

Ashwinee PANDA

[WEBSITE](#)

[EMAIL](#)

EXPERIENCE

25 -	Staff RS	TogetherAI	RL	
24 -26	Postdoc	UMD	PRETRAINING	with Tom Goldstein
24	AI Research	Capital One	PRETRAINING	
20 - 24	PhD (ECE)	Princeton	POSTTRAINING	with Prateek Mittal
16 - 20	B.S./M.S. (EECS)	UC Berkeley	AI SYSTEMS	with Joey Gonzalez

AWARDS

25	Outstanding Paper Award at ICLR 2025
25	OpenPhilanthropy Grants (PI, \$310,000) for <i>Encoded Reasoning</i>
25	OpenPhilanthropy Grants (PI, \$310,000) for <i>Dataset Optimization</i>
25	OpenPhilanthropy Grants (PI, \$218,000) for <i>Efficient Reasoning</i>
24	OpenAI Superalignment Fast Grant (PI, \$200,000) for <i>Shallow Alignment</i>
24	Far AI Grant (PI, \$150,000) for <i>Dataset Optimization</i>
<=20	Gordon Wu Fellowship, LAUNCH Grand Prize, YC Hackathon First Prize

PUBLICATIONS (BEST PAPER, ORAL, SPOTLIGHT)

Reasoning and Reinforcement Learning

o-KL	Ashwinee P. O train-infer mismatch for Open-weight MoE RL in Open-source code
XORL	Ashwinee P., TogetherAI Team XORL: Extensible Orchestration for Reinforcement Learning
Rollouts	Juzheng Z, ... , Ashwinee P., Tom G. Learning from Mixed Rollouts: Logit Fusion as a Bridge Between Imitation and Exploration
Encoded	Vatsal B., Tom G., Ashwinee P. Not All LLM Reasoning is Visible in the Chain-of-Thought
Legibility	Dipika K., Jack H., Ashwinee P. Legibility and Visible Completeness Form a Basis For Classifying CoT Behaviors
Efficient	Dipika K., Ashwinee P. Reasoning Models Reason Inefficiently <i>NeurIPS 25 Workshops</i>

Inference

MTP	John K., ..., Micah G., Ashwinee Panda, Tom G. Multi-Token Prediction via Self-Distillation
Prefetch	Vivan M., Prajwal S., Abhinav B., Tom G., Ashwinee Panda Speculating Experts Accelerates Inference for Mixture-of-Experts
Sinks	Pedro S., Xijun W., Ashwinee P., Micah G., Ronen B. Tom G. David J. Identifying and Evaluating Inactive Heads in Pretrained LLMs <i>ICLR 2026</i>

Pretraining

Benchmax	Hong-Min C., Neel J., ..., Tom G., Ashwinee P. Benchmaxxing: Selecting Data that Targets Many Benchmarks with Scalable Gradient-Based Methods
--------------------------	--

DenseMoE	Ashwinee P., Vatsal B., Zain S., ..., Tom G., Supriyo C. Dense Backpropagation Improves Training for Sparse MoEs <i>NeurIPS 25</i>
Gemstones	Sean M., John K., David M., ..., Micah G., Ashwinee P., Tom G. Gemstones: A Model Suite for Multi-Faceted Scaling Laws <i>NeurIPS 25</i>
FSA	Foreign Sparse Attention: Effective Distillation into Sparse Attention Vijaykaarti S., Tom G., Ashwinee P. <i>ICML 25 Workshops</i>
DS-Opt	Scalable Dataset Optimization Hong-Min C., Vivan M., Jiachen W., Tom G., Ashwinee P. <i>ICML 25 Workshops</i>
StructMoE	Zain S., Ashwinee P., Benjamin T., Stephen R., Sambit S., Supriyo C. Dynamic FFNs Improve Representation Learning in Transformer Pre-training
MoE-CPT	Benjamin T., Charles J., Zain S., Ashwinee P., ..., Irina R. Continual Pre-training of MoEs: How robust is your router? <i>TMLR 25</i>
Post-training	
Safety	Xiangyu Qi, Ashwinee P., Kaifeng L., ..., Ahmad B., Prateek M., Peter H. Safety Alignment Should be Made More Than Just a Few Tokens Deep <i>ICLR 25, Best Paper</i>
Guardians	Monte H., Vatsal B., Neel J., ..., Bayan B., Ashwinee P., Tom G. DynaGuard: Realtime Content Moderation With User-Defined Policies <i>ICLR 2026</i>
LoRI	Juzheng Zhang, Jiacheng You, Ashwinee P., Tom Goldstein LoRI: Reducing Cross-Task Interference in Multi-Task LoRA <i>COLM 25</i>
Refusal	Neel J., ..., Ashwinee P., Micah G., Tom G. Refusal Tokens: A Simple Way to Control Refusal Messages <i>COLM 25</i>
LoTA	Ashwinee P., Berivan I., Xiangyu Q., Sanmi K., Tsachy W., Prateek M. Lottery Ticket Adaptation: Mitigating Destructive Interference in LLMs <i>ICML-WANT 24 Best Paper</i>
Privacy	
Auditing	Ashwinee P.* , Xinyu Tang* , Milad N., Chris C., Prateek M. Privacy Auditing of LLMs <i>ICLR 25</i>
DP-ZO	Xinyu Tang* , Ashwinee P.* , Milad N., Saeed M., Prateek M. Private Fine-tuning of LLMs with Zeroth-order Optimization <i>TPDP 24 Oral</i> , TMLR 25
DP-Scaling	Ashwinee P.* , Xinyu Tang* , Vikash S., Saeed M., Prateek M. A New Linear Scaling Rule for Private Adaptive HPO <i>ICML 25</i>
Phishing	Ashwinee P., Chris C., Zhengming Z., Yaoqing Y., Prateek M. Teach LLMs to Phish: Stealing Private Information from LLMs <i>ICLR 24</i>
DP-ICL	Tong Wu* , Ashwinee P.* , Tianhao Wang* , Prateek M. Privacy-Preserving In-Context Learning for LLMs

DP-RandP	ICLR 24
	Xinyu Tang*, Ashwinee P.* , Prateek M.
	DP Image Classification by Learning Priors from Random Processes
	<i>NeurIPS 23 Spotlight</i>
Neurotoxin	Zhengming Zhang*, Ashwinee P.* , Linyue S., Yaoqing Y., ... Prateek M.
	NeuroToxin: Durable Backdoors in Federated Learning
	<i>ICML 22 Oral</i>
SparseFed	Ashwinee P. , Saeed M., Arjun B., Supriyo C., Prateek M.
	SparseFed: Mitigating Model Poisoning Attacks in FL via Sparsification
	<i>AISTATS 22</i>
FetchSGD	Daniel Rothchild*, Ashwinee P.* , Enayat U., Nikita I., Joey G., Raman A.
	FetchSGD: Communication-Efficient Federated Learning with Sketching
	<i>ICML 20</i>
Multimodal	
FineGRAIN	Kevin H., Micah G., Vikash S., Gowthami S., Ashwinee P. , Tom G.
	FineGRAIN: Evaluating Failure Modes of T2I Models with VLM Judges
	<i>NeurIPS 25 Spotlight</i>
Video	Yuxin Wen, Jim Wu, Ajay Jain, Tom Goldstein, Ashwinee P.
	Analysis of Attention in Video Diffusion Transformers
	<i>ICML 25 Workshops</i>
AdvVLM	Xiangyu Qi*, Kaixuan H.*, Ashwinee P. , Mengdi W., Prateek M.
	Introducing Vision into LLMs Expands Attack Surfaces
	<i>AAAI 24 Oral</i>
DP-Diffusion	Vikash S.*, Ashwinee P.* , Ashwini P., Xinyu T., Saeed M., Mung C., Zico K., Prateek M.
	DP Generation of High Fidelity Samples From Diffusion Models
	<i>ICML 23 Workshops</i>

INVITED TALKS

SEP '25	Dense Backpropagation <i>xAI</i>
JUL '25	Expanding Bottlenecks in LLM Scaling <i>Essential AI</i>
JUN '25	Scalable Safety <i>Scale AI</i>
JUN '25	Worst-Case Membership Inference of LLMs <i>Google</i>
MAY '25	Scalable Safety <i>International Symposium on Trustworthy Foundation Models at MBZUAI</i>
APR '25	Safety Oversight via Reasoning <i>OpenAI</i>
MAR '25	Expanding Bottlenecks in LLM Scaling <i>Cartesia</i>
FEB '25	Expanding Bottlenecks in LLM Scaling <i>AllenAI (AI2)</i>
SEP '24	Lottery Ticket Adaptation <i>Google Federated Learning Seminar</i>
SEP '24	Privacy Auditing of LLMs <i>Google Privacy Seminar</i>
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 <i>Sun Lab, Berkeley</i>
NOV '23	Challenges in Data-Driven Alignment of Large Language Models <i>SPYLab, ETH Zurich</i>
OCT '23	New Directions in Differentially Private Machine Learning <i>Meta CAS</i>
SEP '23	Challenges in Data-Driven Alignment of Large Language Models <i>University of Maryland, College Park</i>
SEP '23	Challenges in Augmenting Large Language Models with Private Data <i>SL² Lab, UIUC</i>
SEP '23	Improving the Privacy Utility Tradeoff in Differentially Private Machine Learning with Prior Information <i>SECRIT, University of Michigan</i>
APR '23	Improving the Privacy Utility Tradeoff in Differentially Private Machine Learning with Public Data <i>Apple</i>
MAR '23	Google Privacy Seminar (click for talk recording) <i>Google</i>
JUN '22	Challenges and Directions in Privacy Preserving Machine Learning <i>Microsoft Research Cambridge</i>
MAY '22	Towards Trustworthy Machine Learning <i>Meta AI</i>
JAN '22	Federated Learning for Forecasting <i>Ohmconnect</i>
NOV '21	Building Federated Learning Systems at Scale <i>Liftoff AI</i>
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)

I have served as a reviewer 20+ times, receiving recognition for my reviewing efforts at ICML, ICLR, and NeurIPS. I have served as an AC on 10+ occasions for ICML, ICLR, NeurIPS, COLM and ACL.

ICML26-25 (AC),24*,23-19; **NeurIPS**25*,24,23*,21,**ICLR** 26(AC),25*,24,23,19,**ACL**25 (AC),23, **TMLR**24,**AISTATS**22, **SATML**23

Advising

I have been fortunate to have the opportunity to advise a number of talented students in Tom Goldstein's group during my time as a postdoctoral fellow at UMD.

[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](#), Vijaykaarti Sundarapandian, Rifaa Quadri, Abhimanyu Hans