EpiCache: Enabling Long-Term Conversational Agents on Memory-Constrained Devices

Minsoo Kim
(Hanyang Univ)

SEMINAR ANNOUNCEMENT

- **Date & Time :** ***Feb. 6, 2026, 10:00~12:00***
- **Place :** ***Room 911, ITBT Bldg., Hanyang University***
- **Abstract:** Long-term conversational agents promise personalized and contextually coherent outputs, but remain impractical on mobile devices due to the unbounded memory growth from the KV cache. We introduce EpiCache, a training-free KV cache management framework that treats long conversations as a set of coherent episodes. By constructing episode-specific KV caches and dynamically matching each query to the most relevant episode, EpiCache preserves long-term conversational memory under fixed memory budgets.
- **Bio:** Minsoo Kim is a Ph.D. candidate in Electronics Engineering at Hanyang University. He received his B.S. from Hanyang University and has been a research intern at Qualcomm AI Research and Apple, where he worked on memory-efficient multimodal large language models. His research focuses on developing algorithms for deploying LLMs in resource-constrained environments, with interests in quantization, knowledge distillation, multimodality and long-context optimization.

Contact : kimtaeuk@hanyang.ac.kr **(컴퓨터소프트웨어학부 김태욱 교수)**