

SAED REZAYI, PhD

NLP RESEARCH SCIENTIST • TRUSTWORTHY EVALUATION OF LANGUAGE MODELS

saed.rezaei@gmail.com | 404-388-1414 | Media, PA, USA
saedr.github.io | linkedin.com/in/srezayi | [Google Scholar](https://scholar.google.com/citations?user=saedr)

SUMMARY

NLP research scientist working on trustworthy evaluation of language models. My research asks whether an evaluation measures what it claims to measure and treats fairness, explainability, and robustness as instances of it rather than separate topics. I build methods that separate genuine model behavior from artifacts of the evaluation itself, including cases where a widely used metric turned out to be measuring the evaluator rather than the system under test. PhD in Computer Science; 20+ peer-reviewed publications across ACL/NAACL, IJCAI, SDM, and IEEE venues; one patent pending.

EXPERIENCE

NLP Scientist — National Board of Medical Examiners (NBME)

October 2023 – Present • Philadelphia, PA

The setting for this research: automated scoring of clinical communication in physician licensure assessment, where scores carry real consequences and the measurement faces expert human validation and adversarial pressure. I own the evaluation and explainability research agenda for the deployed system.

- **Evidence sufficiency.** Developed a framework to separate missing evidence from evaluator mismatch, showing that much of the apparent insufficiency came from the evaluator rather than the evidence.
- **Explainability and evidence quality.** Tested whether evidence highlighted by a scoring model actually supported its predictions, and found that correcting evaluator effects changed which explanations appeared most reliable.
- **Evaluation of LLM-generated content.** Built a multi-stage pipeline to check generated clinical cases for format, quality, and duplication, with the same checks designed for use as training rewards.
- **Fairness in automated scoring.** Published a BEA 2026 study evaluating fairness through the different risks of false positives and false negatives rather than relying on a single disparity metric.
- **Fairness and validity.** Developed a broader framework for connecting fairness evaluations to validity claims, downstream risks, and the comparisons used to detect disparities.
- **Robustness to leakage and copying.** Developed simulation-based methods to detect test-content leakage and answer copying, including a temporal detector reaching 0.81 recall at 1% false-positive rate.
- **Research leadership.** Proposed and led an organization-wide explainability initiative covering evidence evaluation, expert validation, cross-model comparisons, and deployment guidance.

Research Intern — Thomson Reuters

June 2022 – August 2022 and January 2023 – October 2023 • Remote

- Investigated how incorporating legal-case graph structure into document representations could improve retrieval of relevant cases.

Research Intern (Machine Learning / NLP) — Adobe Research

June 2020 – August 2020 • Remote

- Built a knowledge-graph-driven framework for search query reformulation and suggestion, combining entity linking with link prediction. Published at IEEE BigData; resulting method is the subject of a pending patent.

Graduate Research Assistant — University of Georgia

September 2018 – May 2023 • Athens, GA

- Doctoral research on knowledge-infused language models: injecting external structured knowledge into pretrained encoders to improve comprehension in low-resource and specialized domains. Produced domain-adapted models for agriculture (AgriBERT, IJCAI) and clinical radiation oncology (ClinicalRadioBERT, MLMI), a knowledge-graph enrichment framework (EDGE, NAACL), and a cross-domain clustering method (XDC, SDM).

SELECTED PUBLICATIONS

Full list on [Google Scholar](#).

- Policy-Sensitive Fairness Evaluation in Automated Scoring of Clinical Communication. Rezayi, S., Ha, L. A., Yaneva, V., Harik, P., Mee, J., Snyder, J. **ACL-BEA 2026**
- Evaluating LLM Workflows for Generating Clinical Communication Assessment Items. Runyon, C., Baldwin, P., Micir, I., Frome, K., Mann, S., Rezayi, S., Evanini, K., Yaneva, V. **ACL-BEA 2026**
- Towards Reliable Generation of Clinical Chart Items: A Counterfactual Reasoning Approach. Li, J., Rezayi, S., Baldwin, P., Harik, P., Yaneva, V. **AIME-Con 2025**
- Automated Scoring of Communication Skills in Physician-Patient Interaction: Balancing Performance and Scalability. Rezayi, S., Ha, L. A., Zhou, Y., Houriet, A., D'Addario, A., Baldwin, P., Harik, P., King, A., Yaneva, V. **ACL-BEA 2025**
- Enhancing Security and Strengthening Defenses in Automated Short-Answer Grading Systems. Yarmohammadtoosky, S., Zhou, Y., Yaneva, V., Baldwin, P., Rezayi, S., Clauser, B., Harik, P. **ACL-BEA 2025**
- Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Questions. Yaneva, V., North, K., Baldwin, P., Ha, L. A., Rezayi, S., Zhou, Y., Harik, P., Clauser, B. **ACL-BEA 2024**
- AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition. Rezayi, S., et al. **IJCAI 2022**

In submission / in preparation

- Interpreting Fairness Evidence in Automated Scoring Through Validity, Risk, and Detection. **under review**
- What Does Rationale Sufficiency Measure? Separating Information Loss from Evaluation-Model Mismatch. **in preparation**

TECHNICAL SKILLS

Evaluation: Benchmark and eval-set design, faithfulness and rationale evaluation (ERASER-style sufficiency/comprehensiveness and its failure modes), perturbation and counterfactual testing, human-annotation protocol design.

Modeling: Transformer fine-tuning (DeBERTa, BERT family), token classification, knowledge distillation, RLVR and verifiable reward design, few-shot and knowledge-infused learning, synthetic data generation pipelines

Engineering: Python, PyTorch, Hugging Face Transformers, AWS SageMaker, SQL, Git, LaTeX; reproducible experiment harnesses and distributed training

EDUCATION

PhD, Computer Science — University of Georgia

2018 – 2023 · Athens, GA

Advisor: Prof. Sheng Li. Area: knowledge-infused language models for machine comprehension.

MS: Computer Science, University of Oregon, 2017

MS: Electrical Engineering, Amir Kabir University of Technology, 2013

BS: Electrical Engineering, Mazandaran University, 2010

SERVICE & RECOGNITION

Program committee: AAAI, MobiQuitous

Reviewer: ACL / ARR, EMNLP, ICML, KDD, UAI, IEEE Sensors Letters, IEEE TBioCAS

Patent: Search query reformulation with entity linking and link prediction (pending, Adobe)