HANNAH BROWN

hsbrown@comp.nus.edu.sg ◇ github.com/hannah-aught

RESEARCH INTERESTS

My research interests lie in trustworthy NLP. Specifically, I'm interested in measuring and improving the fairness and privacy of large language models. Within this area I focus on language generation tasks where models have much more freedom in their outputs and may unexpectedly generate biased/private information.

PROJECTS

Web Agent Benchmark for LLMs, NUS — Aug. 2023 - Present
PI: Kenji Kawaguchi

- Designed realistic and challenging web-related tasks to benchmark LLMs used as agents to complete these tasks
- Designed tasks and traps to test for unsafe behaviors in LLM agents
- Curated data from existing and synthetically generated datasets to create data for these tasks
- Assisted with design and testing of LLM agent benchmarks

Fairness in Automatic Summarization — Sep. 2021 - Present
PI: Reza Shokri

- Designed experiments for measuring different types of bias in automatically generated summaries as compared to their source articles.
- Identified methods to identify the groups discussed in documents, and where in an original article this information appeared.
- Generated summaries from various extractive and abstractive summarizers on the CNN/DailyMail dataset.
- Measured the effect of perturbations to the original articles on generated summaries.

Privacy of Language Models — Sep. 2021 - Jan. 2022
PI: Reza Shokri

- Assisted in writing a paper discussing the privacy concerns represented by language models for submission to FAccT 2022.
- Collected examples of privacy violating that from the Enron email dataset.

Applying NLP to Source Code, UC Davis — July 2020 - July 2021
PI: Prem Devanbu

- Wrote scripts for data collection and analysis of Java source code sourced from Github and SonarCloud.
- Built PyTorch models for classification of static analysis issues collected from SonarCloud.
- Modified CodeSearchNet source code to allow for use of pretrained Word2Vec and FastText embeddings instead of their embedding layer.
- Trained Word2Vec and FastText models on a corpus of Java source code.
- Designed experiments to test the stability of word embeddings from these models dependent on features of the training corpus.

Comparing SAT-Solving to ILP for Computational Biology, UC Davis — June 2019 - June 2020
PI: Dan Gusfield

- Converted ILP formulations tio SAT formulations for two problems in computational biology.
- Wrote python scripts to compare the speed of a SAT solver to that of an ILP solver for each problem.
- Designed experiments to gauge performance of each solver.
- Assisted in writing and submitting paper on our results.

EDUCATION

PhD Student, Computer Science, National University of Singapore — Aug. 2021 - Present
Advisor: Kenji Kawaguchi
Research Focus: Privacy and fairness in natural language processing.
GPA: 4.75/5.0

BAS, Computer Science and Linguistics (Honors), University of California, Davis — Sep. 2018 - June 2021
GPA: 4.0/4.0

AAS, Mathematics and Spanish, Lake Tahoe Community College — Sept. 2015 - June 2018
GPA: 4.0/4.0

PUBLICATIONS

Towards Regulatable AI Systems: Technical Gaps and Policy Opportunities [Paper]
Xudong Shen, **Hannah Brown**, Jiashu Tao, Martin Strobel, Yao Tong, Akshay Narayan, Harold Soh, Finale Doshi-Velez
arXiv Preprint (under submission), 2023

What Does it Mean for a Language Model to Preserve Privacy? [Paper]
Hannah Brown*, Katherine Lee*, Fatemehsadat Mireshghallah*, Reza Shokri*, Florian Tramèr* [Presentation]
FAccT, 2022

Unified SAT-Solving for Hard Problems of Phylogenetic Network Construction [Paper]
Dan Gusfield, **Hannah Brown**
ICCABS, 2021

Comparing Integer Linear Programming to SAT-Solving for Hard Problems in Computational and Systems Biology [Paper]
Hannah Brown, Lei Zuo, Dan Gusfield [Presentation]
AlCoB, 2020

TEACHING

TA - AI Planning and Decision Making (NUS CS5446/CS4246) — Spring 2023, Fall 2023

TA - Trustworthy Machine Learning (NUS CS5562) — Fall 2022

AWARDS AND ACHIEVEMENTS

President's Graduate Fellowship, National University of Singapore — Aug. 2021
Awarded to full-time PhD students who show exceptional promise or accomplishment in research.

Dean's Honors, UC Davis — Sept. 2018-June 2021
Awarded each quarter to full-time students with GPAs in the 12% of their major.

American Association of Unviersity Women Scholarship, AAUW — June 2018
Awarded to non-male students attending community college with the intention to transfer to a university.

CMC^3 **Scholarship**, California Math Council Community Colleges — June 2018
Awarded to qualified and deserving California Community College students who demonstrate promise and interest in the areas of Mathematics and Mathematics Education.

*Equal contribution