

# TEAGAN JOHNSON

johnsont4.github.io linkedin.com/in/johnsont4 teagan.johnson@colorado.edu

## EDUCATION

**University of Colorado, Boulder, CO**

Aug 2025 - Present

Ph.D. in Computer Science

Advised by Dr. Maria Antoniak, my research focuses on LLM training pipelines, LLM reasoning, and computational narratology.

**Carleton College, Northfield, MN**

Sep 2019 - Jun 2023

B.A. in Computer Science & Statistics

Activities/Societies: Sigma Xi Research Society, Chi Alpha Sigma NCAA Student-Athlete Society, Varsity Swim Team (captain).

## PUBLICATIONS

### · **Characterizing Narrative Content in Web-scale LLM Pretraining Data**

**Teagan Johnson**, Elliott Ash, Andrew Piper, Maria Antoniak

*Findings of Empirical Methods in Natural Language Processing (EMNLP) 2026*

### · **Can Counterfactually-Inspired Preprocessing Help Detect Polarization?**

**Teagan Johnson**

*Presented at SemEval 2026, co-located with ACL*

## RESEARCH EXPERIENCE

**College of Engineering and Applied Science, University of Colorado**

Aug 2025 - Present

*Research Assistant - Natural Language Processing under Dr. Maria Antoniak*

*Boulder, CO*

- Building scalable NLP pipelines for web-scale corpora and designing evaluation frameworks for open-ended, creative tasks.
- Implementing data mixing pipelines to evaluate LLM performance on different distributions of pretraining data. This includes frequently working with Common Crawl snapshots (250+ TB), the Dolma corpus (5+ TB), and other large-scale web-based corpora.
- Accepted to EMNLP Findings 2026 (Johnson et al. 2026) for work on web-scale narrative analysis.

**Institute for Pure and Applied Mathematics, UCLA**

Jun - Sep 2023

*Research Assistant - Multifidelity Modeling under Dr. Susana Serna*

*Los Angeles, CA*

- Built multi-fidelity surrogate models for AFRL's Boltzmann simulations of supersonic flow over a wedge, cutting runtime from 48+ hours on 16 supercomputer cores to ~15 minutes on a laptop.
- Developed and benchmarked a POD/SVD reduced-order model, an Euler-equation CFD model, and a space-mapping fusion of both, learning ~19M per-coordinate coefficients via least squares on 20 Boltzmann reference runs.
- Achieved 0.098 L1 error against Boltzmann ground truth; presented results to AFRL and at the Joint Mathematics Meetings 2024.

**Department of Mathematics & Statistics, Carleton College**

Jan - Jun 2023

*Research Assistant - Statistical Graphics under Dr. Adam Loy*

*Northfield, MN*

- Found that including scagnostics (novel scatterplot features at the time) helps determine the fit of a model along with classic statistics such as the variance and mean. Implemented random forest, generalized additive, and support vector machines.
- Our paper won the national 2023 USRESP Competition and we presented our work at the 2023 eUSR Conference.

**Department of Computer Science, Carleton College**

Sept 2022 - Jan 2023

*Research Assistant - Natural Language Processing under Dr. Anna Rafferty*

*Northfield, MN*

- Implemented a loss function that prioritizes fairness using a counterfactual fairness metric and a counterfactual data augmentation method that mitigates bias in toxicity classification.
- The loss function improved fairness by up to 53% and the data augmentation method improved fairness by up to 70% compared to baseline models. Both methods maintained high accuracy.

**Department of Computer Science, Carleton College**

Jun - Nov 2021

*Research Assistant - Software Development under Dr. Aaron Bauer*

*Northfield, MN*

- Developed an educational programming game using ReactJS and Typescript that studies how students learn coding concepts.
- Constructed a building interface and connected it to a Flask database that stores various aspects of gameplay.

## PROFESSIONAL EXPERIENCE

---

### General Motors

Aug 2023 - Aug 2025

#### Machine Learning Engineer

Austin, TX

- Saved ~100 analyst hours/week with a multi-tool agent for supplier issue triage (cross-country travel-time estimation, urgency scoring, summarization), deployed on-prem and served up to 10K requests/day.
- Shipped a GPT-4o-mini OCR-to-JSON invoice parsing pipeline with Pydantic validation in 4 weeks, reaching 89% field-level accuracy; used daily by 100+ buyers for tariff analysis.
- Built a LangGraph NL2SQL agent over GM's supply chain databases on on-prem Kubernetes, achieving 91% accuracy on a 1,000-question manually-curated eval set.
- Fine-tuned BERT-based NER and multi-label risk classifiers and built topic modeling and summarization pipelines for a news-driven supply chain risk app on Azure Kubernetes, serving hundreds of daily users and cutting manual review by ~70%.
- Owned CI/CD (GitHub Actions, Azure DevOps) and monitoring of latency and uptime for 4 production ML apps on Kubernetes.

### National Center for Atmospheric Research

May - Aug 2022

#### Software Engineer Intern

Boulder, CO

- Built an end-to-end search service with Spring Boot and Solr that integrated ingestion, indexing, and query layers for ~10K scientific records across NCAR's 7 research labs; launched to a researcher beta group at the completion of my internship.
- Designed an ingestion-time metadata validation layer to catch corrupted records before indexing, and cut the outdated-record deletion process from hours to seconds.

## AWARDS & TALKS

---

- Parallel talk at IC2S2 2026 (**top ~33% of submissions**) for my work on narratives in LLM pretraining data.
- **Beverly Sears Award (\$1,200)**, 2026. Competitive research grant supporting my work on narratives in LLM pretraining data.
- **First Place**, USRESP Competition (ASA & CAUSE), 2023. "Classification Models for Statistical Graphics."

## TEACHING & MENTORSHIP

---

### Department of Computer Science, University of Colorado, Boulder

Dec 2025 - Present

#### Teaching Assistant

Boulder, CO

- Hold office hours, design exams, and assist with additional administrative tasks for the graduate-level introduction to NLP course.

### Department of Mathematics & Statistics, Carleton College

Aug 2022 - June 2023

#### Teaching Assistant

Northfield, MN

- Assisted up to ten students per day with data science, machine learning, and Bayesian statistics courses, primarily using R.

### Career Center, Carleton College

Aug 2020 - June 2023

#### Career Counselor

Northfield, MN

- Coached up to ten students per day regarding cover letters, resumes, internships, jobs, and funding opportunities.

## TECHNICAL SKILLS

---

|                                  |                                                                                    |
|----------------------------------|------------------------------------------------------------------------------------|
| <b>Languages &amp; Libraries</b> | Python (PyTorch, numpy, transformers, CrewAI, LangGraph), R, SQL, TypeScript, Java |
| <b>ML &amp; AI Paradigms</b>     | RAG, Agentic AI, NER, Topic Modeling, post-training                                |
| <b>LLMs &amp; Agents</b>         | OpenAI, Claude, Qwen, Llama model families (enterprise & personal scale)           |
| <b>Infra &amp; Web</b>           | Kubernetes, Docker, Azure (AKS), REST APIs, React, Flask, Streamlit, Spring Boot   |