Bridging LLM and Big Data with

LMCache

vLLM instances ← LMCache → Contextual Data (text, video, audio)

Fast KV-Cache Blending
EuroSys'25

Fast KV-Cache Delivery
SIGCOMM'24

KV Cache Store