Jeffrey W. Li

jwl2162@cs.washington.edu | https://jeffreywpli.github.io/ | google.scholar.com

INTERESTS

Data-centric ML, (Large-scale) Data Selection and Curation, Weak Supervision

EDUCATION

University of Washington Seattle, WA
PhD in Computer Science *Expected December 2025*
Advisors: Ludwig Schmidt, Alexander Ratner

Carnegie Mellon University Pittsburgh, PA
M.S. in Machine Learning *December 2019*
Advisor: Ameet Talwalkar

Columbia University in the City of New York New York, NY
B.S. in Applied Mathematics *May 2018*
Minors in Computer Science and Economics
Magna Cum Laude

CONFERENCE PAPERS

Jeffrey Li*, Mohammadreza Armandpour*, Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, Fartash Faghri *TiC-LM: A Web-Scale Benchmark for Time-Continual LLM Pretraining.* ACL 2025 Main Track. https://arxiv.org/abs/2504.02107

Samir Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, **Jeffrey Li**, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Jenia Jitsev, Alexandros Dimakis, Gabriel Ilharco, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, Ludwig Schmidt. *Language models scale reliably with over-training and on downstream tasks.* ICLR 2025. https://arxiv.org/abs/2403.08540

Jeffrey Li*, Alex Fang*, Georgios Smyrnis*, Maor Ivgi*, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, Vaishaal Shankar. *DataComp-LM: In search of the next generation of training sets for language models.* To appear in NeurIPS 2024: D&B. https://arxiv.org/abs/2406.11794

Tianyi Zhang, Linrong Cai, **Jeffrey Li**, Nicholas Roberts, Neel Guha, Frederic Sala. *Stronger Than You Think: Benchmarking Weak Supervision on Realistic Tasks.* To appear in NeurIPS 2024: D&B. https://arxiv.org/abs/2406.11794

Thao Nguyen, **Jeffrey Li**, Sewoong Oh, Ludwig Schmidt, Jason Weston, Luke Zettlemoyer, Xian Li. *Better Alignment with Instruction Back-and-Forth Translation.* To appear in EMNLP Findings. https://arxiv.org/abs/2406.11794

Jeffrey Li, Jieyu Zhang, Ludwig Schmidt, Alexander Ratner. *Characerizing the Impacts of Semi-supervised Learning for Weak Supervision.* NeurIPS, 2023.
`https://openreview.net/forum?id=Z8TjsPFBSx`

Valerie Chen*, **Jeffrey Li***, Joon-sik Kim, Gregory Plumb, Ameet Talwalkar. *Interpretable Machine Learning: Moving from mythos to diagnostics.* ACM Queue, 2022.
`https://queue.acm.org/detail.cfm?id=3511299`

Jeffrey Li*, Vaishnavh Nagarajan*, Gregory Plumb, Ameet Talwalkar. *A Learning Theoretic Perspective on Local Explainability.* ICLR, 2021.
`https://openreview.net/forum?id=7aL-OtQrBWD`

Jeffrey Li, Mikhail Khodak, Sebastian Caldas, Ameet Talwalkar. *Differentially Private Meta-Learning.* ICLR, 2020.
`https://openreview.net/forum?id=rJgqMRVYvr`

EXPERIENCE

University of Washington Pittsburgh, PA
Graduate Research Assistant *October 2020 - Present*
Worked with Alexander Ratner and Ludwig Schmidt on projects related to data-centric ML, including weak supervision and data curation for LLMs

Apple Seattle, WA
ML Research Intern *April 2023 - June 2025*
Explored data curation and time-continual pre-training of LLMs

Snorkel AI Redwood City, CA
Research Intern *June 2023 - August 2023*
Explored weak supervision approaches for building instruction-tuning datasets

Carnegie Mellon University Pittsburgh, PA
Research Assistant *January 2019 - August 2020*
Worked with Ameet Talwalkar on projects studying privacy in the context of meta-learning as well as interpretable machine learning

Donut Technologies New York, NY
Software Engineering Intern *June 2017 - August 2017*
Worked at a startup focused on streamlining new employee onboarding, implemented features such as new event types and creator/editor tracking

Societe Generale New York, NY
Financial Engineering Intern *June 2016 - August 2016*
Automated weekly market monitors and investigated fair-value/momentum hedging to assist in suggesting equity derivative trading strategies

Dartmouth College Hanover, NH
Data Science Research Assistant *June 2015 - August 2015*
Worked with Zhigang Li to apply unsupervised machine learning techniques to explore a dataset about placental metal concentrations

SERVICE & TEACHING

Conference Reviewer

- ICLR 2024, 2025
- MLSys 2022
- NeurIPS 2021

Teaching Assistant

- UW CSEP546: Machine Learning (Winter 2023)
- UW CSE415: Introduction to AI (Winter 2022)

Mentorship

- Lingrong Cai (UW-Madison Undergraduate)
- Tianyi Zhang (University of Washington Undergraduate)

Pathmentors Mentor *Aug 2021 – Present*

- Mentor for two high school students pursuing independent studies and projects centered on applications of ML

AI4ALL Project Leader *July 2019*

- Planned a two week project and lecture series to teach low-income high schoolers about CNNs and their application to plant-disease detection

One-to-One Tutor *September 2014-May 2015*

- Matched as a weekly volunteer tutor for a family in an under-resourced school district in New York City

SKILLS

Competent: Python, PyTorch, AWS (EC2 and S3), Unix, Git, Ray
Familiar: Docker, SQL, Java, C, Ruby