

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).

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 an **on-prem OCR-to-JSON invoice parsing pipeline** for unstructured PDF files (89% field-level accuracy); used daily by 100+ buyers for time-sensitive tariff analysis.
- 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 100+ 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 Spring Boot and Solr search service for ~10K scientific records across NCAR's 7 research labs.

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 studying LLM pre- and post-training and designing eval frameworks for open-ended tasks.
- Implementing data mixing pipelines to evaluate LLM performance on different distributions of pretraining data, ingesting Common Crawl snapshots (250+ TB), the Dolma corpus (5+ TB), and other large-scale web corpora.

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 (POD/SVD reduced-order and Euler-equation CFD) for Air Force Research Lab's Boltzmann simulations of supersonic flow, cutting runtime from 48+ hours on 16 supercomputer cores to ~15 minutes on a laptop.

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*

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."