HANNAH BROWN

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

RESEARCH INTERESTS

My research interests lie in trustworthy NLP and regulation of AI. Specifically, I'm interested in measuring and improving the robustness and privacy of large language models and ensuring that technical definitions of these concepts align with social perspectives. I focus on language generation tasks where models have much more freedom in their outputs and may unexpectedly generate biased, private, or dangerous information.

PROJECTS

Robustness in LLMs, NUS Jan. 2024 - Present
PI: Kenji Kawaguchi and Michael Shieh

- Designed and tested an effective self-evaluation based defense for adversarial attacks on LLMs that is highly effective against unsafe instructions and adversarial attacks
- Discovered a simple character-addition attack on LLMs and explored what makes this attack effective
- Explored how well pre-trained language models can be aligned for safety using only prompts and no further fine-tuning, discovering that pre-trained LLMs can be more robust to attacks and safety instructions than previoiusly thought
- Assisted in writing three papers on this topic, resulting in one publishe in AAAI '24, one under submission at IJCAI '25, and one under submission at ACL '25.

Web Agent Benchmark for LLMs, NUS Aug. 2023 - June. 2024
PI: Kenji Kawaguchi and Michael Shieh

- 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 - Sep. 2023
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 published in FAccT 2022.
- Developed a framework for what privacy preservation in natural language processing should consider and how this differs from considerations in traditional ML.
- Collected examples of privacy violating data 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.

EDUCATION

PhD Student, Computer Science, National University of Singapore Aug. 2021 - Present
Advisors: Kenji Kawaguchi and Michael Shieh)
Research Focus: Trustworthy 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

SafePrompt: Safer Pre-Trained Models with Only Prompts [Paper]
Hannah Brown, Kenji Kawaguchi, Michael Shieh
Pre-Print, 2025 (Under Review)

Single Character Perturbations Break LLM Alignment [Paper]
Leon Lin*, **Hannah Brown***, Kenji Kawaguchi, Michael Shieh
AAAI, 2025

Self-Evaluation as a Defense Against Adversarial Attacks on LLMs [Paper]
Hannah Brown*, Leon Lin*, Kenji Kawaguchi, Michael Shieh
Pre-Print, 2024 (Under Review)

Prompt Optimization Via Adversarial In-Context Learning [Paper]
Do Long*, Yiran Zhao*, **Hannah Brown***, Yuxi Xie, James Zhao, Nancy Chen, Kenji Kawaguchi, Michael Shieh, Junxian He
ACL, 2024

Can AI be as Creative as Humans? [Paper]
Haonan Wang, James Zou, Michael Mozer, Anirudh Goyal, Alex Lamb, Linjun Zhang, Weijie J Su, Zhun Deng, Michael Qizhe Xie, **Hannah Brown**, Kenji Kawaguchi
Pre-Print, 2024

Directions of Technical Innovation for Regulatable AI Systems [Paper]
Xudong Shen, **Hannah Brown**, Jiashu Tao, Martin Strobel, Yao Tong, Akshay Narayan, [Extended Preprint]
Harold Soh, Finale Doshi-Velez
Communications of the ACM, 2024

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

*Equal contribution

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

Teaching Assistant - AI Planning and Decision Making (NUS CS5446/CS4246) Spring 2023, Fall 2023, Spring 2024

Teaching Assistant - Trustworthy Machine Learning (NUS CS5562) Fall 2022

SERVICE

Reviewer - AAAI 2025

Reviewer - ACL 2023

Reviewer - EMNLP 2023

Reviewer - EMNLP Industry Track 2023, 2024

Reviewer - FAccT 2022

Reviewer - GenLaw Workshop 2023, 2024

Reviewer - ICLR 2024, 2025

Reviewer - ICML 2024, 2025

Reviewer - IJCAI 2025

Reviewer - NAACL Industry Track 2025

Reviewer - NeurIPS 2023, 2024, 2025

Ethics Reviewer - NeurIPS 2023, 2024

Volunteer - ACL 2024

AWARDS AND ACHIEVEMENTS

Top 30% reviewer, NeurIPS 2024 Dec. 2024 Awarded to reviewers judged to be in the top 30% of all reviewers by review quality by program chairs.

President's Graduate Fellowship, National University of Singapore Aug. 2021 - Present
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.