Afra Feyza Akyürek

(412) 773-2342
akyurekafra@gmail.com

PhD Candidate in Computer Science

Google Scholar
Personal Webpage
GitHub: feyzaakyurek
LinkedIn: afrafeyzaakyurek

I am broadly interested in LLM post-training, evaluation and continual learning from feedback. Most recently, I worked on online reinforcement learning algorithms (OnlineRubrics and RL4F). Most recently, I have led the curation of a large-scale reasoning benchmark for professional domains at Scale AI called **P**rofessional **R**easoning Bench (PRBench).

Education

PhD in Computer Science, *Boston University* — September 2019 — March 2025
MSc in Statistics, *Carnegie Mellon University* — August 2018 — July 2019
BSc in Computer Engineering, *Koc University*, *Ranked 3rd in graduating class.* — September 2014 — July 2018

Work Experience

Research Scientist — **April 2025 — Present**
Scale AI — *San Francisco, CA*

- Post-training algorithms research and synthetic data generation.
- Advising internal human data programs.
- Leading the creation of PRBench for evaluating LLMs in professional reasoning.
- Investigating capabilities of customer LLMs and proposing data programs to address deficiencies.

Research Intern — **September 2022 — December 2022**
Allen Institute for Artificial Intelligence (AI2) — *Seattle, WA*

- Designed and implemented an external critique model trained with reinforcement learning to repair errors in LLM outputs (ACL 2023).

Machine Learning Research Intern — **June 2021 — August 2021**
Apple — *Cambridge, MA*

- Evaluating and aligning large language models for social biases (Findings of NAACL 2022).

Statistical Learning Intern — **May 2019 — July 2019**
Novartis — *East Hanover, NJ*

- Developed machine learning model to predict CAR-T therapy outcomes, accurately forecasting treatment efficacy and safety risks; findings presented to VP leadership.

Publications

- **A.F.A**, Advait Gosai, Chen Bo Calvin Zhang, et al. **PRBench: Large-Scale Expert Rubrics for Evaluating High-Stakes Professional Reasoning. Under review**, 2025

  — Large-scale evaluation of LLMS for reasoning in professional domains.

- MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, **A.F.A**. **Online Rubrics Elicitation from Pairwise Comparisons. Underreview**, 2025

  — Online estimation of rewards via synthetic rubrics.

- **A.F.A**, Ekin Akyürek, Leshem Choshen, Derry Wijaya, Jacob Andreas. **Deductive Closure Training of Language Models for Coherence, Accuracy, and Updatability. Findings of ACL**, 2024

  — Built a self-training algorithm that makes language models significantly more accurate and coherent on factual knowledge.

- **A.F.A**, Eric Pan, Garry Kuwanto, Derry Wijaya. **DUnE: Dataset for Unified Editing. EMNLP**, 2023

  — Designed a new benchmark that tests targeted model editing based on natural language instructions and preferences. I showed surgical model editing methods generalize poorly compared to simple RAG systems.

- **A.F.A**, Ekin Akyürek, Ashwin Kalyan, Peter Clark, Derry Wijaya, Niket Tandon. **RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs**. **ACL**, 2023

  — Demonstrated that a small model (weak) can be trained via RL to give language feedback to large model (strong) during test-time yielding to improvements in strong models accuracy — an early demonstration of how a two-agent system of LLMs could operate.

- **A.F.A**, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Wijaya. **Challenges in Measuring Bias via Open-Ended Language Generation**. **GeBNLP at NAACL (Oral)**, 2022

  — Identified the pitfalls in evaluating social biases in language model generations and recommended a robust evaluation scheme.

- Garry Kuwanto*, **A.F.A***, Isidora Chara Tourni*, Siyang Li*, Alex Jones, Derry Wijaya. **Low-Resource Machine Translation Training Curriculum Fit for Low-Resource Languages**. **PRICAI**, 2023
- **A.F.A**, Sejin Paik, Muhammed Yusuf Kocyigit, Seda Akbiyik, Şerife Leman Runyun, and Derry Wijaya. **On Measuring Biases in Prompt-Based Learning**. **Findings of NAACL**, 2022
- **A.F.A**, Ekin Akyürek, Derry Wijaya and Jacob Andreas. **Subspace Regularizers for Few-Shot Class Incremental Learning**. **ICLR**, 2022

  — Made image classifiers learn continually using a small set of examples and language instructions without much forgetting by a simple regularized objective that achieved SoTA.
- Ekin Akyürek, **A.F.A** and Jacob Andreas. **Learning to Recombine and Resample Data for Compositional Generalization**. **ICLR**, 2021
- Haryo Akbarianto Wibowo, Made Nindyatama Nityasya, **A.F.A**, Suci Fitriany, Alham Fikri Aji, Radityo Eko Prasojo, and Derry Tanti Wijaya. **IndoCollex: A Testbed for Morphological Transformation of Indonesian Colloquial Words**. **Findings ACL-IJCNLP**, 2021
- **A.F.A**, Lei Guo, Randa Elanwar, Margrit Betke, Prakash Ishwar and Derry T. Wijaya. **Multi-label and Multilingual News Framing Analysis**. **ACL**, 2020

  — Built a SoTA classifier to identify the ways different media outlets frame the same events across multiple languages.

Leadership, Honors and Awards

Hariri Institute for Computing, Graduate Student Fellow	May 2021 - December 2024
Best Senior Design, Self-Driving Cars in Unity with Deep Reinforcement Learning	June 2018
Hult Prize, Regional Finalist	March 2017
Society of Women Engineers, President, Koc University	June 2016 - May 2017
31^{st} **Place in National University Entrance Exam** amid 2 million takers in Turkey	2013