Yoonho Lee

yoonholee95@gmail.com | yoonholee.com | google scholar | Stanford, CA

Education

Stanford University — *United States*
Ph.D. Computer Science, Advisor: Chelsea Finn — *2021 - present*

Pohang University of Science and Technology — *South Korea*
M.S. Computer Science, Advisor: Seungjin Choi — *2018*
B.S. Mathematics — *2016*

Publications

Total 4,900+ citations (Google Scholar, Apr 2025).

[28] Jubayer Ibn Hamid*, Ifdita Hasan Orney*, Michael Y. Li, Omar Shaikh, **Yoonho Lee**, Dorsa Sadigh, Chelsea Finn, Noah Goodman. *"SPIRAL: Learning to Search and Aggregate".* Preprint (arXiv:2606.23595)

[27] **Yoonho Lee**, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, Chelsea Finn. *"Meta-Harness: End-to-End Optimization of Model Harnesses".* **COLM 2026**; AI Agents for Discovery in the Wild @ **ACM CAIS 2026** (**Oral**); Agents in the Wild @ **ICML 2026** (**Spotlight**); RL in Big Worlds @ **RLC 2026**; RL Beyond Rewards @ **RLC 2026** (**Spotlight**)

[26] **Yoonho Lee**, Joseph Boen, Chelsea Finn. *"Feedback Descent: Open-Ended Text Optimization via Pairwise Comparison".* MemAgents @ **ICLR 2026** (**Best Paper Runner-Up**); RSI @ **ICLR 2026**

[25] Yuxiao Qu*, Anikait Singh*, **Yoonho Lee***, Amrith Setlur, Ruslan Salakhutdinov, Chelsea Finn, Aviral Kumar. *"RLAD: Training LLMs to Discover Abstractions for Solving Reasoning Problems".* **ICLR 2026**; ES-FoMo @ **ICML 2025** (**Spotlight**); Ram2 @ **COLM 2025** (**Oral**)

[24] **Yoonho Lee**, Jonathan Williams, Henrik Marklund, Archit Sharma, Eric Mitchell, Anikait Singh, Chelsea Finn. *"Test-Time Alignment via Hypothesis Reweighting".* **TMLR 2025**

[23] Yuejiang Liu, Jubayer Ibn Hamid, Annie Xie, **Yoonho Lee**, Maximilian Du, Chelsea Finn. *"Bidirectional Decoding: Improving Action Chunking via Closed-Loop Resampling".* **ICLR 2025**

[22] **Yoonho Lee**, Michelle Lam, Helena Vasconcelos, Michael S. Bernstein, Chelsea Finn. *"Clarify: Improving Model Robustness With Natural Language Corrections".* **UIST 2024**; XAIA @ **NeurIPS 2023**; ICBINB @ **NeurIPS 2023**

[21] Johnathan Xie*, Annie S. Chen*, **Yoonho Lee**, Eric Mitchell, Chelsea Finn. *"Calibrating Language Models With Adaptive Temperature Scaling".* **EMNLP 2024**

[20] Caroline Choi*, **Yoonho Lee***, Annie S. Chen, Allan Zhou, Aditi Raghunathan, Chelsea Finn. *"AutoFT: Learning an Objective for Robust Fine-Tuning".* Distribution Shifts @ **NeurIPS 2023**

[19] Annie S. Chen, **Yoonho Lee**, Amrith Setlur, Sergey Levine, Chelsea Finn. *"Confidence-Based Model Selection: When to Take Shortcuts for Subpopulation Shifts".* Distribution Shifts @ **NeurIPS 2023**

[18] Caroline Choi*, Fahim Tajwar*, **Yoonho Lee***, Huaxiu Yao, Ananya Kumar, Chelsea Finn. *"Conservative Prediction via Data-Driven Confidence Minimization".* **TMLR 2024**; TrustML @ **ICLR 2023**; ME-FoMo @ **ICLR 2023**

[17] Annie S. Chen*, **Yoonho Lee***, Amrith Setlur, Sergey Levine, Chelsea Finn. *"Project and Probe: Sample-Efficient Domain Adaptation by Interpolating Orthogonal Features".* **ICLR 2024** (**Spotlight**, top 5% of submissions); TrustML @ **ICLR 2023** (**Oral**); ME-FoMo @ **ICLR 2023**

[16] Johnathan Wenjia Xie, **Yoonho Lee**, Annie S. Chen, Chelsea Finn. *"Self-Guided Masked Autoencoders for Domain-Agnostic Self-Supervised Learning".* **ICLR 2024**

[15] Eric Mitchell, **Yoonho Lee**, Alexander Khazatsky, Christopher D Manning, Chelsea Finn. *"DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature".* **ICML 2023** (**Oral**, top 2% of submissions)

[14] **Yoonho Lee***, Annie S. Chen*, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, Chelsea Finn. *"Surgical Fine-Tuning Improves Adaptation to Distribution Shifts".* **ICLR 2023**

[13] **Yoonho Lee**, Huaxiu Yao, Chelsea Finn. *"Diversify and Disambiguate: Out-of-Distribution Robustness via Disagreement".* **ICLR 2023**

[12] **Yoonho Lee**, Chelsea Finn, Stefano Ermon. *"Relaxing the Kolmogorov Structure Function for Realistic Computational Constraints".* Information-Theoretic Principles in Cognitive Systems @ **NeurIPS 2022**

[11] Balhae Kim, Jungwon Choi, Seanie Lee, **Yoonho Lee**, Jung-Woo Ha, Juho Lee. *"On Divergence Measures for Bayesian Pseudocoresets".* **NeurIPS 2022**

[10] Huaxiu Yao*, Caroline Choi*, Bochuan Cao, **Yoonho Lee**, Pang Wei Koh, Chelsea Finn. *"Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time".* **NeurIPS 2022** (Datasets & Benchmarks track)

[9] **Yoonho Lee**, Wonjae Kim, Wonpyo Park, Seungjin Choi. *" Discrete Infomax Codes for Supervised Representation Learning".* **Entropy 2022** (special issue: Theory and Applications of Information Processing Algorithms)

[8] Giung Nam*, Jongmin Yoon*, **Yoonho Lee**, Juho Lee. *"Diversity Matters When Learning From Ensembles".* **NeurIPS 2021**

[7] Minkyo Seo*, **Yoonho Lee***, Suha Kwak. *"On the Distribution of Penultimate Activations of Classification Networks".* **UAI 2021**

[6] **Yoonho Lee**, Juho Lee, Sung Ju Hwang, Eunho Yang, Seungjin Choi. *"Neural Complexity Measures".* **NeurIPS 2020**

[5] Juho Lee*, **Yoonho Lee***, Jungtaek Kim, Eunho Yang, Sung Ju Hwang, Yee Whye Teh. *"Bootstrapping Neural Processes".* **NeurIPS 2020**

[4] Wonjae Kim, **Yoonho Lee**. *"Learning Dynamics of Attention: Human Prior for Interpretable Machine Reasoning".* **NeurIPS 2019**

[3] Juho Lee, **Yoonho Lee**, Yee Whye Teh. *"Deep Amortized Clustering".* Sets and Partitions @ **NeurIPS 2019** (**Oral**)

[2] Juho Lee, **Yoonho Lee**, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, Yee Whye Teh. *"Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks".* **ICML 2019**

[1] **Yoonho Lee**, Seungjin Choi. *"Gradient-based meta-learning with learned layerwise metric and subspace".* **ICML 2018**

Fellowships and Grants

HAI Google Cloud Credits Award, Stanford HAI, $15,000	2025–2026
OpenAI Superalignment Fellowship, OpenAI, $150,000 (50/2700 proposals)	2024
Microsoft Accelerate Foundation Models Research Grant, Microsoft Research, $20,000	2023–2024
HAI Google Cloud Credits Award, Stanford HAI, $15,000	2022–2023
KFAS Doctoral Fellowship, Korea Foundation for Advanced Studies	2021–present
Presidential Science Scholarship, Korea Student Aid Foundation	2012–2016

Mentoring

Zhengxu (Jason) Yan	2025-current
Teresa Zhang	2025-current
Roshen Nair	2025-current
Victor Kolev	2025
Ryan Park	2025
Jonathan Williams (Next: PhD at Princeton)	2023-2024
Jubayer Ibn Hamid (Next: PhD at Stanford)	2024-2025
Johnathan Wenjia Xie (Next: Tesla)	2024
Fahim Tajwar (Next: PhD at CMU)	2023
Caroline Choi (Next: PhD at Stanford)	2022

Professional Service

Workshop organizer, NeurIPS Workshop on Distribution Shifts (2022, 2023)

Reviewer: NeurIPS (18-25, workshops 24-25), ICML (19-25, position papers 26, workshops 24), ICLR (21-26), AAAI (24-25), AISTATS (19-22), IJCAI (19-21), ACML (19-20), ME-FoMo@ICLR (23), TrustML@ICLR (23).

Selected Talks

Meta-Harness: End-to-End Optimization of Model Harnesses

Max Planck Institute	Jul. 2026
Stanford University (Diyi Yang's group)	Jun. 2026
ACM CAIS 2026 AI Agents for Discovery in the Wild Workshop (Oral)	May 2026
Stanford University (Ludwig Schmidt's group)	Apr. 2026
Tsinghua University	Apr. 2026
MILA / ServiceNow	Apr. 2026
POSTECH	Apr. 2026
Iterative Improvement for Large-Scale Generative Models, SPIGM Workshop @ NeurIPS 2025 (Invited Talk)	Dec. 2025
DetectGPT + Clarify, Moveworks	Dec. 2024
Clarify: Improving Model Robustness With Natural Language Corrections, UIST 2024 (Oral)	Oct. 2024
Distribution-Shift Robustness via Diversity and Non-Learning, Centre for Frontier AI Research	Aug. 2023
Surgical Fine-Tuning Improves Adaptation to Distribution Shifts, MosaicML	Aug. 2023

Teaching Experience

Teaching Assistant, CS330 Deep Multi-Task and Meta Learning, Stanford University	Sep. 2023 - Dec. 2023
Teaching Assistant, CS330 Deep Multi-Task and Meta Learning, Stanford University	Sep. 2022 - Dec. 2022
Teaching Assistant, Deep Learning, POSCO Group	Mar. 2017 - Jun. 2018
Teaching Assistant, Machine Learning for Business, Samsung Electronics	Sep. 2017 - Dec. 2017
Teaching Assistant, AI Job Training, POSTECH Institute of AI	Mar. 2017 - Jun. 2017
Teaching Assistant, CSED101 Programming and Problem Solving, POSTECH	Mar. 2017 - Jun. 2017