Pengrun Huang

La Jolla, CA
peh006@ucsd.edu Website Google Scholar GitHub

EDUCATION

University of California, San Diego **La Jolla, CA**
Ph.D. in Computer Science and Engineering 09/2023 – 09/2027 (expected)
Advisors: Kamalika Chaudhuri and Yu-Xiang Wang
Research Interests: LLM Privacy and Confidentiality; Privacy Attacks and Defenses; LLM Evaluation and Benchmarking; Training-Data Auditing; Watermarking; Statistical Machine Learning

University of Michigan – Ann Arbor **Ann Arbor, Michigan**
Master of Science - Mathematics 01/2022 – 12/2022
Bachelor of Science - Honors Mathematics 09/2019 – 05/2022

PUBLICATION & PREPRINT

InferPriv: Evaluating Inferential Privacy Leakage in LLM Assistants
Pengrun Huang, Yiwei Fu, Luoxi Meng, Kamalika Chaudhuri.

Alignment Defends LLMs from Property Inference Attacks
Pengrun Huang, Chhavi Yadav, Ruihan Wu, Kamalika Chaudhuri. Preprint, 2026.

Dataset Watermarking with Provable Black-Box Detection
Pengrun Huang, Kamalika Chaudhuri, Yu-Xiang Wang. Preprint, 2026.

DPrivBench: Benchmarking LLMs' Reasoning for Differential Privacy
Erchi Wang, **Pengrun Huang**, Eli Chien, Om Thakkar, Kamalika Chaudhuri, Yu-Xiang Wang, Ruihan Wu. NeurIPS 2026 (Evaluation & Dataset).

Can We Infer Confidential Properties of Training Data from LLMs?
Pengrun Huang, Chhavi Yadav, Kamalika Chaudhuri, Ruihan Wu. NeurIPS 2025, spotlight(3%)

Better Membership Inference Privacy Measurement through Discrepancy
Ruihan Wu*, **Pengrun Huang***, Kamalika Chaudhuri. ArXiv 2024.

Conditional differential measurement error: partial identifiability and estimation
Pengrun Huang, Maggie Makar. NeurIPS 2022, workshop on causal machine learning with real-world impact.

RESEARCH EXPERIENCE

Property Inference Attacks and Defenses for LLMs

- Introduced PropInfer, a benchmark task for evaluating whether confidential dataset-level properties can be inferred from fine-tuned LLMs.
- Designed a generation-based attack for black-box model access and a shadow-model attack for grey-box access, revealing that fine-tuned LLMs can leak confidential dataset-level properties of their training data.
- Applied DPO and GRPO for LLM alignment, developing post-training defenses that mitigate property inference attacks while preserving model utility and requiring no access to or modification of original training data.

Dataset Watermarking with Provable Black-Box Detection

- Developed a dataset watermarking framework for detecting benchmark contamination and unauthorized data use using only text generated by closed-model APIs.
- Established a distribution-free, model-agnostic statistical test with provable false-positive control.
- Demonstrated reliable detection even when watermarked data comprises only approximately 1% of fine-tuning tokens, while preserving benchmark utility and semantic integrity across multiple LLM families and benchmark datasets.

TECHNICAL SKILLS

Languages: Python, C++, MATLAB, LaTeX
Machine Learning/LLM: Hugging Face Transformers, TRL,PEFT, vLLM, scikit-learn, NumPy, PyTorch.
Tools: Git, Linux, Weights & Biases

AWARDS

Jacobs School of Engineering Fellowship, UC San Diego 2023-2024
University Honors, University of Michigan 2019-2022
Chancellor's Reward Scholarship in University of Massachusetts-Amherst 2018-2019