Yoonho Lee

yoonholee95@gmail.com | yoonholee.com | google scholar | Oct 2026

Education

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

POSTECH *South Korea*
M.S. Computer Science, Advisor: Seungjin Choi *2018*
B.S. Mathematics *2016*

Experience

Google Research
Student Researcher, Hosts: Ching-An Cheng, Allen Nie *2026*

AITRICS
Research Scientist (alternative military service) *2020–2021*

Kakao Corporation
Research Scientist (alternative military service) *2018–2020*

Publications

* equal contribution

[34] Anikait Singh*, Teresa Zhang*, **Yoonho Lee***, Roshen Nair, Aviral Kumar, Chelsea Finn. *"Librarian Agents: Training LLMs to Curate Long-Context Workspaces that Anticipate Future Queries".* **Preprint** (arXiv:2610.22273)

[33] Qizheng Zhang, Changxiu Ji, Isaac Sun, Yuetai Li, Shubhangi Upasani, Sherry Ruan, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, **Yoonho Lee**, Yuzhen Mao, Genghan Zhang, Rulin Shao, Qiuyang Mang, Andy Dimnaku, Changran Hu, Radha Poovendran, Kunle Olukotun. *"Stateless Language Agents: Scaling Long-Horizon Automated Research".* **Preprint** (arXiv:2610.07625)

[32] Sohyeon Kim*, **Yoonho Lee***, Bo Liu, Dayoon Ko, Rulin Shao, Seungone Kim, Graham Neubig, Pang Wei Koh, Aakanksha Chowdhery, Akari Asai, Omar Khattab, Yejin Choi, Gunhee Kim, Chelsea Finn. *"ScholarCatalyst: A Benchmark for Retrieving Papers That Inspire New Research".* **Preprint** (arXiv:2610.02202)

[31] Peng Xia, Rujun Han, Zifeng Wang, Yanfei Chen, Yufan Zhuang, **Yoonho Lee**, Chengsong Huang, Han Yu, Zhongying CuiZhu, Yifei Ming, Huaxiu Yao, Burak Gokturk, Tomas Pfister, Chen-Yu Lee. *"RRSI: Regularized Recursive Self-Improvement of Agent Harnesses".* **Preprint** (arXiv:2609.24972)

[30] Haechan Kim, **Yoonho Lee**, Gisang Lee, Chelsea Finn, Kangwook Lee. *"WHALE: A Simple Recipe for Joint Harness-Weight Optimization".* **Preprint** (arXiv:2609.00196)

[29] 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)

[28] **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**)

[27] **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**

[26] 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**)

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

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

[23] **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**

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

[21] Yueqi Wang*, **Yoonho Lee***, Pallab Basu, Juho Lee, Yee Whye Teh, Liam Paninski, Ari Pakman. *"Amortized Probabilistic Detection of Communities in Graphs".* SPIGM @ **ICML 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 3% 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**; DistShift @ **NeurIPS 2022**; ICBINB @ **NeurIPS 2022**

[13] **Yoonho Lee**, Huaxiu Yao, Chelsea Finn. *"Diversify and Disambiguate: Out-of-Distribution Robustness via Disagreement".* **ICLR 2023**; PODS @ **ICML 2022**; SCIS @ **ICML 2022**

[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); Shift Happens @ **ICML 2022**

[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

Laude Slingshots Four, Laude Institute, for ScholarCatalyst 2026

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, full ride and living allowance 2012–2016

Selected Talks

Rising Stars in AI, University of Michigan AI Symposium (Invited Talk) Oct. 2026

Text Optimization as a Path to Self-Improving Systems

- Stanford CS300 Seminar Oct. 2026
- COLM 2026 Self-Improving Agents Social (Invited Talk) Oct. 2026

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

Mentoring

Sohyeon Kim 2026–present

Teresa Zhang 2025–present

Roshen Nair 2025–present

Elijah Song 2026

Zhengxu (Jason) Yan 2025–2026

Victor Kolev 2025

Ryan Park (Next: Chai Discovery) 2025

Jubayer Ibn Hamid (Next: PhD at Stanford) 2024–2025

Jonathan Williams (Next: PhD at Princeton) 2023–2024

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)

Area chair. Meta-Agents Workshop @ NeurIPS (2026).

Reviewer. NeurIPS (2018–2026, workshops 2024–2025), ICML (2019–2026, position papers 2026, workshops 2024), ICLR (2021–2027), AAAI (2024–2025), AISTATS (2019–2022), IJCAI (2019–2021), ACML (2019–2020), ME-FoMo@ICLR (2023), TrustML@ICLR (2023).

Teaching Experience

Guest Lecturer, CS 590/790 Transfer Learning, UNC Chapel Hill Apr. 2024

Teaching Assistant, CS330 Deep Multi-Task and Meta Learning, Stanford University Fall 2022, Fall 2023

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