Samee Arif

samee.arif@lums.edu.pk • Website • Google Scholar • Github • Huggingface

RESEARCH INTEREST

My research focuses on Natural Language Processing (NLP), Educational Technologies, Human-Centered AI, and Synthetic Data Generation.

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

Multi-Agent Workflow for Iterative Self-improvement *Oct 2024 - Present*

- Working on using LLMs in multi-agentic workflows for iterative self-improvement.
- Implemented an iterative process for preference optimization dataset generation using LLM Feedback Loop and LLM-as-a-Judge. This approach generates an initial DPO dataset, fine-tunes a model, and then iteratively uses the fine-tuned model to create subsequent datasets.

With a Grain of SALT: Are LLMs Fair Across Social Dimensions? *Aug 2024 - Present*

- Present an analysis of biases in LLMs across gender, religion, and race.
- Introduced a methodology for generating a bias detection dataset using seven bias triggers: General Debate, Positioned Debate, Career Advice, Story Generation, Problem-Solving, Cover-Letter Writing, and CV Generation.
- Evaluated LLMs in three languages, English, German, and Arabic to explore how language influences bias manifestation.
- The research paper is **in submission at NAACL 2025**.

The Art of Storytelling: Multi-Agent Generative AI for Dynamic Multimodal Narratives
July 2024 - Sep 2024

- Developed a multi-agent system integrating LLM, Text-to-Speech (TTS), Text-to-Music (TTM), and Text-to-Video (TTV) models to generate interactive, multimodal narratives for children.
- Applied Propp's 31 narrative functions and Freytag's Pyramid to structure story generation, with Llama-3.1-8b as the story generator.
- Evaluated LLM-generated stories (GPT-4, Llama-3.1, and Gemma-2), TTS models (XTTSv2 and StyleTTS 2), and TTV model (CogVideoX-5b).
- The research paper is **in submission at Coling 2025**.

WER We Stand: Benchmarking Urdu ASR Models *July 2024 - Sep 2024*

- Evaluated three ASR model families—Whisper, MMS, and Seamless-M4T—on read and conversational speech datasets for Urdu.
- Developed the first conversational speech dataset for benchmarking Urdu ASR models, consisting of 471 recordings captured in real-world settings.
- Demonstrated that Whisper-large achieved the lowest WER (17.86) on conversational speech, while Seamless-large outperformed all models on read speech with a WER of 17.09.
- Highlighted error patterns, including common wrong words and error types like insertions, deletions, and substitutions, focusing on the need for robust Urdu text normalization.
- The research paper is **in submission at Coling 2025**.

The Fellowship of the LLMs: Multi-Agent Workflows for Synthetic Preference Optimization Dataset Generation
June 2024 - August 2024

- Evaluated multi-agent workflows for LLM-as-evaluators and LLM-as-generators modules to generate synthetic preference optimisation datasets using Llama-3.1, Gemma-2, and GPT-4 families.
- Tested LLM-as-a-Judge, Jury, and LLM Debate, to identify the most effective LLM-as-evaluator strategy.
- Demonstrated the effectiveness of Llama-3.1-8b as the generator with Gemma-2-9b as the reviewer, achieving a **71.8% and 73.8% win rate** against single-agent Llama-3.1-8b and Gemma-2-9b, respectively.
- Presented DPO and KTO datasets generated using the LLM Feedback Loop with GPT-4o-as-a-Judge, focused on single-agent improvement. Presented DPO and KTO datasets aimed at improving multi-agent LLM Feedback Loop configurations.
- The research paper is currently **in submission at NAACL 2025**.

Generalists vs. Specialists: Evaluating Large Language Models for Urdu *April 2024 - June 2024*

- Fine-tuned Llama-3, mT5, and XLM-R for 13 Urdu generation and classification tasks.
- Evaluated the fine-tuned models and compared their performance against GPT-4-Turbo and Llama-3-8b as the baseline.
- Presented benchmarking datasets in Urdu designed to evaluate the performance of LLMs as evaluators.
- First authored and published a research paper at **Findings of EMNLP 2024**.

GradSelect and SoPify - Student Counseling Chatbot *Aug 2023 - Present*

- Developed an LLM-powered graduate assistant tool to provide educational counselling.
- Implemented multimodality by fine-tuning and integrating Automatic Speech Recognition (ASR) and TTS models into the system.
- The Chatbot is available on ActualAIz website.

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

- Developed a question-answer corpus for the Urdu language to address the limited resources available in the domain.
- Manually evaluated Seamless M4T and Google Translator for Urdu.
- Introduced EATS - a technique to preserve the answer spans in the translated context paragraphs and employed it to translate the SQuAD2.0 dataset to Urdu.
- Successfully generated 124,745 question-answer pairs and fine-tuned mBERT, XLM-RoBERTa, mT5, and LLaMA-2 on our dataset to achieve an **85.99% F1 Score** and **74.56% Exact Match**.
- First authored and published a research paper at **LREC-Coling 2024**.

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

- Evaluated Optical Character Recognition (OCR) 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), Dr. Awais Athar (EMBL-EBI).
- Working on multi-agent frameworks for LLM, preference optimization, and synthetic dataset generation.

Research Associate | *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. Raza (LUMS), 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*

- Fine-tuned Whisper and MMS ASR model, achieving a **13.01% WER**. Analyzed 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 using my ASR model and LLM for emotion classification.

Arabic Handwriting Recognition *Jan 2023 - May 2023*

- Applied transfer learning techniques to adapt the Urdu OCR model for recognizing handwritten Arabic texrt.
- 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? **In Submission at NAACL 2025**

[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. **In Submission at NAACL 2025**

[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. **In Submission at Coling 2025**

[3]. **Samee Arif**, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, Awais Athar, WER We Stand: Benchmarking Urdu ASR Models. **In Submission Coling 2025**

[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 funding from the OpenAI Research Access Program.

May 2024: The project *Generalists vs Specialists: Evaluating Large Language Models for Urdu* received funding from the OpenAI Research Access Program.