

Minbeom Kim

✉ research@minbeom.kim | 🏠 https://minbeomkim.github.io/ | 📄 Google Scholar | 🌐 mbkimm

Education

Seoul National University

Ph.D. in Artificial Intelligence

Thesis: Controllable Post-Training of Language Models via Distributional Matching

Best Ph.D Dissertation Award

Advisor: [Kyomin Jung](#)

Seoul, South Korea

Sep. 2021 - Aug. 2026

KAIST

B.E. in Industrial and System Engineering

Daejeon, South Korea

Mar. 2014 - Feb. 2018

Professional Experiences

SK hynix Research

Senior Researcher

Working on Recursive Self-Improvement

Bundang, South Korea

Jul. 2026 -

Google

Student Researcher (worked on reliable multi-agent AI systems)

Host: [Long T. Le](#)

California, United States

Oct. 2025 - Jun. 2026

LawZero

Research Intern (worked on Scientist AI: probabilistic safety guarantees for superintelligent AI)

Host: [Yoshua Bengio](#)

Montreal, Canada

Jun. 2025 - Sep. 2025

NAVER Labs Europe

Research Intern (worked on theoretical guarantees for constraint satisfaction in generative models)

Host: [Marc Dymetman](#)

Grenoble, France

Apr. 2024 - Sep. 2024

Agency for Defense Development

Research Officer / 1st lieutenant (duration of military service)

I worked on R&D Projects for National Defense Technology. This job also fulfilled my 3-year mandatory military service in South Korea

Daejeon, South.Korea

June. 2018 - May. 2021

Publications

- [1] **Minbeom Kim**, Lesly Miculicich, Bhavana Dalvi Mishra, Mihir Parmar, Phillip Wallis, Bharath Chandrasekhar, Kyomin Jung, Tomas Pfister, Long T. Le, "LiSA: Lifelong Safety Adaptation via Conservative Policy Induction", *arXiv preprint*
- [2] Jeonghye Kim, Xufang Luo, **Minbeom Kim**, Sangmook Lee, Dongsheng Li, Yuqing Yang, "Rethinking Less Is More: Understanding Reasoning as Strategic Information Allocation under Uncertainty", *arXiv preprint*
- [3] Sangmook Lee, **Minbeom Kim**, Jeonghye Kim, Dohyung Kim, Sojeong Lee, Kyomin Jung, "Are We Measuring Strategy or Phrasing? The Gap Between Surface- and Approach-Level Diversity in LLM Math Reasoning", In: *EMNLP 2026*
In: *ICML 2026 AI4Math Workshop (Spotlight, Top 2.9%)*
- [4] Jeonghye Kim, Xufang Luo, **Minbeom Kim**, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, Yuqing Yang, "Why Does Self-Distillation (Sometimes) Degrade the Reasoning Capability of LLMs?", In: *COLM 2026*
- [5] Kyungmin Min, **Minbeom Kim**, Kang-il Lee, Seunghyun Yoon, Kyomin Jung, "ReflectCap: Detailed Image Captioning with Reflective Multi Agents", In: *ECCV 2026*

- [6] **Minbeom Kim**, Mihir Parmar, Phillip Wallis, Lesly Miculicich, Kyomin Jung, Krishnamurthy Dj Dvijotham, Long T. Le, Tomas Pfister. "CausalArmor: Efficient Indirect Prompt Injection Guardrails via Causal Attribution", In: **ICML 2026**
- [7] Dohyung Kim, **Minbeom Kim**, Jeonghye Kim, Sangmook Lee, Sojeong Lee, Kyomin Jung, "Beyond Normalization: Rethinking the Partition Function as a Difficulty Scheduler for RLVR", In: **ICML 2026**
- [8] Kang-il Lee, Jahyun Koo, Seunghyun Yoon, **Minbeom Kim**, Hyukhun Koh, Dongryeol Lee, Kyomin Jung. "Transductive Program Synthesis", In: **NeurIPS 2025**
- [9] **Minbeom Kim**, Kang-il Lee, Seongho Joo, Hwaran Lee, Thibaut Thonet, Kyomin Jung. "Drift: Decoding-time Personalized Alignments with *Implicit* User Preferences", In: **EMNLP 2025**
- [10] Jeonghye Kim*, Sojeong Rhee*, **Minbeom Kim**, Sangmook Lee, Dohyung Lee, Youngchul Sung, Kyomin Jung. "ReflAct: World-Grounded Decision Making in LLM Agents via Goal-State Reflection", In: **EMNLP 2025**
- [11] Kyeongman Park, **Minbeom Kim**, and Kyomin Jung. "A Character-Centric Creative Story Generation via Imagination", In: **ACL 2025 Findings**
- [12] **Minbeom Kim**, Thibaut Thonet, Jos Rozen, Hwaran Lee, Kyomin Jung, and Marc Dymetman. "Guaranteed Generation from Large Language Models", In: **ICLR 2025**
- [13] **Minbeom Kim**, Hwanhee Lee, Joonsuk Park, Hwaran Lee and Kyomin Jung. "AdvisorQA: A Benchmark for Advice-seeking Question Answering with Collective Intelligence", In: **NAACL 2025**
- [14] Kang-il Lee, **Minbeom Kim**, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung, "VLind-Bench: Measuring Language Priors in Large Vision-Language Models", In: **NAACL 2025 Findings**
- [15] Kyungmin Min, **Minbeom Kim**, Kang-il Lee, Dongryeol Lee, and Kyomin Jung, "Mitigating Hallucinations in LLMs via Summary-Guided Decoding", In: **NAACL 2025 Findings**
- [16] **Minbeom Kim**, Jahyun Koo, Hwanhee Lee, Joonsuk Park, Hwaran Lee and Kyomin Jung. "LifeTox: Unveiling Implicit Toxicity in Advice-seeking Question Answering", In: **NAACL 2024 (Oral, Top 5.9%)**
- [17] **Minbeom Kim**, Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee and Kyomin Jung. "Critic-Guided Decoding for Controlled Text Generation", In: **ACL 2023 (Findings Spotlights)**
- [18] **Minbeom Kim**, Kyeongha Rho, Yong-duk Kim, Kyomin Jung. "Action-driven Contrastive Representation for Reinforcement Learning", In: **PLOS ONE 2022**
- [19] Changyeop Shin, **Minbeom Kim**, Sungho Kim and Youngjung Kim. "Stacked Lossless Deconvolutional Network for Remote Sensing Image Restoration", In: **Journal of Remote Sensing 2020**
- [20] Jianrui Cai, **Minbeom Kim**, Youngjung Kim, et al. "NTIRE 2019 Challenge on Real Image Super-Resolution: Methods and Results", In: **NTIRE @ CVPR 2019**

Awards and Honors

2026	Awards: Best Ph.D Dissertation Award (Seoul National University)	<i>Korea</i>
2024	Awards: Best Summer Intern in Naver Labs Europe	<i>France</i>
2021-2023	Scholarship: Artificial Intelligence Graduate School Program (Seoul National University)	<i>Korea</i>
2022	Awards: Creative AI Research Competition, 2rd (Seoul National University)	<i>Korea</i>

Academic Activities

Academic Services Area Chair at ACL ARR 2026 ~
ACL ARR, ICLR, ICML, NeurIPS 2026
ACL ARR, ICLR, ICML, NeurIPS 2025
ACL ARR, NeurIPS 2024
Several Times Nominated as a Great Reviewer at ACL ARR
Safety Guaranteed Generation from Large Language Models, 2025, **Mila Quebec, Canada**
AI Retreat, 2023, **Seoul National University, South Korea**

Teaching Assistant Programming Methodology (Spring 2022, 2025), **Seoul National University**
IoT, Artificial intelligence, and Big data (IAB) (Fall 2022, 2023), **Seoul National University**