanisms, Prompt Engineering, RAG and Vector Databases, Quantization and other aspects of Efficient Deep Learning.
- Designed assignments, demos, quizzes and reading material. Graded course components and held Office Hours for student support.

Head Teaching Assistant: CS535 - Machine Learning Sep. 2023 – Dec. 2023
LUMS *Lahore, Pakistan*

- Worked as the Head TA for CS535-Machine Learning under Dr. Agha Ali Raza, managing a team of 11 TAs for **a class of over 220 students**.
- Revamped the syllabus for the second half of the course, with the introduction of Sequence Models, Transformers, and Unsupervised Learning.
- Designed course materials, including but not limited to: assignments containing a good blend of theoretical and applied topics, slides for the aforementioned topics, and over 5 manuals for topics meant for a 6-week long project.
- Held weekly tutorials, graded components, and handled logistics.

Machine Learning Engineer May 2023 – Sep 2023
ISSM.ai *Lahore, Pakistan*

- Prepared a survey of the current landscape of Document-Understanding models for the purpose of extracting Invoice Numbers from receipts.
- Designed a pipeline to extract Invoice Numbers from the detection of QR Codes, and written text using OCR, within receipts.
- Rewrote the source code for existing tools to train custom Text Detection and Recognition models for a private dataset of over 15000 images, annotated by the team.
- Improved the pipeline by optimizing the final models on Intel CPUs, using ONNX and OpenVINO, dropping the latency per image from 1700ms to 350ms, with 0.83 Recall and 0.8142 Precision.

Research Assistant June 2023 – Aug. 2023
CITY at LUMS *Lahore, Pakistan*

- Worked as a Research Intern under Dr. Muhammad Tahir on Signal Processing and Machine Learning topics.
- Created learning material for on-boarding students, on topics including PyTorch, CNNs, Autoencoders, Vision Transformers, and relevant topics in between.
- Prepared a survey of the current landscape of using Foundation Models, such as CLIP and GPT-3.5, for Geospatial AI tasks.
- Compared and implemented different Attention Mechanisms across a variety of CNN architectures for Image Classification, and employing these techniques to Spatiotemporal data for Demand Forecasting of taxis in cities.

Deep Learning Intern June 2022 – March 2023
CodeSlash *Lahore, Pakistan*

- Created an Object Tracking system using YOLOv5 and StrongSORT to track Cattle from satellite imagery, with a 0.542 mAP.
- Developed the backend for a Face Recognition system using OpenVINO, MongoDB, FastAPI and Docker, running at 15 FPS.
- Optimized the backend, using multiprocessing in C++, for a Computer Vision system to stitch together frames, from a video stream, of the underside of a car, as it passes over the camera, to detect foreign objects.

Teaching Assistant: CS437 - Deep Learning Jan. 2023 – May 2023
LUMS *Lahore, Pakistan*

- Worked as a Teaching Assistant for the course CS437 Deep Learning under Dr. Murtaza Taj, for a class of 80+ students.
- Held weekly tutorials to teach students how to tackle Deep Learning problems with Python and PyTorch, along with reviewing content and sharing assorted readings.
- Created assignments for fundamentals, CNNs, Autoencoders, RNNs in PyTorch alongside using HuggingFace for NLP tasks.
- Supervised 4 groups for course projects related to the distinct topics of Neural Network Compression, Forest Fire Detection, and Image Segmentation for Crops - this involved creating a survey of previous works, developing a training and evaluation pipeline, and writing a report in academic fashion.

Technical Skills

Languages, Libraries, and Frameworks: Python, C, C++, JavaScript, NumPy, JAX, Pandas, Matplotlib, Seaborn, TensorFlow, PyTorch, Scikit-Learn, HTML/CSS, React, Node.js, MongoDB, PostgreSQL, MySQL, Express, FastAPI, Transformers, OpenAI SDK, LangChain, AutoGen
Skills and Other Tools: Linear Algebra, Google Colab/Jupyter Notebooks, Computer Vision, NLP, Machine/Deep Learning, Calculus, MATLAB and elementary signal processing, Elementary Frontend Development

Volunteering and Leadership

LUMS Boxing Team Summer 2022 – Summer 2024

Captain *LUMS*

- Coached and held training sessions for a team of 35+ members.
- Organized events, coordinating with the administration for a smooth execution.
- Developed teamwork and leadership skills through planning, and solving challenges in a high-pressure environment.

LUMS Data Science Society Winter 2020 – Spring 2023

Director *LUMS*

- Developed notebooks on Python, Numpy and Pandas to teach beginners the tools of elementary Data Analysis.
- Assisted with the preparation of a national Data Hackathon which was sponsored by Kaggle.
- Headed a workshop to teach Python to non-programmers.

Juno Circle Astronomy Society Winter 2019 – Summer 2020

Vice President *Cedar College*

- Designed and co-headed the Mathematics module in the national STEM olympiad Scinnova IV.
- Participated in organization of the Cedar College intraschool STEM competition Scientia.

Achievements

- Graduated with the High Distinction Award from LUMS.
- Chosen as one of the three lab leads for CSaLT under Dr. Agha Ali Raza.
- Dean's Honors List 2020-2023.
- Chosen as the LUMS Boxing Team Vice Captain for 2022, and Captain in 2023.
- Top 2% of 250000+ users on the Codewars platform.
- Dean's Honor Roll throughout A Levels.
- Winners of the Karachi Grammar STEM Olympiad 2020.
- Winners of the Hypercube STEM Olympiad 2018.
- Runners-up of the Karachi Grammar Math Olympiad 2019.
- Runners-up of the Karachi Grammar STEM Olympiad 2019.