

Dani Roytburg

droytbur@andrew.cmu.edu
djroytburg.github.io

November 28, 2025

Summary	Master's student in the Machine Learning Department at Carnegie Mellon University focused on reliable and interpretable AI, multi-agent training paradigms, and activation-based model analysis.		
Education			
M.S.	Carnegie Mellon University, School of Computer Science Machine Learning	2026	
B.S.	Emory University, College of Arts and Science Computer Science; Quantitative and Social Sciences	2025	
Positions	Carnegie Mellon University Pittsburgh, PA	Research Assistant, Language Technologies Institute	September 2025–present
	Develop multi-agent adversarial reinforcement learning environments to improve legibility of reasoning traces.		
Martian Research	Remote	Research Fellow	May 2025–present
	Engineer activation profiles of language model evaluators to suppress self-preferential bias; winner of a mechanistic interpretability hackathon.		
Emory University	Atlanta, GA	Research Assistant, Digital Humanities Lab	November 2021–May 2025
	Designed and trained large-scale graph-text models; evaluated adversarial robustness of LLMs; built historical correspondence networks and dashboards.		
ZS Associates	Evanston, IL	Internship - Business Technology Solutions	June 2024–August 2024
	Automated ETL pipelines and implemented synthetic-data imputation for compromised client data.		
Cloverpop	Chicago, IL	Intern — Machine Learning	January 2024–May 2024
	Deployed decision-management systems and structured-prediction models on meeting tran-		

scripts.

McCormick and Company
Hunt Valley, MD

Intern — Global Marketing and Sales Data Science
June 2023-August 2023

Built Amazon sales data visualizations and automated analytics with GCP BigQuery and BERT-based embedding search.

Northeastern University
Boston, MA

Research Assistant, Mediabase Lab
June 2022-October 2022

Extended NER to long-tail ethnic cuisines using Bon Appetit data.

Publications

Conference Papers

Dani Roytburg, Matthew Bozoukov, Hongyu Fu, Matthew Nguyen, Jou Barzdukas, and Narmeen Fatimah Oozeer. Breaking the mirror: Examining self-preference in llm evaluators through activation-based representations. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL: <https://arxiv.org/abs/2509.03647>

Dani Roytburg, Deborah Olorunisola, Sandeep Soni, and Lauren Klein. Words and action: Modeling linguistic leadership in # blacklivesmatter communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1690–1703, 2025. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/35895>

Thesis

Daniel Roytburg. Generative argument mining: Pretrained language models are argumentative text parsers. Undergraduate thesis, Emory University, 2025. URL: <https://etd.library.emory.edu/concern/etds/5t34sm11w?locale=en>

Recognition

Awards

Emory University Dean's List, 2022, 2024-2025
Winner, Martian Research Mechanistic Interpretability Hackathon, 2025

Skills

Python Java R JavaScript Typescript SQL PyTorch JAX HuggingFace Transformers scikit-learn spaCy networkx D3.js React.js GCP Docker MySQL