DAVID HEINEMAN

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

EDUCATION

Georgia Institute of Technology 2020 - 2024
B.S. Computer Science. GPA: 3.9

Teaching Assistant. Natural Language Processing, CS 4650 (Fall 2023)
Teaching Assistant. Design & Analysis of Algorithms, CS 3510 (Fall 2021 & 2022)
Awards: College of Computing Outstanding Undergrad Research Award, President's Undergrad Research Grant

EXPERIENCE

Allen Institute for AI | *Pre-doctoral Young Investigator, Advisor: Dr. Jesse Dodge & Prof. Hannaneh Hajishirzi* 2024 - Present
- *Ongoing work with the AllenNLP team on LLM training and evaluation.*

Amazon Web Services, EC2 | *Software Engineering Intern* Summer 2023
- Built prototype LLM service to query AWS documentation, account knowledge to solve root causes of CloudWatch alarms
- 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.
- Implemented novel method of chaining LLM queries in parallel to explore AWS account details in a graph-like structure which improved generation pre-processing time from 30s to 5s.
- Used parameter efficient fine-tuning (LoRA) to adapt proprietary LLMs to developer-friendly responses using internal AWS data.

Georgia Tech NLPx Lab | *NLP Research Assistant, Advisor: Prof. Wei Xu* 2021 - 2024
- *Decoding Methods for LLMs:* Built high-performance decoding library for evaluation across 25+ open-source LLMs on 12 metrics, including custom multi-GPU distributed inference with up to 50 NVIDIA A40 GPUs simultaneously.
- *Fine-grained Evaluation for LLMs:* Performed the first large scale fine-grained text evaluation to exhaustively consider *every* linguistic transformation performed by an LLM. Built an open-source library (`thresh.tools`) for analysis, modeling and agreement calculations for 10 fine-grained text generation studies, including a novel span-based evaluation metric.
- *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.
- Implemented distributed MapReduce-like algorithms for large n-gram datasets for real-time paraphrase generation.

Amazon Web Services, CloudWatch | *Software Engineering Intern* 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 workloads
- **Tools:** *Java, TypeScript, React.js, Kubernetes, PowerShell, Lambda, EC2, DynamoDB, internal tools*

Patientco | *Machine Learning Engineering Intern* Summer 2021
- Created a large-scale data ingestion pipeline 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
- **Tools:** *PyTorch, Jenkins, AWS Sagemaker, Docker, SQL, Apache Airflow*

PUBLICATIONS

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

Towards a Path Dependent Account of Category Fluency
David Heineman, Reba Koenen, Shashak Varma. *Proceedings of CogSci, 2024 (Oral)*

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

Edit-level Simplification Evaluation using SALSA
David Heineman, Yao Dou, Mounica Maddela, Wei Xu. *Proceedings of EMNLP, 2023*

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

SKILLS

Languages: Python • C++ • Go • Rust • CUDA • TypeScript • JavaScript • C • SQL • Java • Swift • OCaml

Frameworks: PyTorch • NumPy • Fairseq • NLTK • SciKit • HuggingFace

Tools: Git • PostgreSQL • Redis • Ansible • Docker • Kubernetes • UNIX • Jenkins • AWS • Apache Spark

Web: React • Node • Vue • Angular • Tailwind • D3 • Flask • Django

Coursework: Computer Architecture • Information Security • Machine Learning • Database Implementation