(+1) 512-992-4955
seohyeon.cha@utexas.edu
Website | LinkedIn

Seohyeon Cha Ph.D. student

Research Summary

I study how to make AI models accurate and adaptable when memory, computation, and communication are limited. My work spans **low-bit LLM inference**, **memory-efficient continual learning**, and **distributed learning and inference**, with a focus on balancing model quality and deployment cost across devices and systems.

Education

The University of Texas at Austin Austin, TX
Ph.D. in Electrical and Computer Engineering 08/24–**09/28** (expected)

- Advisor: Prof. Haris Vikalo
- **Selected Coursework:** Deep Reinforcement Learning, Convex Optimization, Integer Programming, Statistical Machine Learning, Diffusion Models for Generative AI. (GPA: 4.0/4.0)

Korea Advanced Institute of Science and Technology (KAIST) Daejeon, Korea
M.S. in Electrical Engineering 2022–2024

- Advisor: Prof. Joonhyuk Kang
- **Selected Coursework:** Bayesian Nonparametric Methods for Machine Learning, Information Theory, Pattern Recognition. (GPA: 4.17/4.3)

B.S. in Electrical Engineering (GPA: 4.03/4.3, Summa Cum Laude) 2017–2022

Research Experience

Graduate Research Assistant, UT Austin 2024–Present

- **LLM quantization:** Developed CoreQ to limit quantization error propagation and balance PTQ accuracy against computation through mismatch correction and successive rounding, without retraining (**NeurIPS 2026 Spotlight**).
- **Efficient adaptation:** Balanced new-task adaptation and knowledge retention under memory limits; addressed misalignment in federated LoRA (**TMLR, ICLR, ICML 2026**).
- **Inference systems:** Optimized model placement and workload routing across device, edge, and cloud under latency, communication, and resource constraints (**ICMLCN 2026**).

Graduate Researcher, KAIST 2022–2024

- **Model pruning for heterogeneous federated learning:** Co-developed NeFL, using widthwise and depthwise model scaling to construct nested submodels for resource-constrained clients; also studied generative knowledge transfer (**IEEE TMC 2025, 2026**).

Industry Experience

Research Intern, Kilby Labs, Texas Instruments Summer 2026

- **Fixed-point training:** Designed memory-efficient training algorithms for microcontrollers with tight memory and compute budgets.
- **Continual learning:** Developed selective parameter updates to reduce the memory cost of on-device adaptation to changing data.

Undergraduate Summer Intern, SK hynix Summer 2019

Selected Publications

1. **S Cha**, H Chen, D Kim, H Zhang, K S Chan, G de Veciana, and H Vikalo, "CoreQ: Learning-Free Mismatch Correction and Successive Rounding for Quantization," *NeurIPS*, 2026. **Spotlight (top 5.1% of accepted papers).**
2. **S Cha***, H Chen*, and H Vikalo, "FedProTIP: Task-Agnostic Federated Continual Learning via Replay-Free Gradient Projection," *TMLR*, 2026.
3. **S Cha**, K S Chan, G de Veciana, and H Vikalo, "Joint Model Onloading and Offloading for Hierarchical Multi-Task Inference," *ICMLCN*, 2026.
4. H Kang*, **S Cha***, and J Kang, "GeFL: Model-Agnostic Federated Learning With Generative Models," *IEEE Transactions on Mobile Computing*, 2026.

*Equal contribution.

Additional Publications

1. H Zhang, D Kim, **S Cha**, and H Vikalo, "FedRot-LoRA: Mitigating Rotational Misalignment in Federated LoRA," *ICML*, 2026.
2. D Kim, **S Cha**, H Chen, C Wang, and H Vikalo, "Quantized Gradient Projection for Memory-Efficient Continual Learning," *ICLR*, 2026.
3. H Kang, **S Cha**, J Shin, J Lee, and J Kang, "NeFL: Nested Model Scaling for Federated Learning With System Heterogeneous Clients," *IEEE Transactions on Mobile Computing*, 2025.
4. **S Cha**, H Kang, and J Kang, "On the Temperature of Bayesian Graph Neural Networks for Conformal Prediction," *NeurIPS Workshop on New Frontiers in Graph Learning*, 2023.

Honors and Awards

- **Best Project Award**, ML on Real-World Networks, UT Austin 2024
  Project: Distributed Continual Learning Using Gradient Projection
- **Korean Government Scholarship**, KAIST Graduate Program 2022–2024
- **National Science and Engineering Scholarship**, Academic Excellence 2019–2021
- **Korean Government Scholarship**, KAIST Undergraduate Program 2017–2018

Teaching and Mentoring

Undergraduate Individual Study Assistant, KAIST 2023

Teaching Assistant, KAIST 2022–2023

- EE205 Data Structures and Algorithms for Electrical Engineering; EE966 M.S. Seminar / Colloquium.

Counseling Assistant, KAIST Sep 2022–Feb 2023

Freshman Tutor, KAIST 2018–2019

- MAS101 Calculus I and MAS102 Calculus II.

Skills

Programming: Python, C++, MATLAB, LaTeX

ML frameworks: PyTorch, TensorFlow, Hugging Face

Model optimization: Post-training quantization, quantization-aware training, widthwise/depthwise model pruning, fixed-point/integer training, parameter-efficient fine-tuning

Tools: Linux, Git, Pandas

Languages: Korean (native), English (proficient)