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 the intersection of LLM pretraining data curation and narratology.

Carleton College, *Northfield, MN* Sep 2019 - Jun 2023
B.A. in Computer Science & Statistics
Activities/Societies: Sigma Xi Research Society, American Mathematics Society, Society for Industrial and Applied Mathematics, Chi Alpha Sigma NCAA Student-Athlete Society, Varsity Swim Team (captain)

PROFESSIONAL EXPERIENCE

General Motors Aug 2023 - Aug 2025
Machine Learning Engineer *Austin, TX*

- Designed a multi-tool AI agent using CrewAI to automate supplier issue triage. Developed tools for travel time estimation, urgency scoring, and summarization. Achieved 100 analyst hours saved weekly; deployed on-prem with 24/7 uptime.
- Built OCR-to-structured-data pipeline using GPT-4o-mini for tariff invoice parsing. Designed robust post-processing with Pydantic for JSON output. Delivered production-ready system in $\leq$30 days, adopted daily by hundreds of buyers.
- Developed LangGraph-based NL2SQL chatbot for querying GM's complex supply chain databases. Integrated with Streamlit UI and deployed on-prem Kubernetes. Delivered dev-ready system in 2 weeks; enabled iterative feedback from business users.
- Fine-tuned NER and multi-label classification models on custom supply chain risk datasets. Built BERT-based topic modeling and extractive summarization pipelines. Deployed containerized NLP microservices on Azure Kubernetes, reducing manual review workload by 70%.

National Center for Atmospheric Research May - Aug 2022
Software Engineer Intern *Boulder, CO*

- Developed a new search engine for NCAR using Spring Boot that enables scientists to efficiently find scientific resources produced by each of NCAR's seven research labs.
- Ensured full accuracy across 20,000+ search results by designing and implementing a robust metadata validation process.
- Improved the efficiency of metadata file deletion by up to 99%, allowing scientists to quickly delete outdated and incorrect files.

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 non-traditional evaluation frameworks for open-ended tasks.
- Implementing language model remixing pipelines to evaluate LLM performance on different mixes of pretraining data.
- Analyzed $\sim$3M documents within Dolma (an open-source pretraining dataset) using custom theory-inspired narrative features.

Institute for Pure and Applied Mathematics, UCLA Jun - Sep 2023
Research Assistant - Multifidelity Modeling under Dr. Susana Serna *Los Angeles, CA*

- Worked with the Air Force Research Laboratory (AFRL) to replace high-fidelity models that simulate the flow of supersonic air over a wedge with a faster, comparably accurate multi-fidelity model, resulting in 99% faster runtime.
- Combined two lower-fidelity models that utilized dimension reduction techniques (proper orthogonal and singular value decomposition) and numerically solved fluid dynamics equations into a more accurate surrogate model.
- Reduced the runtime of AFRL's supersonic airflow modeling approach from 48 hours to 13 minutes.

Department of Mathematics & Statistics, Carleton College Jan - Jun 2023
Research Assistant - Statistical Graphics under Dr. Adam Loy *Northfield, MN*

- Determined that scagnostics are up to 80% more effective at determining the fit of a model than classic statistics such as the variance and mean by implementing random forest, generalized additive, and support vector models.
- 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.

PUBLICATIONS

- **Characterizing Narrative Content in Web-scale LLM Pretraining Data**
  Teagan Johnson, Elliot Ash, Andrew Piper, Maria Antoniak
  Preprint, under review (2026).
- **Can Counterfactually-Inspired Preprocessing Help Detect Polarization?**
  Teagan Johnson
  Presented at SemEval 2026, co-hosted at ACL

AWARDS

- Awarded a **parallel talk session** at the 12th International Conference on Computational Social Science (IC2S2 2026) 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

Languages	Python, R, SQL, TypeScript, Java
ML & NLP	PyTorch, BERT, Transformers, NER, Text Classification, Topic Modeling, word2vec
LLMs & Agents	CrewAI, LangGraph, GPT-4o-mini, RAG, Prompt Engineering, Pydantic
Infra & Web	Kubernetes, Docker, Azure (AKS), REST APIs, React, Flask, Streamlit, Spring Boot