

Yongjin Yang

Ph.D. Student, University of Toronto & Vector Institute

dyyjkd@gmail.com | [Website](#) | [Google Scholar](#) | [GitHub](#) | [LinkedIn](#)

RESEARCH INTERESTS

Post-training and self-improvement for language models, with a focus on exploration, adaptive curricula, knowledge distillation, reinforcement learning, memory-augmented agents, and continual learning.

EDUCATION

University of Toronto (UofT)

Ph.D. in Computer Science; Adviser: Zhijing Jin

Toronto, Canada

Sep 2025–Present

Korea Advanced Institute of Science and Technology (KAIST)

M.S. in Graduate School of Artificial Intelligence; Advisers: Se-Young Yun, Kimin Lee

Seoul, South Korea

Mar 2023–Feb 2025

GPA: 4.18/4.3; ranked 2nd in the department

Seoul National University (SNU)

B.S. in Electrical and Computer Engineering

Seoul, South Korea

Mar 2017–Feb 2023

GPA: 4.12/4.3; summa cum laude

SELECTED PUBLICATIONS IN LLM POST-TRAINING AND SELF-IMPROVEMENT

*Equal contribution. †Corresponding authors.

- [1] **Yongjin Yang**, Jiarui Liu, Yinghui He, Lezhen Zhang, Bernhard Schölkopf, Zhijing Jin[†]. Transferability for General Reasoning: An Automated Curriculum for Multi-Domain RLVR. *Preprint; ICML 2026 RLxF Workshop (Oral; selected among four best papers from 222 submissions)*. [\[paper\]](#) [\[code\]](#)
- [2] **Yongjin Yang**^{*}, Sinjae Kang^{*}, Juyong Lee, Dongjun Lee, Se-Young Yun[†], Kimin Lee[†]. Automated Skill Discovery for Language Agents through Exploration and Iterative Feedback. *Preprint*. [\[paper\]](#)
- [3] **Yongjin Yang**, Yinghui He, Jiarui Liu, Zhijing Jin. Making Complex Reasoning Student-Friendly: A Hybrid Distillation Framework for Small Language Models. *The 1st Workshop on Scaling Post-Training for LLMs at ICLR 2026*. [\[paper\]](#)
- [4] Yujin Kim^{*}, Namgyu Ho^{*}, Sangmin Hwang^{*}, Joonkee Kim, **Yongjin Yang**, Sangmin Bae, Seungone Kim, Jaehun Jung, Se-Young Yun[†], Hwanjun Song[†]. LLM-as-a-Tutor: Policy-Aware Prompt Adaptation for Non-Verifiable RL. *Preprint*. [\[paper\]](#)
- [5] Yinghui He, Simran Kaur, Adithya Bhaskar, **Yongjin Yang**, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, Sanjeev Arora. Self-Distillation Zero: Self-Revision Turns Binary Rewards into Dense Supervision. *Preprint; ICML 2026 RLxF Workshop (Oral; selected among four best papers)*. [\[paper\]](#) [\[code\]](#)
- [6] Woogyoul Jin, Taywon Min, **Yongjin Yang**, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, Kimin Lee. Entropy-Aware On-Policy Distillation of Language Models. *The Forty-Third International Conference on Machine Learning (ICML 2026)*. [\[paper\]](#) [\[code\]](#)
- [7] Jiarui Liu, Lechen Zhang, **Yongjin Yang**, Yinghui He, Yingheng Wang, Weihao Xuan, Zhijing Jin, Mona T. Diab. MixSD: Mixed Contextual Self-Distillation for Knowledge Injection. *Preprint*. [\[paper\]](#) [\[code\]](#)
- [8] **Yongjin Yang**^{*}, Euiin Yi^{*}, Jongwoo Ko, Kimin Lee[†], Zhijing Jin[†], Se-Young Yun[†]. Revisiting Multi-Agent Debate as Test-Time Scaling: A Systematic Study of Conditional Effectiveness. *Preprint; ICML 2025 Multi-Agent Systems Workshop*. [\[paper\]](#) [\[code\]](#)

ADDITIONAL PUBLICATIONS

- [9] Punya Syon Pandey, **Yongjin Yang**, Jiarui Liu, Zhijing Jin. CORE: Measuring Multi-Agent LLM Interaction Quality under Game-Theoretic Pressures. *The 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026 Main)*. [\[paper\]](#)
- [10] Yejin Son, **Yongjin Yang**^{*}, Ryan Faulkner, Matt Ratto, Seungwon Lim, Youngjae Yu, Zhijing Jin. The Wedge Questions: Latent Cultural Boundaries in LLMs via Persona Projection Divergence. *Preprint; ICML 2026 Pluralistic Alignment Workshop*. [\[paper\]](#)
- [11] Segyu Lee^{*}, Boryeong Cho^{*}, Hojung Jung^{*}, Seokhyun An, Juhyeong Kim, Jaehyun Kwak, **Yongjin Yang**, Sangwon Jang, Youngrok Park, Wonjun Chang, Se-Young Yun. UniSAFE: A Comprehensive Benchmark for Safety Evaluation of Unified Multimodal Models. *Preprint*. [\[paper\]](#) [\[code\]](#)
- [12] Younwoo Choi^{*}, Changling Li^{*}, **Yongjin Yang**, Zhijing Jin. Agent-to-Agent Theory of Mind: Testing Interlocutor Awareness among Large Language Models. *The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025 Main)*. [\[paper\]](#) [\[code\]](#)
- [13] David Guzman Piedrahita, **Yongjin Yang**, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, Zhijing Jin. Corrupted by Reasoning: Reasoning Language Models Become Free-Riders in Public Goods Games. *Conference on Language Modeling (COLM 2025); ACL 2025 REALM Workshop (Oral)*. [\[paper\]](#) [\[code\]](#)
- [14] Tergel Munkhbat^{*}, Namgyu Ho^{*}, Seo Hyun Kim^{*}, **Yongjin Yang**, Yujin Kim, Se-Young Yun. Self-Training Elicits Concise Reasoning in Large Language Models. *Findings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025 Findings)*. [\[paper\]](#) [\[code\]](#)
- [15] Haneul Yoo, **Yongjin Yang**, Hwaran Lee. Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding. *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025 Main); NeurIPS 2024 Workshop on Red Teaming GenAI*. [\[paper\]](#) [\[code\]](#)
- [16] **Yongjin Yang**^{*}, Sihyeon Kim^{*}, Hojung Jung, Sangmin Bae, Sangmook Kim, Se-Young Yun[†], Kimin Lee[†]. Automated Filtering of Human Feedback Data for Aligning Text-to-Image Diffusion Models. *The Thirteenth International Conference on*

Learning Representations (ICLR 2025). [paper] [code]

- [17] **Yongjin Yang**, Haneul Yoo, Hwaran Lee. MAQA: Evaluating Uncertainty Quantification in LLMs Regarding Data Uncertainty. *Findings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025 Findings)*. [paper] [code]
- [18] **Yongjin Yang***, Jongwoo Ko*, Se-Young Yun. Towards Difficulty-Agnostic Efficient Transfer Learning for Vision-Language Models. *The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024 Main)*. [paper] [code]
- [19] **Yongjin Yang***, Sihyeon Kim*, SangMook Kim*, Gyubok Lee, Se-Young Yun, Edward Choi. Towards Unbiased Evaluation of Detecting Unanswerable Questions in EHRSQL. *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*. [paper]
- [20] **Yongjin Yang**, Taehyeon Kim, Se-Young Yun. Leveraging Normalization Layer in Adapters With Progressive Learning and Adaptive Distillation for Cross-Domain Few-Shot Learning. *The 38th Annual AAAI Conference on Artificial Intelligence (AAAI 2024)*. [paper]
- [21] **Yongjin Yang***, Joonkee Kim*, Yujin Kim*, Namgyu Ho, James Thorne[†], Se-Young Yun[†]. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023 Findings)*. [paper] [code]
- [22] Minchang Kim*, **Yongjin Yang***, Jung Hyun Ryu, Taesup Kim. Meta-Learning with Adaptive Weighted Loss for Imbalanced Cold-Start Recommendation. *The 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023, Oral)*. [paper] [code]

INDUSTRY RESEARCH EXPERIENCE

ServiceNow AI Research <i>Research Intern, UI Assist Team</i> Developing memory-augmented and continual-learning methods for language-model agents, investigating how interaction experience can be stored and reused across model parameters, external memory, and agent workflows.	Montreal, Canada Feb 2026–Sep 2026
Naver AI Lab <i>Research Intern, Language Research Team</i> Conducted research on uncertainty quantification and code-switching red-teaming for large language models, resulting in publications at NAACL 2025 Findings and ACL 2025 Main.	Seongnam, South Korea Mar 2024–Jun 2024
Samsung Electronics <i>Software Engineering Intern, Mobile Experience Team</i> Designed and implemented an automated logging framework for measuring execution times across Android System APIs.	Seoul, South Korea Jul 2021–Aug 2021

EARLIER RESEARCH PROJECTS

Lightweight Edge-Device Technology via Federated Learning , KAIST OSI Lab Developed federated-learning pipelines for object detection and edge-device image classification in smart-farm settings; supported by the Electronics and Telecommunications Research Institute (ETRI).	Mar 2023–Dec 2024
Skin Color Estimation , SNU 3D Vision Lab Developed a skin-color estimation framework using 3D geometry and deep learning; supported by COSMAX.	Jan 2022–Jun 2022

AWARDS AND GRANTS

Connaught International Scholarship , University of Toronto Entrance scholarship for five international students.	Apr 2025
Best Paper Award , Korea Software Congress For <i>FedCON: Model Contrastive Federated Learning on Re-Identification</i> .	Dec 2023
Korea National Science and Technology Scholarship Full-ride academic excellence scholarship.	Mar 2017–Feb 2023

ACADEMIC SERVICE

Reviewer: ICLR (2025, 2026); NeurIPS (2025, 2026); ICML (2026); COLM (2025, 2026); AAAI (2025); Transactions on Machine Learning Research (TMLR); ACL Rolling Review (Feb, Jun, Oct, Dec 2024; Feb, May, Jul, Oct 2025; Jan 2026); DPFM at ICLR (2024, 2025); Multi-Agent Systems at ICML (2025).

TEACHING

Teaching Assistant , Introduction to Programming Language, University of Toronto	Sep 2025–Dec 2025
Departmental Tutor , Introduction to Algorithm, Seoul National University	Sep 2021–Feb 2022
Group Tutor , Principles of Physics I, Seoul National University	Mar 2018–Aug 2018

PATENT

YoungMin Kim, Hojun Jang, **Yongjin Yang**, Junho Lee. “Method, Device and Computer-Readable Medium for Color Estimation in Image.” Korean Patent Application No. 10-2022-0055230.