Novelty Assessment Report

Paper: Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective
PDF URL: https://openreview.net/pdf?id=R7fv5NWfMm
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

KV caching is a fundamental technique for accelerating Large Language Model (LLM) inference by reusing key-value (KV) pairs from previous queries, but its effectiveness under limited memory is highly sensitive to the eviction policy. The default Least Recently Used (LRU) eviction algorithm struggles with dynamic online query arrivals, especially in multi-LLM serving scenarios, where balancing query load across workers and maximizing cache hit rate of each worker are inherently conflicting objectives. We give the first unified mathematical model that captures the core trade-offs between KV cache eviction and query routing. Our analysis reveals the theoretical limitations of existing methods and leads to principled algorithms that integrate provably competitive randomized KV cache eviction with learning-based methods to adaptively route queries with evolving patterns, thus balancing query load and cache hit rate. Our theoretical results are validated by extensive experiments across 4 benchmarks and 3 prefix-sharing settings, demonstrating improvements of up to **6.92$\times$** in cache hit rate, **11.96$\times$** reduction in latency, **14.06$\times$** reduction in time-to-first-token (TTFT), and **77.4%** increase in throughput over the state-of-the-art methods.

Core Task Landscape

This paper addresses: **KV Cache Eviction and Query Routing in Multi-LLM Serving Systems**

A total of **25 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **KV Cache Management and Eviction Strategies**
- **Query Distribution and Routing**
- **Distributed and Disaggregated LLM Serving Architectures**
- **Multi-Stage and Multi-Model Serving Pipelines**
- **Scheduling and Resource Management**
- **Surveys and Overviews**
- **Non-LLM Serving Systems**

Complete Taxonomy Tree

- KV Cache Eviction and Query Routing in Multi-LLM Serving Systems Survey Taxonomy
- KV Cache Management and Eviction Strategies
  - Cache Eviction Policies and Optimization ★ (2 papers)
  - [0] Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective (Anon et al., 2026) View paper
  - [1] Queue management for slo-oriented large language model serving (Archit Patke, 2024) View paper
  - Cache Storage and Memory Hierarchy (2 papers)
  - [2] Flashinfer: Efficient and customizable attention engine for llm inference serving (Ye Zihao, 2025) View paper
  - [4] Strata: Hierarchical context caching for long context language model serving (Xie Zhi-qiang, 2025) View paper
  - Context and Semantic Cache Reuse (2 papers)
  - [7] Ic-cache: Efficient large language model serving via in-context caching (Yu Yifan, 2025) View paper
  - [10] Online Context Caching for Distributed Large Language Models Serving (Bin Gao, 2025) View paper
  - Cache Sharing and Security (2 papers)
  - [14] CIFLEX: Contextual Instruction Flow for Sub-task Execution in Multi-Turn Interactions with a Single On-Device LLM (Juntae Lee, 2025) View paper
  - [22] SafeKV: Safe KV-Cache Sharing in LLM Serving (K Chu, n.d.) View paper
  - Heterogeneous Precision and Quantized Cache Management (1 papers)
  - [21] FineServe: Precision-Aware KV Slab and Two-Level Scheduling for Heterogeneous Precision LLM Serving (Choi Seungbeom, 2025) View paper
- Query Distribution and Routing
  - Multi-Region and Cross-Region Load Balancing (1 papers)
  - [6] SkyLB: A Locality-Aware Cross-Region Load Balancer for LLM Inference (Xia Tian, 2025) View paper
  - Cache-Aware Request Routing (2 papers)
  - [13] Unified Gateway Architecture For Multi-Tenant Large Language Model Serving (Cheekuri, 2025) View paper
  - [17] Scalable Scheduling and Intelligent Resource Optimization for Efficient Large Language Model Inference Acceleration (Cheung, 2025) View paper
  - Adaptive and Learning-Based Routing (1 papers)
  - [18] Systematic Technical Survey on LLMOps: Lifecycle, Tools, Challenges, and Emerging Practices (Ã□zer, 2025) View paper

- Distributed and Disaggregated LLM Serving Architectures
  - Disaggregated Prefill-Decode Architectures (1 papers)
  - [12] BanaServe: Unified KV Cache and Dynamic Module Migration for Balancing Disaggregated LLM Serving in AI Infrastructure (He, 2025) View paper
  - Distributed Inference and Partitioning (3 papers)
  - [3] A Scalable Approach to Distributed Large Language Model Inference (Yihua, 2025) View paper
  - [5] Infinite-LLM: Efficient LLM Service for Long Context with DistAttention and Distributed KVCache (Lin Bin, 2024) View paper
  - [11] Efficient distributed LLM inference with dynamic partitioning (Ong, 2024) View paper
  - Heterogeneous Deployment and Cost Optimization (1 papers)
  - [16] Cauchy: A Cost-Efficient LLM Serving System through Adaptive Heterogeneous Deployment (Y Zhang, 2025) View paper
- Multi-Stage and Multi-Model Serving Pipelines
  - Multi-Stage Inference Pipelines (2 papers)
  - [8] An Adaptive Vector Index Partitioning Scheme for Low-Latency RAG Pipeline (J Kim, 2025) View paper
  - [15] Understanding and Optimizing Multi-Stage AI Inference Pipelines (Bambhaniya, 2025) View paper
  - Adapter and Expert-Specialized Serving (1 papers)
  - [24] ExpertWeave: Efficiently Serving Expert-Specialized Fine-Tuned Adapters at Scale (Shi Ge, 2025) View paper
- Scheduling and Resource Management (1 papers)
  - [23] LLM Inference Scheduling: A Survey of Techniques, Frameworks, and Trade-offs (Morgan Heisler, 2025) View paper
- Surveys and Overviews (2 papers)
  - [9] A Survey of LLM Inference Systems (Pan James, 2025) View paper
  - [20] Demystifying LLM Serving Pipeline: From Prompt to Response (Kumar, 2025) View paper
- Non-LLM Serving Systems (2 papers)
  - [19] Understanding Diffusion Model Serving in Production: A Top-Down Analysis of Workload, Scheduling, and Resource Efficiency (Y Lin, 2025) View paper
  - [25] Cache-based Executive Request Dispatching Method in The Distributed Workflow System (Bo Lv, 2021) View paper

Narrative

Core task: KV cache eviction and query routing in multi-LLM serving systems. The field addresses the dual challenge of managing memory-intensive key-value caches during inference and intelligently distributing queries across multiple large language model instances. The taxonomy reveals several major branches: KV Cache Management and Eviction Strategies focuses on policies that decide which cached tokens to retain or discard, often trading memory footprint against recomputation cost; Query Distribution and Routing explores how to assign incoming requests to appropriate model replicas or endpoints; Distributed and Disaggregated LLM Serving Architectures examines system designs that separate compute from storage or span multiple nodes; Multi-Stage and Multi-Model Serving Pipelines considers workflows involving cascades of smaller and larger models; Scheduling and Resource Management tackles broader orchestration questions such as batching and GPU allocation; and Surveys and Overviews provide synthetic perspectives on inference optimization. Representative works like Strata Hierarchical Caching[4] and Infinite-LLM[5] illustrate memory hierarchy techniques, while SkyLB[6] and Unified Gateway[13] exemplify routing and load-balancing approaches.

A particularly active line of work centers on cache eviction policies that balance hit rates with service-level objectives, as seen in Queue Management SLO[1], which integrates eviction decisions with queueing dynamics. Another contrasting theme is hierarchical or tiered caching—Strata Hierarchical Caching[4] and In-Context Caching[7] explore multi-level storage to exploit locality—versus online adaptive schemes like Online Context Caching[10] that adjust eviction on the fly. The original paper, Randomization KV Caching[0], sits squarely within the Cache Eviction Policies and Optimization cluster, proposing randomized selection mechanisms to reduce deterministic bias in token retention. Its emphasis on stochastic eviction distinguishes it from deterministic priority schemes in Queue Management SLO[1] and from the hierarchical staging strategies in Strata Hierarchical Caching[4], offering a complementary angle on managing cache churn under variable workloads.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Queue management for slo-oriented large language model serving

Authors: Archit Patke, Dhemath Reddy, Saurabh Jha, Haoran Qiu, Christian Pinto, et al. (9 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ KV cache of batch requests, allowing execution to resume from the last decoding iteration. â¦ its associated virtual queue, thus ensuring the distribution of requests across all the serving â¦

Relationship Analysis

Both papers belong to the Cache Eviction Policies and Optimization category, addressing KV cache management under memory constraints in LLM serving systems. They share overlapping concerns with cache eviction strategies and their impact on system performance metrics like latency and throughput. However, the original paper focuses on the joint optimization of randomized eviction policies (RLT) and learning-based query routing (LBGR) across multiple LLMs with theoretical competitive ratio analysis, while the candidate paper appears to emphasize SLO-oriented queue management and virtual queue mechanisms for request distribution, representing a different approach to managing cache and workload in multi-LLM serving.

Contributions Analysis

Overall novelty summary. The paper proposes a unified mathematical model integrating KV cache eviction with query routing in multi-LLM serving, alongside two algorithms: Randomized Leaf Token (RLT) eviction and Learning-Based Greedy Routing (LBGR). It resides in the 'Cache Eviction Policies and Optimization' leaf, which contains only one sibling paper (Queue Management SLO). This leaf sits within the broader 'KV Cache Management and Eviction Strategies' branch, indicating a moderately sparse research direction focused specifically on eviction policy design rather than storage formats or semantic reuse.

The taxonomy reveals neighboring leaves addressing complementary aspects: 'Cache Storage and Memory Hierarchy' explores tiered memory layouts (e.g., Strata Hierarchical Caching), 'Context and Semantic Cache Reuse' targets similarity-based retrieval, and 'Cache-Aware Request Routing' examines routing algorithms that consider cache states. The original work bridges eviction and routing—a cross-cutting concern—by jointly optimizing both dimensions. Its randomized eviction strategy contrasts with deterministic priority schemes in sibling work and hierarchical staging approaches in adjacent leaves, positioning it at the intersection of cache management and query distribution.

Among seven candidates examined, no contribution was clearly refuted. The unified model (one candidate examined) and LBGR routing (two candidates) show no overlapping prior work in the limited search. RLT eviction (four candidates examined) likewise encountered no

refutations, suggesting that randomized token-level eviction combined with load-aware routing represents a relatively unexplored angle. The search scope—seven papers from top-K semantic matches—is narrow, so these findings reflect novelty within a constrained sample rather than exhaustive field coverage.

Given the limited search and sparse taxonomy leaf, the work appears to occupy a distinct niche by unifying eviction and routing under a single framework. The absence of refutations across all contributions, combined with the small sibling set, suggests the approach is not directly anticipated by the examined literature. However, the narrow candidate pool means adjacent or emerging work outside the top-seven matches could still present relevant comparisons.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Unified mathematical model for KV cache-aware load balancing

Description: The authors present the first formal model that jointly captures the tensions between local KV cache eviction policies and global query routing across multiple LLMs. This model decomposes end-to-end latency into service time and queuing delay, enabling principled analysis of the makespan minimization problem.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. BanaServe: Unified KV Cache and Dynamic Module Migration for Balancing Disaggregated LLM Serving in AI Infrastructure

URL: View paper

Brief Assessment

BanaServe[12] focuses on disaggregated LLM serving with dynamic module migration and global KV cache stores, not on unified mathematical models for KV cache eviction and query routing trade-offs as formulated in the original paper.

Contribution 2: Randomized Leaf Token (RLT) eviction algorithm

Description: The authors propose RLT, a randomized eviction algorithm that uniformly selects unmarked leaf tokens for eviction. This approach achieves an O(log n) competitive ratio, which is exponentially better than the O(n) ratio of existing LRU-based policies, and is proven optimal among randomized algorithms.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Marconi: Prefix caching for the era of hybrid llms

URL: View paper

Brief Assessment

Marconi[26] focuses on prefix caching for hybrid LLMs (attention + SSM layers) with admission/eviction policies based on reuse likelihood and FLOP efficiency, not on randomized eviction algorithms for prefix-sharing KV cache structures.

2. Transcending Cost-Quality Tradeoff in Agent Serving via Session-Awareness

URL: View paper

Brief Assessment

Session-Aware Agent Serving[29] focuses on agent serving systems with session-aware KV cache management and model cascading, not on randomized eviction algorithms for prefix-sharing structures. The candidate's eviction policy (ETA-based) differs fundamentally from RLT's randomized approach.

3. Prefix and Output Length-Aware Scheduling for Efficient Online LLM Inference

URL: View paper

Brief Assessment

Prefix Output Scheduling[28] focuses on prefix-aware and output length-aware scheduling for request routing across GPUs, not on randomized eviction algorithms for KV cache management within individual workers.

4. Learned Prefix Caching for Efficient LLM Inference

URL: View paper

Brief Assessment

Learned Prefix Caching[27] focuses on learned eviction policies using conversation continuation prediction for prefix caches, not randomized eviction algorithms for prefix-sharing KV cache structures. The technical approaches are fundamentally different: RLT uses randomization to achieve competitive ratios in radix tree structures, while Learned Prefix Caching[27] uses ML-based prediction of conversation patterns.

Contribution 3: Learning-Based Greedy Routing (LBGR) algorithm

Description: The authors introduce LBGR, which uses online linear regression to estimate per-LLM end-to-end latency by combining service time estimation, exponentially decaying queue load tracking, and learned residual correction. Queries are then routed greedily to the LLM with minimal predicted latency, providing adaptability to evolving query patterns.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Chameleon: predictable latency and high utilization with queue-aware and adaptive source routing

URL: View paper

Brief Assessment

Chameleon[31] focuses on network routing with queue-aware path selection for latency guarantees in datacenters, not on LLM inference routing with KV cache state estimation and online regression for load balancing.

2. Efficient Cache Retrieving Scheme in VANET via Online Reinforcement Learning

URL: View paper

Brief Assessment

VANET Cache Retrieving[30] focuses on cache retrieval scheduling in vehicular networks using double-ε online reinforcement learning for RSU prioritization, not on multi-LLM query routing with latency prediction based on service time and queue load estimation.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Randomization Boosts KV Caching, Learning Balances Query Load: A Joint Perspective View paper
- [1] Queue management for slo-oriented large language model serving View paper
- [2] Flashinfer: Efficient and customizable attention engine for llm inference serving View paper
- [3] A Scalable Approach to Distributed Large Language Model Inference View paper
- [4] Strata: Hierarchical context caching for long context language model serving View paper
- [5] Infinite-LLM: Efficient LLM Service for Long Context with DistAttention and Distributed KVCache View paper
- [6] SkyLB: A Locality-Aware Cross-Region Load Balancer for LLM Inference View paper
- [7] Ic-cache: Efficient large language model serving via in-context caching View paper
- [8] An Adaptive Vector Index Partitioning Scheme for Low-Latency RAG Pipeline View paper
- [9] A Survey of LLM Inference Systems View paper
- [10] Online Context Caching for Distributed Large Language Models Serving View paper
- [11] Efficient distributed LLM inference with dynamic partitioning View paper
- [12] BanaServe: Unified KV Cache and Dynamic Module Migration for Balancing Disaggregated LLM Serving in AI Infrastructure View paper
- [13] Unified Gateway Architecture For Multi-Tenant Large Language Model Serving View paper
- [14] CIFLEX: Contextual Instruction Flow for Sub-task Execution in Multi-Turn Interactions with a Single On-Device LLM View paper
- [15] Understanding and Optimizing Multi-Stage AI Inference Pipelines View paper
- [16] Cauchy: A Cost-Efficient LLM Serving System through Adaptive Heterogeneous Deployment View paper
- [17] Scalable Scheduling and Intelligent Resource Optimization for Efficient Large Language Model Inference Acceleration View paper
- [18] Systematic Technical Survey on LLMOps: Lifecycle, Tools, Challenges, and Emerging Practices View paper
- [19] Understanding Diffusion Model Serving in Production: A Top-Down Analysis of Workload, Scheduling, and Resource Efficiency View paper
- [20] Demystifying LLM Serving Pipeline: From Prompt to Response View paper
- [21] FineServe: Precision-Aware KV Slab and Two-Level Scheduling for Heterogeneous Precision LLM Serving View paper
- [22] SafeKV: Safe KV-Cache Sharing in LLM Serving View paper
- [23] LLM Inference Scheduling: A Survey of Techniques, Frameworks, and Trade-offs View paper
- [24] ExpertWeave: Efficiently Serving Expert-Specialized Fine-Tuned Adapters at Scale View paper
- [25] Cache-based Executive Request Dispatching Method in The Distributed Workflow System View paper
- [26] Marconi: Prefix caching for the era of hybrid llms View paper
- [27] Learned Prefix Caching for Efficient LLM Inference View paper
- [28] Prefix and Output Length-Aware Scheduling for Efficient Online LLM Inference View paper
- [29] Transcending Cost-Quality Tradeoff in Agent Serving via Session-Awareness View paper
- [30] Efficient Cache Retrieving Scheme in VANET via Online Reinforcement Learning View paper
- [31] Chameleon: predictable latency and high utilization with queue-aware and adaptive source routing View paper