Thesis Proposal: Efficient KV Cache Reuse for Multi-Document Retrieval-Augmented Generation

Zhipeng Zhang and Dmitry Ilvovsk
HSE University
{chzhan.chzhipen, dilvovsky}@hse.ru

Abstract

Retrieval-Augmented Generation (RAG) systems face efficiency bottlenecks in prefill due to attention mechanism, and traditional KV cache only accelerates decoding. In this context, reusing document-level KV cache computed for retrieved documents in previous sessions during the prefill stage appears to be a natural way to amortize computation, but it raises serious correctness challenges due to position and context misalignment across queries and sessions. This research proposes a multi-document KV cache reuse framework for multi-document RAG workloads across queries and sessions to resolve position misalignment and context misalignment, preserving accuracy while eliminating document-specific quadratic complexity in prefill. Theoretical analysis will establish conditions under which multi-document KV cache reuse remains stable and close to full recomputation, providing principled guarantees for both efficiency and accuracy. These results will enable deployment in existing RAG pipelines without architectural changes or model retraining. Crucially, to ensure robustness in real-world deployments, validation will extend beyond standard benchmarks to include noise-robustness tests and domain-specific workloads (e.g., legal). The research aims to empirically confirm these guarantees and demonstrate that substantial prefill speedups can be achieved without materially degrading task-level performance.

1 Introduction

RAG (Gao et al., 2024) combines Large Language Models (LLMs) (Zhao et al., 2025) with external knowledge sources to tackle knowledge-intensive tasks. In a typical RAG pipeline, the system first retrieves several chunks from external corpora based on the current user query. Then, the system prompt (p), the user query (q), and the retrieved documents ($\{d_i\}_{i=1}^{n}$) are concatenated into a long input sequence ($L = [p, q, d_1, \ldots, d_n]$), which is processed by LLMs through the self-attention mechanism (Vaswani et al., 2023). When processing such a long sequence, the model typically runs a prefill stage, where the entire input context is consumed once to compute hidden states and populate the KV cache, followed by a decoding stage, where output tokens are generated autoregressively conditioned on this cached state. In this proposal, multi-document RAG refers to settings where each query is served with multiple (often long) retrieved documents and where the same documents are repeatedly used across different queries and sessions, rather than being consumed once in a single request as in standard RAG systems. In such multi-document RAG workloads, the computational cost of the prefill stage grows quadratically with the sequence length, becoming a key performance bottleneck when systems are deployed with long documents and high concurrency.

KV cache has become a standard technique for accelerating autoregressive decoding. By caching the KV representations of previously processed tokens, the complexity of per-step decoding can be reduced from $O(L^2)$ to $O(L)$, where L is the length of input tokens. However, this optimization mainly targets decoding phase and does not mitigate the quadratic prefill overhead for processing long contexts for the first time. In scenarios involving multi-turn conversations, enterprise knowledge-based Q&A, or applications with relatively stable document repositories, the same document is often repeatedly retrieved and used. Recomputing full self-attention over these documents in every request wastes significant computational resources and degrades user-perceived latency.

This observation naturally motivates multi-document KV cache reuse. In this work, multi-document KV cache reuse means that during the first prefill of a document, the system computes and stores document-level KV cache for each Trans-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 160–169

former layer and attention head. For subsequent queries, whenever RAG system selects the document again, regardless of its relative position in the current input sequence or changes in the preceding context (p, q, $\{d_i\}_{i=1}^{n}$), the system tries to directly reuse the existing document-level KV cache instead of re-performing a complete self-attention calculation for that document. This reuse happens across requests, across sessions, and under different document combinations, going substantially beyond traditional within-request KV cache reuse. If it can be made reliable, such reuse promises to amortize the document-related quadratic prefill cost over queries, thereby reducing Time-to-First-Token (TTFT), increasing throughput, and improving hardware utilization for long-context RAG systems.

However, existing work indicates that naively reusing document-level KV cache leads to several fundamental issues that threaten model correctness. First, the position is misaligned because the KV cache has an old position encoding, and the position is different in this round. Second, context misalignment, as from the second decoder block onward, the document token's hidden state (and hence its KV cache) depends on the current left context (p, q, and earlier documents), while the KV cache does not undergo this round of conditioning. Third, error amplification and propagation, which results from the stacking of residuals accumulate attention bias (Xiong et al., 2020).

Recent work has explored cross queries and sessions document-level KV cache reuse from both system and algorithmic perspectives. While systems like CacheBlend (Yao et al., 2025) have demonstrated the engineering feasibility of correcting positional encodings (e.g., via inverse RoPE (Su et al., 2023) rotation), they rely on heuristics for stability control. Consequently, these works lack a unified theoretical characterization, failing to clearly define when reuse is safe and how errors arising from it propagate within the model. This theoretical gap makes it difficult to provide interpretable safe reuse boundaries for high-risk applications such as medical or legal Q&A. The research aims to fill precisely this gap. The main technical question is whether, and under what verifiable conditions, one can formalize multi-document KV cache reuse as a controlled structural perturbation of the ideal Transformer computation, and then design positional alignment operators, attention stability bounds, and depth-wise convergence control mechanisms so that multi-document KV cache reuse substantially reduces prefill cost and TTFT while keeping task-level accuracy close to full recomputation.

2 Related Work

2.1 RAG Caching and Efficiency Optimization

The pursuit of efficiency in RAG systems has led to several system-level and algorithmic innovations. RAGCache (Jin et al., 2025) represents a groundbreaking approach to optimizing RAG systems through intelligent caching of intermediate KV states. By organizing cached documents in a knowledge tree structure and implementing a prefix-aware Greedy-Dual-Size-Frequency replacement policy, RAGCache demonstrates that caching frequently accessed documents can reduce TTFT by up to 4× and improve throughput by 2.1× compared to standard vLLM and Faiss integrations. This work provides strong empirical evidence for the performance benefits of reusing computed KV states, but it largely assumes that cached KV is functionally correct once retrieved and does not analyze how reuse interacts with positional encodings or changing left contexts, nor does it specify conditions under which such reuse is safe.

A concurrent approach to RAG optimization is presented in CacheBlend (Yao et al., 2025), which tackles the latency problem in the prefill stage by reusing pre-computed KV cache for multiple text chunks. Unlike methods that only reuse caches when they form the input prefix, CacheBlend enables the reuse of KV cache regardless of their position in the current input. To address the critical issue of missing cross-attention with preceding texts, it selectively recomputes the KV cache for a small subset of tokens to update each reused cache. This approach allows the extra recomputation delay being pipelined with KV cache retrieval, enabling the use of slower, higher-capacity storage without increasing inference latency. CacheBlend demonstrates substantial performance gains, reducing TTFT by 2.2-3.3× and increasing inference throughput by 2.8-5× compared to full recomputation, without compromising generation quality. However, the choice of which tokens to recompute and how much approximation is acceptable remains heuristic, and the method does not provide a formal characterization of when partial recomputation keeps attention distributions and outputs within a

controlled deviation from ideal full recomputation.

Further expanding the efficiency frontier, LightMem (Fang et al., 2025) introduces a cognitive-inspired memory system for LLMs that processes information through sensory, short-term, and long-term memory stages. Its offline updating mechanism for long-term memory, decoupled from online inference, enhances adaptability and can reduce token usage by up to 117×. This bio-inspired architecture offers a broader perspective on efficient knowledge management, but it mainly concerns how to organize and update memory rather than how to safely reuse document-level KV states under changing prompts and document orderings.

Recent work on TeleRAG (Lin et al., 2025) introduces lookahead prefetching to optimize multi-turn RAG conversations. By prefetching relevant IVF clusters during pre-generation and employing GPU-CPU hybrid vector search, TeleRAG demonstrates the benefits of overlapping retrieval with generation. However, this approach focuses on optimizing the retrieval step rather than addressing the computational burden of processing retrieved documents, and it does not consider the correctness or stability of reusing precomputed KV cache for those documents.

2.2 KV Cache Management and Efficient Inference Techniques

This proposal builds on foundational research in KV cache management for accelerating LLM inference. The vLLM system (Kwon et al., 2023) employs PagedAttention to manage KV cache in non-contiguous memory blocks, enabling efficient memory sharing and reducing fragmentation. Similarly, SGLang (Zheng et al., 2023) identifies and reuses intermediate states across different requests within GPU memory. While these systems optimize KV cache management within a single request or across similar requests, they do not specifically address the unique challenges of reusing document-level caches across different queries and sessions in RAG environments.

Recent efforts have explored more aggressive strategies for KV cache reuse across requests. CacheGen (Liu et al., 2024) compresses the KV cache to reduce its memory footprint and transmission overhead, enabling efficient reuse in bandwidth-constrained environments. However, such compression-based methods inevitably introduce approximation errors that may propagate across layers and affect output fidelity. These approaches underscore the inherent trade-off between efficiency and accuracy in KV cache management. The trade-off is currently handled empirically, without explicit guarantees on how compression-induced perturbations influence multi-layer attention and final predictions.

For long-context inference, StreamingLLM (Xiao et al., 2024) maintains stable performance for infinite-length inputs without fine-tuning by preserving attention sinks and a sliding window of recent tokens. Its analysis of attention distribution patterns in extended contexts provides useful insight into which attention heads are most sensitive to positional and contextual variations. However, StreamingLLM targets streaming input scenarios rather than offline reuse of document-level KV cache across sessions and does not provide a general theory of when precomputed KV cache can be safely reused under new input configurations.

2.3 Synthesis and Positioning of Research Proposal

Existing literature shows that retrieval-enhanced LLMs are developing towards increasingly complex caching and reuse mechanisms. It is worth noting that RAGCache and CacheBlend do involve the cross-session reuse of document-level KV cache, which reflects the important research value of this direction. RAGCache achieves the reuse of document KV cache across different sessions through a knowledge tree structure and prefix-aware cache replacement strategy, but its core assumption is that cached documents can be reused directly. CacheBlend solves the problem of cross-attention loss when reusing non-prefix positions by selectively recomputing the KV cache of some tokens. However, these methods have a common theoretical limitation, which they are all based on engineering heuristics rather than strict theoretical guarantees.

Compared to the empirical approaches of RAGCache and CacheBlend, our research is fundamentally different. Firstly, our research formulates the KV cache reuse problem as a mathematical problem involving controlled structured perturbations. While CacheBlend implicitly uses inverse rotation, I propose to formalize this as a Position Propagation Operator that provides theoretically exact alignment for RoPE and ALiBi. Secondly, stability analysis based on the Softmax Jacobian matrix (Qi et al., 2023) provides quantifiable bounds on the changes in attention distribution. Thirdly, deep

error propagation control, through contraction analysis and an adaptive gating mechanism, ensures that inter-layer errors do not amplify unboundedly. Therefore, my framework provides provable security guarantees. Not only do I achieve similar efficiency gains, but more importantly, I ensure that, under certain conditions, the quality of the reused output remains theoretically bounded compared to a completely recomputed result.

3 Problem Analysis and Solution Framework

3.1 Problem Formalization

This study formalizes multi-document KV cache reuse as a structured perturbation problem relative to the ideal Transformer's attention mechanism. In ideal computation, the system constructs queries ($Q = W_Q \cdot x$), keys ($K = W_k \cdot x$), and values ($V = W_V \cdot x$) at each layers and attention heads based on the exact concatenated input $x = [p, q, d_1, \ldots, d_n]$, forming the standard attention distribution $\text{Softmax}((Q \cdot K^T)/\sqrt{d_k}) \cdot V$.

In the cache reuse paradigm, the keys ($\tilde{K}$) and values ($\tilde{V}$) for documents are assembled from precomputed caches, while the keys and values for the prompt (p) and query (q) are computed in real time. This computational divergence alters attention logits from the ideal $s = (q^T \cdot k)/\sqrt{d_k}$ to the perturbed $\hat{s} = (q^T \cdot \tilde{k})/\sqrt{d_k}$.

Through rigorous mathematical analysis, we decompose the total deviation into two independent components, including position-induced discrepancy, which arises because cached documents appear at different positions in the new input sequence compared to their original computation, and context-conditioned discrepancy, which originates from a deeper architectural dependency. Starting from the second decoder block, document token hidden states (and their corresponding KV cache) depend on the current left context (p, q, and earlier documents), while cached KV states lack this round of contextual conditioning. This precise identification and separation of the two discrepancy sources provide the theoretical foundation for designing targeted solutions to address them effectively.

3.2 Proposed Solutions

3.2.1 Position Alignment

The research will design a family of position transport operators to achieve exact or provably approximate alignment for major positional encoding schemes. The feasibility of this approach is supported by the algebraic properties of modern position encodings and recent engineering validations in systems like CacheBlend.

RoPE's Block Rotation Correction
In RoPE, each position i is associated with a rotation matrix R_i:

$$R_i = \begin{pmatrix} \cos(i\theta_0) & -\sin(i\theta_0) \\ \sin(i\theta_0) & \cos(i\theta_0) \end{pmatrix}$$

Where θ_0 is a rotation angle derived from the dimension d. To map cached keys from old position θ_{cache} to new position θ_{actual}, the research will construct position transmission operator $R(\theta_{\text{actual}}) \cdot R(\theta_{\text{cache}})^{-1}$. Since R_i is an orthogonal rotation matrix, this operator is mathematically fully invertible ($R^{-1} = R^T$), ensuring that position information can be corrected exactly without numerical approximation, as empirically utilized in CacheBlend.

ALiBi's Relative Distance Bias Reconstruction
The research will leverage ALiBi's (Press et al., 2022) property of applying positional biases only during the attention score calculation. The reuse mechanism will dynamically reconstruct the linear, distance-based additive bias $b_{(i,j)}$ based on the new relative positions, ensuring positional scoring is exact.

Absolute Position Linear Correction
Construct position feature bases $\phi(\text{pos})$ to decompose positional effects as $W_K(x + \text{PE}(\text{pos})) \approx W_K(X) + A_l\phi(\text{pos})$ (similarly for V). Implement "subtract old, add new" corrections $K_{\text{new}} \approx K_{\text{cache}} - A_l\phi(\text{pos}_{\text{cache}}) + A_l\phi(\text{pos}_{\text{actual}})$, with uniform residual bounds over long contexts.

3.2.2 Context Stability

A stability analysis theory is developed based on the Jacobian of the Softmax function to bound the difference between ideal attention outputs and those using reused KV cache.

For a query vector with bounded norm $\|q\|_2 \leq B_q$, perturbations Δk_j in key vectors induce logit perturbations $\Delta s = (q \cdot k_j^T)/\sqrt{d_k}$. By evaluating the Jacobian $J(p) = \text{diag}(p) - pp^T$ at the ideal attention distribution p, a data-dependent constant $c(p)$ is derived such that the L1 deviation of the attention distribution satisfies $\|\tilde{p} - p\|_1 \leq c(p) \cdot \max_j |\Delta s_j|$. Notably, sharper attention distributions (with larger margins μ) yield smaller $c(p)$, tightening the bound.

Token-level bounds are further elevated to document-level guarantees. By partitioning attention indices per document, the total attention mass $M_i(d) = \sum_{j \in d} p_i(j)$ for document i is bounded via a weighted sum of per-token score perturbations. This provides theoretical assurance for document-level semantic consistency.

Additionally, the margin preservation condition is proposed. If the Top-1 logit margin μ exceeds twice the maximum score perturbation, the Top-1 token identity (and under mild aggregation assumptions, the Top-1 document) remains unchanged. The value vector V's impact is decomposed into a probability shift term and a value perturbation term, both bounded by estimable norms. This quantifies the full error propagation path from attention to output.

3.2.3 Error Propagation Control

To prevent error amplification across Transformer layers, contraction theory is combined with an adaptive gating mechanism. Each Transformer layer is abstracted as $h_{l+1} = h_l + F_l(h_l)$, and the error propagation formula is derived as $\epsilon_{l+1} \leq L_{\text{res},l} \cdot (\epsilon_l + \delta_l)$, where ϵ_l is the input representation deviation at layer l and δ_l is the attention error bounded by stability analysis. Attention outputs are decomposed into prefix contributions (from p and q) and document contributions, and a document gating scalar $\gamma_l \in (0, 1]$ is introduced to scale the document portion in the residual.

By constructing the inequality $L_{\text{res},l} \cdot ((1 - \alpha_l) + \gamma_l \alpha_l) < 1$ (where α_l is the document contribution ratio), the feasible range for γ_l is determined to ensure contractive layer-wise mappings. For deep networks where convergence conditions are hard to satisfy, a "Top-R lightweight recomputation" variant is analyzed: Only recomputing the Top-R layers for document tokens using the full current left context (p, q, and earlier documents). This drastically reduces δ_l in critical final layers to near zero, tightening the end-to-end geometric deviation bound. Through this multi-level control strategy, reuse-induced errors are bounded within acceptable limits even for deep networks, providing reliability guarantees for practical deployment.

4 Methodology

4.1 Key Technical Components

The research introduces three core mathematical and algorithmic tools to ensure provably safe KV cache reuse.

A family of positional transformation operators enables precise position alignment: for RoPE, relative rotations (including inverse-rotation followed by forward rotation) map positional information exactly; for ALiBi, attention scores are recalibrated via bias reconstruction leveraging its position-neutral K/V properties; for absolute positional encoding, layer-wise linear calibration functions are designed with unified residual bounds, all applicable at minimal inference cost to convert position-induced errors into zero or bounded perturbations.

An attention stability toolkit based on the Softmax Jacobian provides theoretical guarantees by deriving data-dependent L1 bounds adaptive to attention sharpness, aggregating token-level perturbations into document-level mass deviation bounds and enforcing ranking invariance via a margin preservation condition under limited perturbations.

A depth-aware convergence control mechanism elevates local stability to end-to-end representation drift guarantees by decomposing attention outputs into prefix/document contributions, constructing feasible ranges for per-layer gating factors γ_l to satisfy contraction conditions $L_{\text{res},l} \cdot ((1 - \alpha_l) + \gamma_l \alpha_l) < 1$, and combining this with a "Top-R lightweight recomputation" strategy triggered by bound violations or margin diagnostics to ensure controlled error propagation in deep networks.

4.2 Algorithm Overview

The methodology establishes a unified pipeline from theory to practice through positional transformation, stability bounds, and convergence control.

In the offline phase, KV cache is built for each document, layer, and attention head: for RoPE, either "content keys" or "pre-rotated keys" are cached; for absolute positional encoding, layer-wise linear calibration functions g_l^K / g_l^V are estimated via least squares or self-distillation over target model/tokenizer pairs, with unified residual bounds derived as positional certificates. Key constants (e.g., query norms $\|q\|$, data-dependent constants $c(p)$, operator norm bounds) are calibrated, and layer-wise error budgets ϵ_l (summing to a global ϵ) are allocated geometrically, while "Top-R lightweight recomputation" thresholds are set based on attention margins to handle high-risk scenarios.

In the online phase, given prompt p, query q, and retrieved documents, the system computes

$Q/K/V$ for p and q, retrieves document caches, applies positional transformations (RoPE: inverse-rotate pre-rotated keys to content keys then re-rotate to new positions; ALiBi: bias reconstruction; absolute positioning: "subtract old, add new" calibration with certified parameters), assembles $\tilde{K}/\tilde{V}$ for attention computation, and dynamically monitors safety via parallel ideal recomputation on select positions. When document attention ratios are small and convergence conditions are met, minimal γ_l satisfying $L_{\text{res},l} \cdot ((1-\alpha_l) + \gamma_l \alpha_l) < 1$ is chosen per layer; when document margins shrink or bounds show negative slack, "Top-R lightweight recomputation" reduces δ_l in upper layers.

To address the efficiency concerns regarding the fallback mechanism, the system employs an "Adaptive Trigger" policy. The bounds derived from the Jacobian analysis act as the runtime decision metric. Specifically, we define a stability threshold τ; if the estimated error bound ϵ_l exceeds τ, it indicates that the reused cache has deviated dangerously from the ideal distribution. This event automatically triggers the Top-R recomputation for the affected tokens. This ensures that recomputation is reserved for high-risk inputs where the cache quality is demonstrably degraded, thereby maximizing the effective speedup.

Parameters (e.g., γ_l, R) are selected based on measurable quantities - γ_l minimizes information suppression within feasible ranges, R is chosen via margin diagnostics (typically 2-4 layers suffice to near-zero δ_l) - ensuring positional alignment is exact or provably approximate, fully decoupling position-induced terms while stability and convergence mechanisms control remaining discrepancies.

5 Validation Strategy

5.1 Experimental Setup

This research will build a comprehensive experimental verification system, with systematic planning from model selection and benchmark to specific implementation details. In model selection, we plan to adopt mainstream open-source causal decoder-only models with moderate parameter scales. Specifically, we plan to use Meta's Llama and Alibaba's Qwen families as the main experimental subjects, focusing on variants in the 7B–30B range that can be hosted on a single NVIDIA Tesla A100 80Gb GPU without excessive engineering effort. However, to address concerns regarding the generalizability of findings to larger scales where attention dynamics may differ, we will also conduct verification experiments on some 70B parameter's model using multi-GPU tensor parallelism. These specific experiments will focus on validating the stability bounds and error propagation theories in high-parameter regimes, ensuring the proposed framework remains effective for larger LLMs.

To ensure the proposed method generalizes across diverse real-world complexities, we will employ a comprehensive benchmark suite covering multi-hop reasoning, noise robustness, and domain-specific challenges. Specifically, we will utilize FRAMES (Krishna et al., 2025) to evaluate standard multi-hop reasoning accuracy, while incorporating RAGBench (Friel et al., 2025) to assess the system's resilience to noisy contexts and the stability bounds' ability to suppress error propagation. Furthermore, we will include BillSum (Kornilova and Eidelman, 2019) to verify the framework's adaptability to specialized terminology and extremely long documents with complex internal references.

For each instance, we will instantiate the same RAG pipeline in different modes. We will not only compare against the standard full-recomputation baseline but also against state-of-the-art engineering heuristics, specifically CacheBlend, which serves as a strong baseline for selective recomputation. By comparing our theoretically guided reuse against CacheBlend's heuristic approach, we aim to demonstrate that our method provides a superior Pareto trade-off between accuracy retention and computational speedup.

5.2 Evaluation Metrics

To comprehensively evaluate the effectiveness of proposed KV cache reuse method, we plan to establish a multi-dimensional evaluation index system. This system will not only focus on the overall performance of the system, but also deeply analyze the specific effects of each component of the algorithm to ensure the comprehensiveness and depth of the evaluation.

For system performance, we will focus on measuring three core metrics, including TTFT, system throughput, and inter-token time. TTFT reflects user-perceived response latency. We will detail the time from request submission to the generation of the first token, focusing on analyzing the speedup achieved by KV cache reuse compared

to a complete recalculation and the CacheBlend baseline. System throughput reflects the system's overall processing power. We will test the maximum request rate the system can handle under various load conditions, which is crucial for evaluating the feasibility of this approach in real-world deployments. The inter-token time metric is primarily used to assess the smoothness of the generation phase. Additionally, to explicitly monitor the cost of the fallback mechanism, we will introduce the Recomputation Trigger Rate (RTR)-the percentage of layers/tokens requiring recomputation-and the Effective Speedup. Regarding the trade-off between cache I/O and computation, we expect to replicate the efficiency patterns observed in RAGCache (up to $4\times$ TTFT reduction and $2.1\times$ throughput improvement). We will measure the "Effective Speedup" by explicitly tracking the end-to-end latency including the cache retrieval time managed by the RAGCache-integrated backend.

For cache efficiency evaluation, we will conduct a more in-depth analysis. In addition to traditional cache hit rate statistics, we will also incorporate cache quality assessment, which includes analyzing the contribution of cached documents to the accuracy of the final answer, as well as the reuse patterns of cached content across requests. By systematically adjusting cache capacity, we plan to quantitatively study the relationship between cache size and performance improvements, providing a reference for actual system deployment.

Generation quality assessment will be aligned with this focus on comparing KV cache reuse against full recomputation and heuristics. Concretely, on FRAMES and the additional datasets, we will compute the official task metrics (e.g., answer correctness, factuality, and domain-specific scores). We will specifically analyze the "Accuracy Drop" relative to full recomputation for both our method and CacheBlend. We hypothesize that our method, protected by stability bounds, will maintain an accuracy profile significantly closer to the "Gold Standard" (full recomputation) than the heuristic approximations used in CacheBlend, particularly in noise-heavy or domain-specific scenarios.

5.3 Theoretical Validation

The theoretical verification phase will systematically validate the various theoretical components of this KV cache reuse framework, to ensure its mathematical rigor and practical reliability and verify the correctness of the theoretical components through carefully designed experiments and explore their practical performance.

First, for the core issue of position alignment, we will design detailed verification experiments. For each of the three major position encoding schemes (RoPE, ALiBi, and absolute position encoding), we will verify the accuracy of their corresponding position transfer operators. Specifically, for RoPE and ALiBi, two encoding schemes that allow for precise correction, we will verify whether the attention calculations adjusted by the position transfer operator can achieve machine-level consistency. For absolute position encoding, since it involves linear approximation, we will focus on verifying whether the actual error is strictly within the theoretically derived residual bounds. These verifications will be conducted by traversing different position indices and testing them in a variety of typical scenarios to ensure the reliability of the position alignment mechanism under various circumstances.

For attention stability verification, we will use controlled variable experiments to test the stability theory based on the Softmax Jacobian matrix. Experiments will simulate perturbations of varying strengths in the key vector, measure the resulting changes in the attention distribution, and compare the measured values with theoretically derived upper bounds. We will test different types of attention distributions, including both sharp and flat ones, to verify the tightness of the theoretical bounds under different circumstances. These experiments not only validate the theory but also provide guidance for parameter tuning in practical systems.

Deep error propagation analysis will be another key validation step. We will track how the hidden state representations of each layer deviate from the ideal path as the network depth increases during request processing. By visualizing the propagation of inter-layer representation drift, we can intuitively demonstrate the effectiveness of the error control mechanism. We will test the impact of different gating factor configurations on error propagation and identify the optimal parameter setting strategy. These experiments will help us gain a deeper understanding of how error accumulates in deep networks and provide a basis for optimizing error control strategies.

Furthermore, we will design specialized boundary condition tests to challenge the theoretical limits. For example, we will construct challenging scenarios with extremely small margins in the at-

tention distribution and test the stability of document ranking in such scenarios. We will also test the performance of this method in edge cases such as processing extremely long contexts and complex interactions between multiple documents. These stress tests not only verify the robustness of the theory but also help identify the limitations of current methods and point the way for future improvements.

6 Research Roadmap

6.1 Theoretical Foundation

The theoretical foundation phase will focus on establishing the theoretical bedrock of this framework and solving the problem of positional misalignment. The expected outputs of this stage are a set of proved theorems (for positional operators and stability inequalities) and a minimal reference implementation for "ideal vs. reused" attention. This stage will be considered sufficient to proceed once the position operators match ideal attention up to a small, predefined tolerance on test cases. Otherwise, the theoretical formulation will be revised before moving on.

6.2 System Integration

The system integration phase is dedicated to addressing the critical challenge of error propagation across transformer layers and integrating all components into a cohesive system. The main outputs of this stage are an end-to-end KV cache reuse prototype and empirical estimates of layer-wise contraction factors under different γ and Top-R settings. The stage will be judged successful when empirical drift curves stay within the error budgets derived in the previous section. If they systematically exceed these budgets, the integration and control scheme will be revised before proceeding.

6.3 Validation and Dissemination

The validation and dissemination phase will be devoted to systematic validation, refinement of the theory, and preparation of the dissertation. The outputs of this stage will be the final theoretical results, a validated implementation, and a comprehensive experimental report suitable for inclusion in the dissertation. This phase will be considered complete once KV cache reuse consistently achieves the targeted efficiency gains while keeping accuracy and measured deviations within the predefined safety margins; otherwise, additional refinement loops between theory and experiments will be performed.

7 Conclusion

The research aims to systematically address the theoretical foundation of multi-document KV cache reuse in RAG systems. We identify the limitations of existing engineering optimization methods (such as RAGCache and CacheBlend) in terms of lack of theoretical guarantees, especially the three core challenges of position misalignment, context misalignment, and error propagation.

The core innovation of the research is to redefine KV cache reuse as a controlled structural perturbation problem and plan to build a complete theoretical framework. Specifically, this research will focus on: developing a universal position transfer operator applicable to different position encoding schemes to achieve accurate or provably approximate position alignment; establishing a data-dependent stability theory based on the Softmax Jacobian matrix to provide quantifiable bounds for changes in attention distribution; designing a deep error propagation control strategy that combines gating mechanisms with selective recomputation to ensure the controllability of inter-layer errors.

Compared with existing work, the unique value of this research lies in providing a solid mathematical foundation for KV cache reuse, rather than proposing another engineering heuristic method. Through systematic theoretical analysis and empirical validation, we plan to advance this research direction from engineering practice to theoretical foundations. This will provide theoretical support for achieving both efficient and reliable RAG systems, particularly in applications requiring extremely high accuracy. We hope to ultimately contribute a theoretically guaranteed KV cache reuse framework to the field of LLMs inference optimization, providing a solid foundation and clear development direction for subsequent researchers.

Limitations

This research also has some significant limitations. Firstly, although the framework is designed to be model-agnostic, its theoretical analysis and empirical verification are limited to casual decoder-only models and only employ mainstream positional coding schemes (e.g. RoPE, ALiBi, and absolute coding). Positional transfer operators and stability bounds are explicitly constructed and calibrated

for these coding schemes. For models employing non-standard or hybrid positional mechanisms, additional derivation and verification are required to obtain the same guarantees.

Secondly, the technical scope of this work is confined to the Generalized Multi-Query Attention (GQA)(Ainslie et al., 2023) mechanism. Our proposed cache reuse framework and theoretical analysis are designed and validated specifically within the computational and memory access patterns of GQA , which serves as a prevalent and representative efficient attention architecture in contemporary LLMs. Consequently, our findings may not directly generalize to models employing other emerging attention computation paradigms, such as the Multi-head Latent Attention (MLA) used in models like DeepSeek-V2 (DeepSeek-AI, 2024). The interplay between KV cache reuse and these alternative attention designs remains an open question for future research.

Acknowledgments

This research was conducted within the framework of the HSE University Basic Research Program. In addition to this institutional framework, this research was also supported in part through the computational resources of the HPC facilities at HSE University. Furthermore, this research received financial support from a grant provided by Huawei Technologies Co., Ltd.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint*, arXiv:2405.04434.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2025. Lightmem: Lightweight and efficient memory-augmented generation. *arXiv preprint*, arXiv:2510.18866.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint*, arXiv:2407.11005.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint*, arXiv:2312.10997.

Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Shufan Liu, Xuanzhe Liu, and Xin Jin. 2025. Ragcache: Efficient knowledge caching for retrieval-augmented generation. *ACM Transactions on Computer Systems*, 44(1):1–27.

Anastassia Kornilova and Vlad Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint*, arXiv:2409.12941.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. *arXiv preprint*, arXiv:2309.06180.

Chien-Yu Lin, Keisuke Kamahori, Yiyu Liu, Xiaoxiang Shi, Madhav Kashyap, Yile Gu, Rulin Shao, Zihao Ye, Kan Zhu, Stephanie Wang, Arvind Krishnamurthy, Rohan Kadekodi, Luis Ceze, and Baris Kasikci. 2025. TeleRAG: Efficient retrieval-augmented generation inference with lookahead retrieval. *arXiv preprint*, arXiv:2502.20969.

Yuhan Liu, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, and Junchen Jiang. 2024. CacheGen: KV cache compression and streaming for fast large language model serving. *arXiv preprint*, arXiv:2310.07240.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint*, arXiv:2108.12409.

Xianbiao Qi, Jianan Wang, and Lei Zhang. 2023. Understanding optimization of deep learning via jacobian matrix and lipschitz constant. *arXiv preprint*, arXiv:2306.09338.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint*, arXiv:2104.09864.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *arXiv preprint*, arXiv:1706.03762.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *arXiv preprint*, arXiv:2309.17453.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. *arXiv preprint*, arXiv:2002.04745.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. 2025. CacheBlend: Fast large language model serving for RAG with cached knowledge fusion. *arXiv preprint*, arXiv:2405.16444.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 1 others. 2025. A survey of large language models. *arXiv preprint*, arXiv:2303.18223.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Jeff Huang, Chuyue Sun, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2023. Efficiently programming large language models using SGLang. *arXiv preprint*, arXiv:2312.07104.