DAVID HEINEMAN

davidheineman.com
david.heineman@gatech.edu
linkedin.com/in/david-heineman

EDUCATION

Georgia Institute of Technology 2024 - 2025 (Expected)
M.S. in Computer Science
Georgia Institute of Technology 2020 - 2024 (Expected)
B.S. in Computer Science

GPA: *3.91* | **Department GPA:** *4.0*, Faculty Honors
Teaching Assistant for Design & Analysis of Algorithms (CS 3510, Fall 2021 & 2022)
Teaching Assistant for Natural Language Processing (CS 4650, Fall 2023)

EXPERIENCE

Data Science Intern | *Amazon Web Services, EC2* Summer 2023

- Built prototype LLM application 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.
- 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
- Used parameter efficient training (i.e., QLoRA) to fine-tune Falcon, LLaMA to align with developer-friendly responses
- **Tools:** *PyTorch, CUDA, HuggingFace TGI, AWS Sagemaker, Docker, Huggingface TRL, internal tools*

Research Assistant | *Georgia Tech NLPx Lab, Advisor: Prof. Wei Xu* 2021 - Present

- **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 for text generation evaluation with LLMs, 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 an LLM. Built an open-source library for analysis, modeling and agreement calculations for fine-grained text generation, including a novel span-based evaluation metric.

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

Software Engineering Intern | *Patientco* 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:** *Tensorflow, Keras, Python, AWS Sagemaker, Horvorod, Docker, SQL, Apache Airflow*

PUBLICATIONS

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*

PROJECTS

- **Huggingface Decoding Visualizer (2023)** - Added visualizer of well known language model decoding algorithms (e.g., sampling, beam-search) for any HF model
- **Hashtag Segmentation API (2022)** - Deployed real-time model inference for segmentation. Created interactive Vue app to demonstrate hashtag segmentation

SKILLS

Languages: Python • C++ • TypeScript • JavaScript • C • Git • SQL • Java

ML Frameworks: PyTorch • CUDA • NumPy • TensorFlow • Keras • Pandas • Fairseq • NLTK • SciKit • Huggingface

Tools: PostgreSQL • DynamoDB • Redis • Ansible • Docker • Kubernetes • UNIX • Jenkins • AWS • Apache Airflow

Web / Mobile: React • Node • Vue • Tailwind • D3 • Flask • Django

Coursework: Low-level Computer Architecture • Information Security • Machine Learning • Database Implementation