# DAVID HEINEMAN

davidheineman.com
davidh@allenai.org
linkedin.com/in/davidheineman

## EDUCATION

**Georgia Institute of Technology** 2020 - 2024
*B.S. in Computer Science*. GPA: 3.9, Highest Honors.
*Undergraduate Thesis:* Edit-based Language Model Evaluation
*Coursework:* Low-level Computer Architecture and Systems, Information Security, Deep Learning, Database Implementation

## RESEARCH EXPERIENCE

**Allen Institute for AI** | *Predoctoral Young Investigator, AllenNLP. Advisor: Jesse Dodge, Kyle Lo* 2024 - Present

- Built evaluation, methodology, scaling details and implementation for OLMo — a fully **O**pen **L**anguage **Mo**del.
- Performed the first large-scale analysis of language model pre-training benchmarks to analyze modeling noise on the validity of experiment results, designed new methodology for reducing experimental noise in LLM pre-training
- Contributed to Fluid Benchmarking a novel method which dynamically selects items with the highest Fisher information for a specific ability level. Fluid benchmarking retains the full statistical power of a benchmark with 50x fewer eval instances
- Help build task scaling laws to predict downstream performance (e.g. MMLU) for multi-million dollar model training runs using models trained with < 1% of the compute of the large model
- Trained and released DataDecide, a suite of 225 LLMs from 4M to 1B parameters for studying pairwise data decisions using small compute budgets. Introduced perplexity-based metrics to run data ablations for capabilities that were otherwise 'emergent'
- Led work to publication at NeurIPS, co-authored publications at COLM, ICML and in submission to ICLR

**Georgia Tech NLP X Lab** | *Undergraduate Research Assistant. Advisor: Wei Xu* 2021 - 2024

- Created flexible decoding library for test-time scaling using MBR decoding: swapping out datasets, metrics and managing experiments, including custom multi-GPU distributed inference with open-source LLMs
- Performed the first large scale fine-grained text evaluation for LLMs to exhaustively consider *every* linguistic transformation performed by the LLM. Built an open-source library for analysis, modeling and agreement calculations for fine-grained text generation, including a novel span-based evaluation metric.
- Help build the `lens-metric` library to load automatic simplification metrics directly from Hugging Face
- Built a multi-stage highlighting, ranking and rating interface for text generation, orchestrated large-scale in-house human evaluation across 26 text simplification systems used to develop state-of-the-art simplification metrics
- Implemented distributed MapReduce-like algorithms for large $n$-gram datasets for real-time paraphrase generation
- Led work to publication at EMNLP, CogSci, co-authored publications at ACL and COLM
- Led discussions of bi-weekly undergraduate reading group (Spring 2021, Fall 2021)

**Service:** Reviewer at NeurIPS 2024, TSAR 2022. Panel speaker at TSAR 2024 (co-located at EMNLP 2024)

## INDUSTRY EXPERIENCE

**Amazon Web Services** | *Software Engineering Intern, EC2* Summer 2023

- Built prototype LLM agent system to autonomously debug and solve CloudWatch alarms
- Used parameter efficient training (LoRA) to fine-tune LLaMA using AWS help forums as data for a preference optimization
- Created first internal deployment of a fine-tuned open-source 40B LLM in AWS Sagemaker, contributed to `bitsandbytes` to fix GPU compatibility errors when performing quantization in EC2 instances with NVIDIA A100s
- Proposed new method of running asynchronous LLM queries in parallel to pull and consolidate account information which improved retrieval latency by 5x, reducing wait time for responses.
- Built internal manual inspection dashboard to evaluate performance and an automatic classifier for harmful responses

**Amazon Web Services** | *Software Engineering Intern, CloudWatch* Summer 2022

- Developed a new feature for CloudWatch Application Insights to monitor processes running on EC2 instances
- Worked with EC2 Windows experts to identify breakpoints for customers' Microsoft SQL and .NET workloads, predict these common errors using CloudWatch metrics and provide insights to reduce problem resolution time
- On-boarded existing AWS customers like Koch Industries, Moody's to monitor processes on critical Windows applications

**Patientco** | *Software Engineering Intern, ML* Summer 2021

- Created a large-scale data ingestion pipeline for training ML models to predict the likelihood of bill payments
- Used AWS Sagemaker to productionalize models to re-train and label 500,000 unique bills daily (5% of U.S. healthcare bills)
- Developed a novel sequence-based approach for detecting anomalies in insurance charges, used by providers like Piedmont to identify and fix errors in adjudication for patient bills

**Open-source Contributions:** `thresh`, `bitsandbytes`, `vLLM`, `OLMo` / `olmes`, `terminal-bench`, `harbor`

## PUBLICATIONS

**Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces**
Mike A Merrill, Alexander Glenn Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, ..., **David Heineman**, ..., Ryan Marten, Yixin Wang, Alex Dimakis, Andy Konwinski, Ludwig Schmidt.
*preprint*

**Signal and Noise: A Framework for Reducing Uncertainty in Language Model Evaluation**
**David Heineman**, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, Jesse Dodge.
*NeurIPS, 2025 (Spotlight, Top 5%)*

**Fluid Language Model Benchmarking**
Valentin Hofmann, **David Heineman**, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, Noah A. Smith.
*COLM, 2025 (Oral, Top 5%)*

**Establishing Task Scaling Laws via Compute-Efficient Model Ladders**
Akshita Bhagia*, Jiacheng Liu*, Alexander Wettig, **David Heineman**, Oyvind Tafjord, Ananya Harsh Jha, Luca Soldaini, Noah A. Smith, Dirk Groeneveld, Pang Wei Koh, Jesse Dodge, Hannaneh Hajishirzi.
*COLM, 2025*

**Evaluating LLMs on Chinese Idiom Translation**
Cai Yang, Yao Dou, **David Heineman**, Xiaofeng Wu, Wei Xu.
*COLM, 2025*

**2 OLMo 2 Furious**
Pete Walsh*, Luca Soldaini*, Dirk Groeneveld*, Kyle Lo*, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, ..., **David Heineman**, ..., Ali Farhadi, Noah A. Smith, Hannaneh Hajishirzi.
*COLM, 2025*

**DataDecide: How to Predict Best Pretraining Data with Small Experiments**
Ian Magnusson*, Nguyen Tai*, Ben Bogin*, **David Heineman**, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, Jesse Dodge.
*ICML, 2025*

**Improving Minimum Bayes Risk Decoding with Multi-Prompt**
**David Heineman**, Yao Dou, Wei Xu.
*EMNLP, 2024*

**Towards a Path Dependent Account of Category Fluency**
**David Heineman**, Reba Koenen, Sashank Varma.
*CogSci, 2024*

**Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation**
**David Heineman**, Yao Dou, Wei Xu.
*EMNLP: System Demonstrations, 2023*

**Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA**
**David Heineman**, Yao Dou, Mounica Maddela, Wei Xu.
*EMNLP, 2023*

**LENS: A Learnable Evaluation Metric for Text Simplification**
Mounica Maddela*, Yao Dou*, **David Heineman**, Wei Xu.
*ACL, 2023*

## HONORS & AWARDS

- NSF CS Graduate Fellowship (3-year support of $159,000; 5th cohort of research fellows)
- Georgia Tech College of Computing (CoC) Outstanding Undergraduate Research Award (1 of 3000+ CS undergraduates)
- Zell Miller Scholarship ($40,000; Full Undergraduate Tuition)
- 1st place in Listen Labs Berghain Challenge (1 of 1300 entrants)
- Entrepreneur First residency ($12,000 stipend; 1st San Francisco cohort)
- President's Undergraduate Research Award ($1,500; PURA Research Grant)
- 2nd place in AGI House Open Source AI Hackathon ($1,000; from Prime Intellect)
- Tinker Beta User ($150 in credits; from Thinking Machines)

## TEACHING

(CS 4650) **Natural Language Processing** Fall 2023
*Teaching Assistant (TA)* | *Instructor: Alan Ritter*

(CS 3510) **Design & Analysis of Algorithms** Fall 2022
*Teaching Assistant (TA)* | *Instructor: Gerandy Brito & Dana Randall*

(CS 3510) **Design & Analysis of Algorithms** Fall 2021
*Teaching Assistant (TA)* | *Instructor: Jacob Abernathy*