

Jiayu Qin

PhD Student in Computer Science and Engineering, University at Buffalo

✉ jiayuin@buffalo.edu 🌐 jiayujack.github.io 📄 github.com/jiayujack

Professional Summary

PhD student focused on **LLM serving systems**, **LLM post-training/RL**, and **AI for scientific applications**. Recent work includes **RebaseKV**, a training-free multi-agent LLM serving architecture that reduces TTFT by up to $3.9\times$; **BETA-RL**, a Bayesian reward-weighting method for multi-reward LLM RL; and **FinTOA**, an LLM-based financial topic-attention framework linking news narratives to macro factors and asset-pricing signals.

Education

University at Buffalo, SUNY

Ph.D. in Computer Science and Engineering

Buffalo, NY

08/2022 – Present

Fudan University

B.Eng. in Electrical Engineering

Shanghai, China

09/2018 – 06/2022

Second-class Scholarship (Top 5%)

Selected Research Experience

RebaseKV: Optimistic Cache Alignment for Multi-Agent LLM Serving **Under Review, 2026**

- Proposed **RebaseKV**, a training-free serving architecture that hides repeated prefill latency in multi-agent LLM workflows through **Optimistic Cache Alignment**.
- Designed a dual-stream execution strategy: decode a short optimistic prefix from inherited, semantically relevant KV states while asynchronously materializing the exact prompt-conditioned cache, then rebase onto the fresh cache for standard decoding.
- Achieved **up to $3.9\times$ TTFT reduction** across GSM8K, MMLU, and HumanEval while preserving answer-level accuracy; restored role-specific programming behavior from **5%** under stale reuse to **97%** after rebasing.

BETA-RL: Bayesian Reward Weighting for Multi-Reward LLM RL **Under Review, 2026**

- Developed **BETA-RL**, a Bayesian uncertainty-aware reward weighting framework that learns simplex-constrained per-reward precision weights for GRPO/GDPO-style LLM post-training.
- Proved that GDPO batch normalization makes the absolute scale of naive Beta-KD-style weights unidentifiable, and resolved the failure mode with $\beta_k = K \cdot \text{softmax}(\alpha)_k$ plus entropy regularization.
- Improved Pass@1 over GDPO on four math-reasoning benchmarks with DeepSeek-R1-Distill-Qwen-1.5B: **+2.74 average** without per-task tuning and **+5.21 average** when combined with reward conditioning, including **+80% relative** on AIME-24.

FinTOA: LLM Agent for Financial News Topic Extraction and Market Prediction **2026**

- Designed an **LLM-driven topic-attention agent** to extract themes from 40+ years of financial news (1979–2023), converting unstructured text into quantitative time-series features; the project was awarded **USD 100K** from Empire AI.
- Translated topic trends into quantitative attention scores that capture evolving investor focus on market drivers, then incorporated these features into time-series models for historical return prediction.
- Evaluated attention features in 6-month rolling-window stock-market backtests from 1979 to 2023, achieving **17% annual return**, **1.25 Sharpe ratio**, and **30.9% max drawdown**.

Dynamic KV-Cache Compression for LLM Serving **2025**

- Developed an incremental SVD-based framework to measure dataset-level singular spectra of KV caches across layers and models, enabling **globally informed low-rank projections** for LLM serving.
- Introduced a **Normalized Effective Rank** metric to predict per-layer compressibility and relate it to perplexity degradation, supporting **layer-aware, data-driven KV policies** instead of fixed global ratios.
- Built tooling that emulates realistic serving workloads across domains and context lengths to study **memory–bandwidth–quality** trade-offs for deployment.

Biomedical Knowledge Model for Molecule–Protein Interaction Prediction

ICLR 2026

- Developed a **multimodal biomedical foundation model** for molecule–protein interaction (MPI) prediction under **scarce labeled data**.
- Constructed a large-scale **biomedical knowledge graph** integrating molecular structures, protein sequences, pathways, and gene interaction networks.
- Proposed an **optimal transport-based alignment** method to infer pseudo-labels for unlabeled molecule–protein pairs using known biological interaction patterns.
- Demonstrated improved accuracy and **zero-shot generalization** on virtual screening and protein retrieval tasks compared to state-of-the-art MPI models.

GeoMamba-SE(3): Efficient Molecular Representation Learning

AISTATS 2026

- Proposed **GeoMamba-SE(3)**, a Mamba-based state-space architecture for molecular property prediction with **linear-time complexity** and practical **SE(3) stability**.
- Introduced **local-frame scalarization** and **dual-stream selective SSMs** to model molecular graphs and 3D geometry without heavy equivariant tensor representations.
- Enforced **statistical SE(3) invariance** via orbit-based contrastive learning, kernel mean embedding (MMD), and invariant risk minimization.
- Achieved **state-of-the-art performance** on MoleculeNet and QM9 benchmarks while reducing training cost by $\sim 6\times$ and inference memory by over 50%.

Probability Contrastive Learning for Graph Representation Learning

NeurIPS 2024

- Identified and formalized the **False Positive/Negative Pair problem** in graph contrastive learning caused by stochastic graph augmentations.
- Proposed a **Bayesian probability-based contrastive learning framework** that learns pairwise importance weights and automatically down-weights spurious positive and negative pairs.
- Outperformed existing approaches on **13/15 MoleculeNet** benchmarks and achieved state-of-the-art results on **8/12 QM9** molecular property prediction tasks.

Professional Experience

Applied Scientist Intern, Amazon

05/2026 – Present

- Research on **efficient LLM/LLM-agent inference acceleration**, including KV-cache analysis, compression, and evaluation under deployment-style constraints.

Visiting Researcher, Harvard University

02/2025 – 08/2025

- Collaborated on efficiency and reliability of large foundation models in realistic drug-discovery serving environments, including evaluation pipelines and failure analysis.

Visiting Researcher, MBZUAI

08/2024 – 01/2025

- Worked on Mamba/Transformer-based molecular representation learning with a focus on **memory- and throughput-aware** sequence modeling.

Machine Learning Engineer Intern, Galasports Co.

06/2023 – 08/2023

- Built and deployed prediction models on large user-behavior streams; improved pipelines, monitoring, and latency, gaining hands-on experience with **production ML systems**.

Technical Skills

LLM & Inference: KV-cache internals, prefix/non-prefix cache reuse, multi-agent serving, LoRA, post-training quantization.

Post-training & RL: GRPO/GDPO, multi-reward RL, reward modeling, reward conditioning, HuggingFace TRL.

Systems & Tools: Python, PyTorch, CUDA/Triton, HuggingFace, vLLM, SGLang, DeepSpeed.

ML & Theory: Representation learning, contrastive learning, graph learning, optimal transport, multimodal/scientific ML.