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

RESEARCH EXPERIENCE

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

- Built new evaluation benchmarks, methodology 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 sources of modeling noise on the validity of experiment results, designed new metrics for experiments with <1% compute compared to the large model
- Technical contributions to projects fitting scaling laws, fast benchmarking, data mixing, mixture-of-experts LLM training
- Co-author on work to appear at ICML, work COLM, and led work currently under review at NeurIPS

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

- **Controllable Text Diffusion:** Managed large scale training of text diffusion models on GPU clusters to replicate controllable text diffusion experiments. Developed a prototype controllable diffusion method for sequence-level control at decoding time by using an automatic metric as part of the control signal
- **Minimum Bayes Risk Prompting:** Created flexible decoding library for swapping out datasets, metrics and managing experiments, including custom multi-GPU distributed inference with open-source LLMs
- **Fine-grained Text Generation Analysis:** Performed the first large scale fine-grained text evaluation 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. Contributed to the `lens-metric` library to load automatic simplification metrics directly from Huggingface
- **Rank & Rate Evaluation:** 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
- 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)

PUBLICATIONS

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.
under review at NeurIPS, 2025

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

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

INDUSTRY EXPERIENCE

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

- Built prototype LLM application to query AWS documentation, account knowledge to solve root causes of CloudWatch alarms
- Used parameter efficient training (i.e., QLoRA) to fine-tune Falcon, LLaMA to align with developer-friendly responses
- Created first internal deployment of fine-tuned open-source 40B LLM on a Sagemaker endpoint, contributed to `bitsandbytes` to fix a GPU compatibility error when performing quantization in the `G4` class of EC2 instances
- Proposed new method of chaining T5 queries in parallel to explore account details in a graph-like structure which improved generation pre-processing time from 30s to 5s
- Built an internal manual inspection scheme to evaluate performance and automatic identification system 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

HONORS & AWARDS

- NSF CS4GradsUS Fellowship ($159,000; 5th cohort of research fellows)
- Georgia Tech College of Computing (CoC) Outstanding Undergraduate Research Award (1 of 3000+ CS undergraduates)
- 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)

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*

OPEN SOURCE CONTRIBUTIONS

Thresh | *thresh.tools* | *github.com/davidheineman/thresh* 2023

- Created a highly customizable in-browser span alignment and annotation builder, with questions rendered in a recursive tree structure. Thresh currently supports 13 languages, paragraph annotation and a variety of question and span-selection options
- Implemented 12 well known fine-grained annotation interfaces into standardized YML templates and easy-to-load JSON datasets
- Built an integrated Python library to load arbitrary data collected with Thresh and manage large-scale annotation jobs

Hashtag Segmentation API | *github.com/davidheineman/hashtag-master-web-demo* 2022

- Deployed real-time inference for neural text segmentation, allowing projects to query an API for integration in future work
- Created interactive Vue app to demonstrate the hashtag segmentation task and new methodology

Individual Contributions: `bitsandbytes`, `vLLM`, `OLMo-core`