Ashwinee Panda

website Email

Experience

24 - 25	Postdoc	**UMD College Park**	AI Training	with Tom Goldstein
24	Intern	**Capital One**	Pretraining	
20 - 24	PhD	**Princeton**	AI Safety	with Prateek Mittal
20	B.S./M.S.	**UC Berkeley**	AI Systems	with Joey Gonzalez

Publications

Safety	Xiangyu Qi, **Ashwinee Panda**, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, Peter Henderson Safety Alignment Should be Made More Than Just a Few Tokens Deep *ICLR 2025*
Auditing	**Ashwinee Panda**[*], Xinyu Tang[*], Milad N., Chris C., Prateek M. Privacy Auditing of LLMs *ICLR 2025*
DP-ZO	Xinyu Tang[*], **Ashwinee Panda**[*], Milad N., Saeed M., Prateek M. Private Fine-tuning of LLMs with Zeroth-order Optimization *TPDP 2024 Oral, TMLR 2025*
DP-Scaling	**Ashwinee Panda**[*], Xinyu Tang[*], Vikash S., Saeed M., Prateek M. A New Linear Scaling Rule for Private Adaptive HPO *ICML 2024 Poster*
Phishing	**Ashwinee Panda**, Chris C., Zhengming Z., Yaoqing Y., Prateek M. Teach LLMs to Phish: Stealing Private Information from LLMs *ICLR 2024 Poster*
DP-ICL	Tong Wu[*], **Ashwinee Panda**[*], Tianhao Wang[*], Prateek M. Privacy-Preserving In-Context Learning for LLMs *ICLR 2024 Poster*
DP-RandP	Xinyu Tang[*], **Ashwinee Panda**[*], Prateek M. DP Image Classification by Learning Priors from Random Processes *NeurIPS 2023 Spotlight*
AdvVLM	Xiangyu Qi[*], Kaixuan H.[*], **Ashwinee Panda**, Mengdi W., Prateek M. Introducing Vision into LLMs Expands Attack Surfaces *AAAI 2024 Oral*
Neurotoxin	Zhengming Zhang[*], **Ashwinee Panda**[*], Linyue S., Yaoqing Y., Prateek M., Joey G., Kannan R., Michael M. NeuroToxin: Durable Backdoors in Federated Learning *ICML 2022 Oral*
SparseFed	**Ashwinee Panda**, Saeed M., Arjun B., Supriyo C., Prateek M. SparseFed: Mitigating Model Poisoning Attacks in FL via Sparsification *AISTATS 2022 Poster*
FetchSGD	Daniel Rothchild[*], **Ashwinee Panda**[*], Enayat U., Nikita I., Ion S., Vladimir B., Joey G., Raman A. FetchSGD: Communication-Efficient Federated Learning with Sketching *ICML 2020 Poster*

Preprints

DenseMoE	**Ashwinee Panda**, Vatsal Baherwani, Zain Sarwar, Benjamin Thérien, Stephen Rawls, Sambit Sahu, Supriyo Chakraborty, Tom Goldstein Dense Backpropagation Improves Routing for Sparsely-Gated Mixture-of-Experts *NeurIPS 2024 ENSLP/OPT Workshops*
Scaling	Sean Michael McLeish, John Kirchenbauer, David Yu Miller, Siddharth Singh, Abhinav Bhatele, Micah Goldblum, **Ashwinee Panda**, Tom Goldstein Gemstones: A Model Suite for Multi-Faceted Scaling Laws
StructMoE	Zain Sarwar, **Ashwinee Panda**, Benjamin Thérien, Stephen Rawls, Sambit Sahu, Supriyo Chakraborty StructMoE: Augmenting MoEs with Hierarchically Routed Low Rank Experts *NeurIPS 2024 ENSLP Workshop*
MoE-CPT	Benjamin Thérien, Charles-Étienne Joseph, Zain Sarwar, **Ashwinee Panda**, Anirban Das, Shi-Xiong Zhang, Stephen Rawls, Sambit Sahu, Eugene Belilovsky, Irina Rish Continual Pre-training of MoEs: How robust is your router?
Refusal	Neel Jain, Aditya Shrivastava, Chenyang Zhu, Daben Liu, Alfy Samuel, Anoop Kumar, **Ashwinee Panda**, Micah Goldblum, Tom Goldstein Refusal Tokens: A Simple Way to Control Refusal Messages *NeurIPS 2024 SafeGenAI Workshop*
T2I Eval	Kevin David Hayes, Micah Goldblum, Vikash Sehwag, Gowthami Somepalli, **Ashwinee Panda**, Tom Goldstein FineGRAIN: Evaluating Failure Modes of Text-to-Image Models with Vision Language Model Judges
LoTA	**Ashwinee Panda**, Berivan I., Xiangyu Q., Sanmi K., Tsachy W., Prateek M. Lottery Ticket Adaptation: Mitigating Destructive Interference in LLMs *ICML 2024 ES-FoMO Oral, ICML 2024 WANT Best Paper*
DP-Diffusion	Vikash S.[*], **Ashwinee Panda**[*], Ashwini Pokle, Xinyu Tang, Saeed M., Mung Chiang, J Zico Kolter, Prateek M. Differentially Private Generation of High Fidelity Samples From Diffusion Models ICML 2023 GenAI Workshop

Awards

24	OpenAI Superalignment Fast Grant (PI)
24	Far AI Grant (PI)
20	Gordon Wu Fellowship
18	LAUNCH Grand Prize
18	Y Combinator Hackathon First Prize

Invited Talks

Feb '25	Expanding Bottlenecks in LLM Scaling *AllenAI (AI2)*
Sep '24	Lottery Ticket Adaptation *Google Federated Learning Seminar*
Sep '24	Privacy Auditing of LLMs *Google Privacy Seminar*
May '24	Challenges in Adapting LLMs to Private Data Google Privacy Seminar (click for talk recording)
Nov '23	New Privacy Attacks on Large Language Models *Sun Lab, Berkeley*
Nov '23	Challenges in Data-Driven Alignment of Large Language Models *SPYLab, ETH Zurich*
Oct '23	New Directions in Differentially Private Machine Learning *Meta CAS*
Sep '23	Challenges in Data-Driven Alignment of Large Language Models *University of Maryland, College Park*
Sep '23	Challenges in Augmenting Large Language Models with Private Data *SL^2 Lab, UIUC*
Sep '23	Improving the Privacy Utility Tradeoff in Differentially Private Machine Learning with Prior Information *SECRIT, University of Michigan*
Apr '23	Improving the Privacy Utility Tradeoff in Differentially Private Machine Learning with Public Data *Apple*
Mar '23	Google Privacy Seminar (click for talk recording) *Google*
Jun '22	Challenges and Directions in Privacy Preserving Machine Learning *Microsoft Research Cambridge*
May '22	Towards Trustworthy Machine Learning *Meta AI*
Jan '22	Federated Learning for Forecasting *Ohmconnect*
Nov '21	Building Federated Learning Systems at Scale *Liftoff AI*
Nov '21	Practical Defenses Against Model Poisoning Attacks Google (click for talk recording)

Service

Organizing

ICLR 2025 Sparsity in LLMs Workshop (Lead Organizer)

Teaching

2023	Teaching Assistant for COS/ECE 432 at Princeton
2019	Course Staff for CS 189 (Machine Learning) at UC Berkeley
2018	Teaching Assistant for CS 70 and CS 189 at UC Berkeley
2017	Course Staff for CS 70 at UC Berkeley

Peer Reviewing (* denotes Best Reviewer Award)

ICML25 (AC),24*,23,22,21,20,19; **NeurIPS**24,23*,21,**ICLR** 24*,23,19,**ACL**23, **TMLR**24,**AISTATS**22, **SATML**23

Advising

Sukriti Paul, Kevin Hayes, Pedro Sandoval, David Miller, Sean McLeish, Vatsal Baherwani, Neel Jain, Alex Stein, John Cava, Vivan Madan, Jie Li, Yuxin Wen, Ryan Synk, Monte Hoover, Khalid Saifullah, Juzheng Zhang, John Kirchenbauer, Hongmin Chu