Haris Riaz

Google Scholar | **Email:** hriaz@arizona.edu | **Github:** hriaz17 | **LinkedIn:** harisriaz17
Java (8 yrs), C (9 yrs), Scala (2 yrs), Python (9 yrs), SQL (7 yrs)

Education

University of Arizona (CGPA: 3.91 /4.0) Tucson, AZ
PhD in Computer Science, Minor in Cognitive Science *Jan. 2022 – Present*
Advisor: Mihai Surdeanu
Research Interest: ***Strategies for making LLMs reason Causally, Faithfully & Interpretably.***

National University of Sciences & Technology (CGPA: 3.75/4.0) Islamabad, PK
Bachelor of Science in Computer Science. *Sep. 2017 – June 2021*
Thesis: ***Handwritten Sequence Recognition with Time Series Transformers.***

Experience

Applied Scientist Intern May 2024 – September 2024
Amazon Web Services *Arlington, VA*

- Interned with **Amazon Science Bedrock** team
- Focused on techniques for scaling **Diverse Synthetic Data** generation for **Large Language Model (LLM)** training.
- Developed **Meta-algorithms** for creating synthetic data from scratch (using random seed keywords) or creating synthetic data by leveraging existing corpora with a **RAG-Guided-Search** approach.
- Created a high quality, fully synthetic & **formally diverse Instruction Tuning dataset** dataset of **25 million tokens**.
- Proposed and implemented metrics for measuring the **formal diversity** of LLM generated synthetic data.
- This work resulted in a paper currently under review for **ACL 2025.**

Graduate Data Scientist Intern June 2023 – August 2023
Kaiser Permanente *San Francisco, CA*

- Focused on the development of efficient NLP models for identifying social health indicators from clinical text, aiding in comprehensive patient health assessments.
- Built a diverse labeled training dataset of **over 1 million** provider notes using **heuristic annotations**, **zero-shot learning with LLMs**, **human-in-the-loop active learning** and multiple **weak supervision** strategies.
- Used this dataset to train an NLP model that achieved over a **90% F1 score** in expert evaluations, a state of the art in pinpointing key social determinants of health.
- Successfully deployed the model into production, now capable of performing batch inference on millions of provider notes monthly.
- Developed and optimized high-performance **distributed queries** for the **Epic Clarity database**, enabling efficient retrieval and processing of billions of records for advanced machine learning analysis.
- Gained hands-on experience with **Hive**, **AzureML**, **Azure Data Factory & Azure Synapse Pipelines**, **Snorkel Flow/Snorkel AI**, **John Snow Labs' NLP** and **Docker** deployment.

Graduate Research Assistant May 2022 – June 2023
University of Arizona *Tucson, AZ*

- Researched novel ways to understand people's opinions through text "Predicting What the Locals will Predict" (PWLP).
- Worked on an **unsupervised "co-ranking"** method for identifying local domain experts based upon graphical representations of named entities, and random walk algorithms for ranking those entities.
- This project is part of the **DARPA Habitus** initiative.

Publications

ELLEN: *Extremely Lightly Supervised* Learning *For Efficient* Named *Entity Recognition*. *In The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (**LREC-COLING 2024**).*

Best of Both Worlds: A Pliable and Generalizable Neuro-Symbolic Approach for Relation Classification. *In The 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (**NAACL 2024 Findings**).*

Deep neural network techniques in the calibration of space-charge distortion fluctuations for the ALICE TPC. *In The 25th International Conference on Computing in High-Energy Physics (**CHEP 2021**).*

Publications Under Review

MetaSynth: Meta-Prompting Your Large Language Model to Generate Formally Diverse Synthetic Data. ***In Review for ACL 2025**.*

Say Less, Mean More: Leveraging Pragmatics in Retrieval-Augmented Generation. ***In Review for NAACL 2025**.*

Peer Reviewing Experience

- Reviewed **3 papers** for **The Second Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning (Pan-DL), 2023.**
- Reviewed **1 paper** for the **2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024.**
- Currently a reviewer for **ICLR 2025**

Awards

Lightning talk @NeurIPS 2024 MusIML workshop December 2024
Neural Information Processing Systems (NeurIPS 2024) *Vancouver, BC*
- Our work on "Leveraging Pragmatics for Retrieval Augmented Generation" is chosen as a contributed lightning talk (oral presentation) at the MusIML workshop co-located with NeurIPS 2024.

College of Science Grad Fellowship Spring 2024
University of Arizona *Tucson, AZ*

Stanford Treehacks 2024 February 2024
Stanford University *Stanford, CA*
- Received a bursary covering all associated costs to participate in the premier U.S. collegiate hackathon, Stanford Treehacks 2024.

AI Talent Bursary May 2022
Alberta Machine Intelligence Institute (AMII) *Edmonton, AB*
- Awarded $1500 CAD for attending AI week organized by Alberta Machine Intelligence Institute (AMII).

Rector's Gold Medal shortlist June 2021
National University of Sciences & Technology *Islamabad, PK*
- Among 3 people shortlisted for Rector's Gold Medal for best final year CS project.

Dean's list 2018 – 2021
National University of Sciences & Technology *Islamabad, PK*
- Dean's list in 6/8 semesters of undergraduate degree.

Turkish Aerospace Industries (TAI) Summer Program July 2020
Turkish Aerospace Academy *Ankara, Turkey*
- Nominated for a fully funded vocational summer program at Turkish Aerospace Academy based upon demonstrated aptitude in technical disciplines.

Skills

Java • Python • C • Scala • SQL • JavaScript • Algorithms • PyTorch • Machine Learning • Convolutional Neural Networks • Transformers • NLP • Image Processing • Time Series • OpenCV • Git • HTML/CSS • D3.js • Maven