Samee Arif

samee.arif@lums.edu.pk | Website | Blog | Google Scholar | GitHub | Huggingface

RESEARCH INTEREST

My research focuses on Natural Language Processing and Machine Learning, specifically Fairness and Biases in AI, LLM Optimization, Multilingual NLP, and Human-Centered AI.

EDUCATION

Bachelor of Science in Computer Science *Sep 2019 - May 2023*

Lahore University of Management Sciences (LUMS)

Relevant Coursework: Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, Speech Processing, Principles and Techniques of Data Science, Computer Vision, Mathematical Foundations for Machine Learning and Data Science, Probability, Calculus I & II, Linear Algebra

RESEARCH EXPERIENCE

Measuring Fairness and Bias Across Social Dimensions *Aug 2024 - Oct 2024*

- Developed (Social Appropriateness in LLM-Generated Text) SALT Dataset with 4,225 prompts across gender (350), religion (1,625), and race (2,250) dimensions to systematically analyze biases in LLMs.
- Designed and implemented seven unique bias triggers (e.g., Career Advice, Story Generation, Debate) to systematically evaluate biases in LLM-generated text across social dimensions.
- Analyzed biases across Llama and Gemma families, revealing consistent polarization, with **Bias Scores ranging from -0.92 to +0.92,** highlighting both favorable and unfavorable group treatment.
- Pioneered the use of LLM-based judgments for scalable bias detection, validated with high inter-annotator agreement of **Cohen's Kappa: 0.72–0.76**.
- Demonstrated language-specific bias variations in multilingual settings, uncovering stronger biases in English and Arabic compared to German models.
- The research paper for this project is currently under review.

Multi-Agent Workflows for Preference Optimization Dataset Generation *Jun 2024 - Aug 2024*

- Proposed a novel multi-agent framework for automated Preference Optimization dataset generation.
- Received an OpenAI Research Grant to implement the framework, leading to an **11% improvement in preference alignment** for fine-tuning large language models.
- Developed and tested LLM-based evaluation methodologies, including Judge, Jury, and Debate strategies, identifying GPT-4o as the most consistent and reliable evaluator.
- Designed a multi-agent workflow using Llama as the generator and Gemma as the reviewer, demonstrating superior dataset generation efficiency and scalability.
- Published generated DPO and KTO datasets alongside the framework's code, ensuring open access for further research and advancing the state of AI optimization methodologies.
- The research paper for this project is currently under review.

Multi-Agent Generative AI for Multimodal Stories *July 2024 - Sep 2024*

- Designed and implemented a novel multi-agent framework combining LLMs, Text-to-Speech, Text-to-Video, and Text-to-Music models to co-create dynamic, multimodal stories for children.
- Integrated Freytag's Pyramid and Propp's 31 Narrative Functions to ensure structurally coherent and engaging narratives tailored for educational and entertainment purposes.
- Conducted human evaluation across three modules (LLM, TTS, TTV) with **inter-rater agreement scores above 0.60**..
- Demonstrated the superiority of smaller LLMs (e.g., Llama-3.1-8b) in generating creative, cost-effective, and child-friendly stories, achieving significant reductions in computational overhead compared to larger models.
- Released datasets for benchmarking child-safe content filtering alongside the system code.
- The research paper for this project is currently under review.

Benchmarking Urdu ASR Models *July 2024 - Sep 2024*

- Led a team of 14 interns and created the first Urdu conversational speech dataset for evaluating Automatic Speech Recognition (ASR) models, including **471 recordings with human-generated transcriptions**.
- Fine-tuned state-of-the-art ASR models using Mozilla's Common Voice and Google's Fleur's datasets, significantly reducing WER across tasks.

- Benchmarked nine models (Whisper, MMS, Seamless-M4T families) on read and conversational speech datasets, achieving **WER as low as 17.09% on read speech** and **17.86% on conversational speech** after fine-tuning.
- Conducted quantitative and qualitative analyses, uncovering persistent challenges like text normalization and overlapping speech handling and highlighting the importance of multimodal learning for low-resource languages.
- Released all models, datasets, and evaluation scripts to the community via GitHub, advancing ASR research for 70+ million Urdu speakers and other low-resource language applications.
- The research paper for this project has been accepted at **COLING 2025**.

Evaluating Large Language Models for Urdu *April 2024 - June 2024*

- Analyzed 15 Urdu datasets and revealed issues of high noise, inconsistent annotations, and limited linguistic diversity, emphasizing the need for improved data quality in low-resource language NLP.
- Conducted quantitative and qualitative comparisons of generalist models (GPT-4-Turbo, Llama-3) and specialist models (XLM-R, mT5, fine-tuned Llama-3) across 14 Urdu NLP classification and generation tasks.
- Received OpenAI Research Grant, enabling evaluation of GPT-4-Turbo alongside Llama-3, leading to deeper insights into low-resource language model capabilities.
- Fine-tuned XLM-R, mT5, and Llama-3 using PyTorch and memory-efficient QLoRA, addressing computational constraints, and **fine-tuned models outperforming generalist models across 7/7 classification tasks**.
- Demonstrated that GPT-4-Turbo outperformed fine-tuned models for generation tasks, achieving **49/50 wins in summarization, paraphrasing, and translation tasks**, highlighting the need for qualitative assessments alongside quantitative metrics.
- The research paper for this project has been accepted at **EMNLP Findings 2024**.

Corpus for Urdu Question Answering *Jan 2023 - Oct 2023*

- Developed UQA, the largest high-quality Urdu QA dataset, featuring **124,745 question-answer pairs**, using a novel EATS (Enclose to Anchor, Translate, Seek) technique for aligning answer spans in machine-translated text.
- Fine-tuned mBERT, XLM-R, and mT5 on UQA, achieving an **85.99% F1 score** and **74.56 Exact Match**, outperforming existing benchmarks.
- Leveraged machine translation to generate the dataset efficiently, led a team of 12 research interns and conducted human evaluation to ensure data quality.
- Authored multiple blog posts detailing model architectures and fine-tuning scripts.
- The research paper for this project has been accepted at **LREC-COLING 2024**.

GradSelect and SoPify - Student Counseling Chatbot *Aug 2023 - May 2024*

- Created an AI-powered tool for educational counseling, assisting students with Statement of Purpose writing and creating a balanced graduate school university list tailored to individual needs.
- Enhanced the system with Automatic Speech Recognition and Text-to-Speech, fine-tuning models to enable seamless voice interaction for accessibility and usability.
- Designed and implemented effective prompt engineering strategies to optimize large language model outputs, ensuring relevant and personalized guidance.
- Developed an intuitive front-end interface using Next.js and React.js, providing a seamless user experience for students and educators.
- The chatbot can be accessed at the ActualAIz website.

Image-to-Speech Pipeline for Urdu Language | *Python* *Sep 2021 - Sep 2022*

- Evaluated Optical Character Recognition models, including Tesseract, EasyOCR, and Kraken on Nastaliq font.
- Established a pipeline to replicate scanned images using data augmentation to generate the dataset.
- Fine-tuned GANs to map the noisy images to clean images as a pre-processing module.
- Implemented a post-processing module based on BERT, Google search engine auto-correction, and conditional random fields to enhance the model accuracy.
- Trained Tesseract to achieve a **1.53% Character Error Rate** and piped it with a fine-tuned TTS model.

WORK EXPERIENCE

Research Associate | *CSaLT (LUMS)* *June 2024 - Present*

- Advisor(s): Dr. Agha Ali Raza (LUMS) and Dr. Awais Athar (EMBL-EBI).
- Working on multi-agent frameworks for LLM, preference optimization, and synthetic dataset generation.

GenAI and Software Engineer | *ActualAIz (LUMS)* *Aug 2023 - June 2024*

- Advisor(s): Dr. Agha Ali Raza, Dr. Ihsan Ayyub Qazi and Dr. Zafar Ayyub Qazi (LUMS).
- Worked on developing a multimodal and multilingual graduate assistant tool leveraging large language models to provide educational counseling.

Research Assistant | *CSaLT (LUMS)* *Aug 2021 - May 2023*

- Advisor(s): Dr. Agha Ali Raza (LUMS) and Dr. Awais Athar (EMBL-EBI).
- Worked on image-to-speech pipeline and Urdu question-answering system.

Teaching Assistant | *Machine Learning (LUMS)* *Fall 2022*

- Oversaw and facilitated learning for a cohort of more than 140 students. Designed and administered course quizzes, assignments, and a project to gauge student understanding and progress.

Teaching Assistant | *Computational Problem Solving (LUMS)* *Fall 2021*

- Managed a 93-student cohort, designed quizzes and labs, and held weekly office hours.

PROJECTS

Speech Technologies *Aug 2023 - Dec 2023*

- Analyzed ASR model quality and inference time, integrated quantization for faster inference, and utilized QLoRA for efficient fine-tuning.
- Trained MMS-TTS and YourTTS, adapting a VITS TTS framework script for training.
- Created a web-based audio annotation tool providing editable transcriptions and timestamps using ASR.

ConvoLense *Aug 2023 - Sep 2023*

- Evaluated speech-based (Wav2Vec2) and text-based (BERT, mT5, GPT, LLaMA) emotion classifiers.
- Used Bark to generate a synthetic conversation dataset between customer and customer representative.
- Established a pipeline for emotion classification using my ASR model and LLM.

Arabic Handwriting Recognition *Jan 2023 - May 2023*

- Applied transfer learning techniques to adapt the Urdu OCR model for recognizing handwritten Arabic text.
- Utilized advanced pre-processing methods, such as skeletonization, to generate a synthetic dataset.

Image Captioning *Jan 2023 - May 2023*

- Conducted an experimental fine-tuning of Swin-Transformer on the Indiana University - Chest X-Rays dataset, exploring its application in medical image analysis.

Fraudulent Job Prediction *Sep 2022 - Dec 2022*

- Trained Logistic Regression, Support Vector Machine, and Random Forest classifiers to identify real versus fake job postings, achieving a **91% Accuracy**.
- Conducted comprehensive data cleaning and exploratory data analysis on the dataset.

Lane Analysis for Autonomous Vehicle *Sep 2022 - Dec 2022*

- Created a lane-change warning system, integrating Lanenet for lane detection and YOLOv7 for vehicle detection.

Learning Management System *Jan 2022 - May 2022*

- Created a platform for schools to manage online education during the pandemic.

Speech-based Language Classifier *Sep 2021 - Dec 2021*

- Recorded voice samples in English, Urdu, and a mix of both languages at 1600MHz.
- Developed and trained a neural network from scratch to classify speech using the recordings dataset.

FoodSwings *Sep 2021 - Dec 2021*

- Implemented food delivery web application.

Neural Network from Scratch *Sep 2021 - Dec 2021*

- Developed a feed-forward neural network from scratch using NumPy and optimized it with Numba JIT.

AWARDS

Dean's Honour List | *LUMS* *Fall 2020*
Dean's Honour List | *LUMS* *Spring 2019*

TECHNICAL SKILLS

Languages | *Python, C/C++, SQL, JavaScript, HTML/CSS*
Frameworks | *React, Node.js, Next.js, FastAPI*
Developer Tools | *Git, Docker, Google Cloud Platform, VS Code, Visual Studio*
Libraries | *pandas, NumPy, Matplotlib, TensorFlow, PyTorch, Keras, transformers, Streamlit*

PUBLICATIONS

[6]. **Samee Arif**, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza. With a Grain of SALT: Are LLMs Fair Across Social Dimensions? **Preprint**

[5]. **Samee Arif**, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza, The Fellowship of the LLMs: Multi-Agent Workflows for Synthetic Preference Optimization Dataset Generation. **Preprint**

[4]. **Samee Arif,** Taimoor Arif, Muhammad Saad Haroon, Aamina Jamal Khan, Agha Ali Raza, Awais Athar, The Art of Storytelling: Multi-Agent Generative AI for Dynamic Multimodal Narratives. **Preprint**

[3]. **Samee Arif**, Aamna Jamal, Mustafa Abbas, Agha Ali Raza, Awais Athar. WER We Stand: Benchmarking Urdu ASR Models. In the **Proceedings of The 31st International Conference on Computational Linguistics** (Core: B, H5-index: 65)

[2]. **Samee Arif**, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza, Generalists vs Specialists: Evaluating Large Language Models for Urdu. In **Findings of Empirical Methods in Natural Language Processing 2024.** November 12-16, 2024, Miami, US. (**EMNLP Ranks 2nd in Computational Linguistics, Core: A*, H5-index: 193**)

[1]. **Samee Arif**, Sualeha Farid, Awais Athar, and Agha Ali Raza, UQA: Corpus for Urdu Question Answering. In **LREC-COLING 2024** – Joint International Conference on Computational Linguistics, Language Resources and Evaluation. May 20-25, 2024, Torino (Italia). (**COLING Ranks 5th in Computational Linguistics | LREC Ranks 6th in Computational Linguistics**)

Research Grants

July 2024: The project *The Fellowship of the LLMs: Multi-Agent Workflows for Synthetic Preference Optimization Dataset Generation* received OpenAI Research Grant.

May 2024: The project *Generalists vs Specialists: Evaluating Large Language Models for Urdu* received OpenAI Research Grant.