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

Publications (Best Paper, Oral, Spotlight)

Reasoning and Reinforcement Learning

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 *NeurIPS 25 Workshops*

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 *ICLR 2026*

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
NeurIPS 25

Gemstones
Sean M., John K., David M., ..., Micah G., **Ashwinee P.**, Tom G.
Gemstones: A Model Suite for Multi-Faceted Scaling Laws
NeurIPS 25

FSA
Foreign Sparse Attention: Effective Distillation into Sparse Attention
Vijaykaarti S., Tom G., **Ashwinee P.**
ICML 25 Workshops

DS-Opt
Scalable Dataset Optimization
Hong-Min C., Vivan M., Jiachen W., Tom G., **Ashwinee P.**
ICML 25 Workshops

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?
TMLR 25

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
ICLR 25, Best Paper

Guardians
Monte H., Vatsal B., Neel J., ..., Bayan B., **Ashwinee P.**, Tom G.
DynaGuard: Realtime Content Moderation With User-Defined Policies
ICLR 2026

LoRI
Juzheng Zhang, Jiacheng You, **Ashwinee P.**, Tom Goldstein
LoRI: Reducing Cross-Task Interference in Multi-Task LoRA
COLM 25

Refusal
Neel J., ..., **Ashwinee P.**, Micah G., Tom G.
Refusal Tokens: A Simple Way to Control Refusal Messages
COLM 25

LoTA
Ashwinee P., Berivan I., Xiangyu Q., Sanmi K., Tsachy W., Prateek M.
Lottery Ticket Adaptation: Mitigating Destructive Interference in LLMs
ICML-WANT 24 Best Paper

Privacy

Auditing
Ashwinee P.*, Xinyu Tang*, Milad N., Chris C., Prateek M.
Privacy Auditing of LLMs
ICLR 25

DP-ZO
Xinyu Tang*, **Ashwinee P.***, Milad N., Saeed M., Prateek M.
Private Fine-tuning of LLMs with Zeroth-order Optimization
TPDP 24 Oral, TMLR 25

DP-Scaling
Ashwinee P.*, Xinyu Tang*, Vikash S., Saeed M., Prateek M.
A New Linear Scaling Rule for Private Adaptive HPO
ICML 25

Phishing
Ashwinee P., Chris C., Zhengming Z., Yaoqing Y., Prateek M.
Teach LLMs to Phish: Stealing Private Information from LLMs
ICLR 24

DP-ICL
Tong Wu*, **Ashwinee P.***, Tianhao Wang*, Prateek M.
Privacy-Preserving In-Context Learning for LLMs

ICLR 24

DP-RandP — Xinyu Tang*, **Ashwinee P.***, Prateek M.
DP Image Classification by Learning Priors from Random Processes
NeurIPS 23 Spotlight

Neurotoxin — Zhengming Zhang*, **Ashwinee P.***, Linyue S., Yaoqing Y., ... Prateek M.
NeuroToxin: Durable Backdoors in Federated Learning
ICML 22 Oral

SparseFed — **Ashwinee P.**, Saeed M., Arjun B., Supriyo C., Prateek M.
SparseFed: Mitigating Model Poisoning Attacks in FL via Sparsification
AISTATS 22

FetchSGD — Daniel Rothchild*, **Ashwinee P.***, Enayat U., Nikita I...Joey G., Raman A.
FetchSGD: Communication-Efficient Federated Learning with Sketching
ICML 20

Multimodal

FineGRAIN — Kevin H., Micah G., Vikash S., Gowthami S., **Ashwinee P.**, Tom G.
FineGRAIN: Evaluating Failure Modes of T2I Models with VLM Judges
NeurIPS 25 Spotlight

Video — Yuxin Wen, Jim Wu, Ajay Jain, Tom Goldstein, **Ashwinee P.**
Analysis of Attention in Video Diffusion Transformers
ICML 25 Workshops

AdvVLM — Xiangyu Qi*, Kaixuan H.*, **Ashwinee P.**, Mengdi W., Prateek M.
Introducing Vision into LLMs Expands Attack Surfaces
AAAI 24 Oral

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
ICML 23 Workshops

Invited Talks

Sep '25	Dense Backpropagation *xAI*
Jul '25	Expanding Bottlenecks in LLM Scaling *Essential AI*
Jun '25	Scalable Safety *Scale AI*
Jun '25	Worst-Case Membership Inference of LLMs *Google*
May '25	Scalable Safety *International Symposium on Trustworthy Foundation Models at MBZUAI*
Apr '25	Safety Oversight via Reasoning *OpenAI*
Mar '25	Expanding Bottlenecks in LLM Scaling *Cartesia*
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)

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 Sundarapandiyan, Rifaa Quadri, Abhimanyu Hans