Boyi Wei

312 Sherrerd Hall, Princeton, NJ 08540
+1 (949) 678-3985 wby@princeton.edu boyiwei.com github.com/boyiwei

Education

Princeton University **August 2023 – July 2028(expected)**
Doctor of Philosophy in Electrical and Computer Engineering *Princeton, NJ*
Advisor: Peter Henderson

Princeton University **August 2023 – April 2025**
Master of Arts in Electrical and Computer Engineering *Princeton, NJ*

University of Science and Technology of China **September 2019 – July 2023**
Bachelor of Science in Applied Physics (Summa Cum Laude) *Hefei, Anhui, China*
GPA: 4.00/4.30 (Top 3%)

Research Interest

My research focuses on building reliable and trustworthy language systems, including:

- Alignment in language models, especially safety alignment.
- Interpretation, understanding, and mitigation of law/policy issues for language models.

Publications and Preprints

1. **Boyi Wei***, Kaixuan Huang*, Yangsibo Huang*, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *ICLR 2024 SeT-LLM (**Best Paper**)/ ICML 2024.*
2. **Boyi Wei***, Weijia Shi*, Yangsibo Huang*, Noah A.Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, Peter Henderson. Evaluating Copyright Takedown Methods for Language Models. *NeurIPS 2024 Datasets and Benchmarks.*
3. Xiangyu Qi*, **Boyi Wei***, Nicolas Carnili, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, Peter Henderson. On Evaluating the Durability of Safeguards for Open-Weight LLMs. *ICLR 2025.*
4. Jakub Łucki, **Boyi Wei**, Yangsibo Huang, Peter Henderson, Florian Tramèr, Javier Rando. An Adversarial Perspective on Machine Unlearning for AI Safety. *NeurIPS 2024 SoLaR Workshop (**Best Paper**) / TMLR.*
5. Tinghao Xie*, Xiangyu Qi*, Yi Zeng*, Yangsibo Huang*, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, **Boyi Wei**, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, Prateek Mittal. SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal Behaviors. *ICLR 2025.*
6. Xiangyu Qi, Yangsibo Huang, Yi Zeng, Edoardo Debenedetti, Jonas Geiping, Luxi He, Kaixuan Huang, Udari Madhushani Sehwag, Vikash Sehwag, Weijia Shi, **Boyi Wei**, Tinghao Xie, Danqi Chen, Pin-Yu Chen, Jeffrey Ding, Ruoxi Jia, Jiaqi Ma, Arvind Narayanan, Weijie J. Su, Mengdi Wang, Chaowei Xiao, Bo Li, Dawn Song, Peter Henderson, Prateek Mittal. AI Risk Management Should Incorporate Both Safety And Security. *arXiv preprint:2405.19524 (2024).*

Experience

Georgia Institute of Technology **January 2022 – November 2022**
Research Intern (Advisor: Cong Hao) *Atlanta, GA*

University of California, Irvine **March 2023 – June 2023**
Research Intern (Advisor: Sitao Huang) *Irvine, CA*

Teaching

COS 568: Systems and Machine Learning **Spring 2025**

Honors and Awards

Francis Robbins Upton Fellowship **September 2023**
National Scholarship **September 2021, September 2022**
USTC Outstanding Student Scholarship (Gold Prize) **October 2021**
Yan Jici Outstanding Student Scholarship **November 2021**

Services

Reviews: ICLR 2025, NeurIPS 2025, SeT-LLM Workshop (ICLR 2024), SoLaR Workshop (NeurIPS 2024), L2M2 Workshop (ACL 2025)
Tutorials: LLMs and Copyright Risks: Benchmarks and Mitigation Approaches (AAAI 2025 / NAACL 2025)