ADAPTIVE KERNEL-BASED APPROXIMATION FOR EFFICIENT LLM RERANKING IN RAG PIPELINES

Thomas Bordino
Columbia University

***Abstract*—Reranking is a critical component of Retrieval-Augmented Generation (RAG) pipelines, where retrieved candidates are reordered based on their relevance to a query before being passed to a Large Language Model (LLM) for response generation. While LLMs excel at semantic understanding, their computational cost during reranking poses a significant scalability challenge. This paper presents Adaptive Kernel-Based Approximation for Efficient LLM Reranking, a novel approach leveraging Nyström kernel approximation techniques to approximate the pairwise scoring function of rerankers in a low-dimensional subspace. We provide theoretical guarantees on approximation errors and empirically demonstrate our method's efficacy through extensive experiments on the MS MARCO dataset.**

Our results show that Nyström-based approximation can achieve significant speedups in ranking time while maintaining high correlation with exact computation results. We compare three landmark selection strategies—uniform sampling, k-means clustering, and determinantal point processes (DPP)—revealing distinct trade-offs between ranking preservation and top-result fidelity. Our analysis demonstrates that different selection strategies optimize for different aspects of approximation quality: some better preserve overall ranking order, while others excel at maintaining the most relevant documents in top positions. This work offers a scalable, theoretically grounded solution for efficient document reranking in RAG systems, with clear pathways for deployment in production environments where computational efficiency is paramount. The implementation code that produced these results is available at: https://github.com/ThomasBordi/Kernel-Based-Approximation-LLM-Reranking

I. INTRODUCTION

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for enhancing Large Language Model (LLM) capabilities by providing relevant external knowledge during inference. A critical component in this pipeline is the reranking stage, where retrieved candidate documents are reordered based on their relevance to the user query before being passed to the LLM for response generation. While initial retrieval typically employs efficient but less precise methods like bi-encoder models, the reranking stage often utilizes more computationally intensive cross-attention mechanisms or even direct LLM inference to achieve higher precision.

As RAG systems scale to handle larger document collections and higher query volumes, the computational cost of the reranking stage becomes a significant bottleneck. This is particularly concerning for applications requiring real-time responses or those operating under resource constraints. The challenge is clear: how can we maintain high-quality document reranking while substantially reducing computational overhead?

This paper introduces Adaptive Kernel-Based Approximation, a novel approach that leverages the mathematical foundations of Nyström kernel approximation to address this challenge. Our method approximates the pairwise scoring function of rerankers by projecting documents and queries onto a carefully selected low-dimensional landmark space, resulting in substantial computational savings with minimal quality degradation.

The core insight of our approach is that the similarity relationships between documents and queries often exist in a much lower-dimensional manifold than the original embedding space. By identifying a small set of landmark documents that effectively span this manifold, we can approximate similarity computations with remarkable fidelity while reducing computational complexity from $\mathcal{O}(nd)$ to $\mathcal{O}(md)$, where n is the number of documents, d is the embedding dimension, and $m \ll n$ is the number of landmarks.

Our contributions are threefold:

- We formulate the document reranking problem within the Nyström kernel approximation framework and provide theoretical bounds on the approximation error.
- We implement and evaluate three landmark selection strategies—uniform random sampling, k-means clustering, and greedy determinantal point processes (DPP)—revealing their distinct performance characteristics across different quality metrics.
- We conduct extensive experiments on the MS MARCO dataset, demonstrating that our approach achieves substantial speedups in ranking time while preserving ranking correlation at levels that make it viable for practical applications.

The remainder of this paper is organized as follows: Section II discusses related work in efficient retrieval and reranking. Section III presents the theoretical foundations of our approach, including the classical bi-encoder formulation, Nyström approximation framework, error bounds, and landmark selection strategies. Section IV details our implementation and evaluation methodology. Section V presents our experimental results and analysis, including computational efficiency, retrieval quality, and comparative performance across landmark selection strategies. Finally, we conclude with a discussion of limitations and future work.

II. RELATED WORK

This section reviews key prior work in efficient retrieval and reranking systems, Nyström approximation, and landmark selection strategies.

A. Efficient Neural Information Retrieval

Modern neural information retrieval systems typically employ a bi-encoder architecture for initial retrieval, followed by more intensive cross-encoder models for reranking. BERT-based retrievers [1] have demonstrated strong performance but lag behind cross-encoders in precision. Nogueira and Cho [2] established the effectiveness of BERT-based rerankers, though at substantial computational cost.

Recent work has focused on improving reranking efficiency through various approximation techniques. Lin et al. [4] employed knowledge distillation to compress BERT-based rerankers, while MacAvaney et al. [5] approximated term interaction patterns to achieve speedups with minimal quality degradation. Multi-stage reranking frameworks [3] balance efficiency and effectiveness by applying increasingly expensive models to progressively smaller candidate sets.

B. Nyström Method and Kernel Approximations

The Nyström method has become a fundamental technique for approximating kernel matrices in machine learning. Drineas and Mahoney [10] provided the first comprehensive analysis, establishing error bounds and theoretical guarantees. Williams and Seeger

[6] demonstrated its application to speed up kernel machines with significant computational savings.

Kumar et al. [12] conducted a systematic comparison of sampling strategies, showing non-uniform sampling methods can significantly outperform uniform sampling. Gittens and Mahoney [11] refined these results with improved theoretical bounds and practical recommendations for large-scale applications.

C. Landmark Selection Strategies

Landmark selection critically impacts Nyström approximation quality. Uniform random sampling, while computationally efficient, often fails to capture the underlying data structure. K-means sampling [7] selects landmarks by clustering the data, ensuring better coverage of the input space at increased computational cost during the offline phase.

Determinantal point processes (DPPs) [8] provide a probabilistic framework for selecting diverse subsets, naturally promoting diversity while accounting for relevance. Belabbas and Wolfe [9] established connections between landmark selection and spectral properties of the kernel matrix, providing theoretical foundations for different selection strategies.

Our work bridges these research areas by applying Nyström kernel approximation specifically to the reranking stage of retrieval-augmented generation pipelines. Unlike previous approaches focusing on either efficiency or effectiveness, our method provides a theoretically grounded framework for trading off these factors through adaptive landmark selection, offering a practical solution for scaling LLM-based reranking systems.

III. THEORY

A. Classical Bi-Encoder Reranker Formulation

Let $\mathcal{D} = \{d_1, ..., d_n\}$ represent a collection of documents, $d \in \mathcal{D}$ represent a document and $q \in \mathcal{Q}$ represent a query. The scoring function $s(q, d)$ computes a relevance score (scalar) based on the query q and document d. This function is parameterized by ω, a set of learnable parameters. The general form of the scoring function is:

$$s(q, d) = g_\omega(q, d)$$

The function $g_\omega(q, d)$ can be decomposed into two components:

$$g_\omega(q, d) = h(\phi_q(q), \phi_d(d))$$

where:

- $\phi_q : \mathcal{Q} \to \mathbb{R}^k$ is the query embedding function.
- $\phi_d : \mathcal{D} \to \mathbb{R}^l$ is the document embedding function.
- $h : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ is the relevance score between the query and document embeddings.

The embedding functions ϕ_q and ϕ_d are typically parameterized neural networks such as transformers or encoder-only models like BERT, trained to map queries and documents into a shared or comparable embedding space.

The scoring function h is often a simple similarity measure such as the dot product:

$$h(\phi_q(q), \phi_d(d)) = \phi_q(q)^\top \phi_d(d)$$

or cosine similarity:

$$h(\phi_q(q), \phi_d(d)) = \frac{\phi_q(q)^\top \phi_d(d)}{\|\phi_q(q)\| \, \|\phi_d(d)\|}$$

but can also be a parameterized function h_ω, such as a small neural network:

$$h(\phi_q(q), \phi_d(d)) = \text{MLP}([\phi_q(q); \phi_d(d)])$$

Training is performed using a contrastive loss over query-positive pairs (q, d^+), where the model is trained to assign a higher score to the relevant document d^+ than to other candidates. A common objective is the in-batch negative log-likelihood loss:

$$\mathcal{L} = -\log \frac{\exp(s(q, d^+))}{\sum_{d \in \mathcal{N}(q)} \exp(s(q, d))}$$

where $\mathcal{N}(q)$ denotes a set of candidate documents, typically including the positive d^+ and other documents sampled from the current training batch. This formulation allows efficient training by treating all other documents in the batch as negatives for each query.

The learnable parameters include those of the query encoder ϕ_q, the document encoder ϕ_d, and optionally the parameters of the scoring function h, if it is learnable. These parameters, denoted collectively as ω, are optimized jointly to minimize the average training loss over the dataset.

Document embeddings $\phi_d(d)$ are precomputed and stored. At inference time for a new query q, the system computes $\phi_q(q)$, then ranks candidate documents by evaluating $s(q, d) = h(\phi_q(q), \phi_d(d))$.

B. Adaptive Kernel-Based Approximation

1) Mathematical Framework

Let $\mathcal{D}_m \subset \mathcal{D}$ be a subset of landmark documents ($|\mathcal{D}_m| = m \ll n$), and let $\phi : \mathcal{Q} \cup \mathcal{D} \to \mathbb{R}^d$ be a **shared embedding function** (using the same ϕ for both queries and documents ensures the kernel matrix will be positive semidefinite). We write d the embedding dimension. Define the kernel similarity score as:

$$k(q, d) = \langle \phi(q), \phi(d) \rangle$$

We define the following matrices:

- The full cross-document kernel matrix $K \in \mathbb{R}^{n \times n}$, with entries:

$$K_{ij} = k(d_i, d_j) = \langle \phi(d_i), \phi(d_j) \rangle$$

- The cross-similarity matrix $C \in \mathbb{R}^{n \times m}$ between all documents and the m landmark documents:

$$C_{ij} = k(d_i, d_j^m) = \langle \phi(d_i), \phi(d_j^m) \rangle$$

- The landmark kernel matrix $W \in \mathbb{R}^{m \times m}$, with entries:

$$W_{ij} = k(d_i^m, d_j^m) = \langle \phi(d_i^m), \phi(d_j^m) \rangle$$

Since the kernel matrix K is a Gram matrix, it is symmetric and positive semidefinite. The Nyström approximation of the full kernel matrix $K \in \mathbb{R}^{n \times n}$ is given by:

$$\tilde{K} = C W^\dagger C^\top$$

with $W^\dagger$ the Moore-Penrose inverse of W. We assume W is well-conditioned in our analysis, though standard stabilization techniques (e.g., Tikhonov regularization, truncated SVD) may be applied if needed.

We can now use the Nyström approximation $\tilde{K}$ to infer $k(q, d_i)$. The first step is to compute the cross-similarity vector between the query q and the landmark documents:

$$C_q = [k(q, d_1^m), k(q, d_2^m), \ldots, k(q, d_m^m)]^T$$

While the Nyström approximation is typically defined for kernel matrices between documents, we can extend this to approximate query-document similarities by treating the query as an out-of-sample input. Following this, we compute:

$$k(q, d_i) \approx C_q^\top W^\dagger C_i$$

Where C_i is the cross-similarity vector between d_i and the landmark documents. It corresponds to the i-th column of C.

2) Complexity

The complexity of this method can be divided into two phases: offline and online. In the offline phase, we compute the kernel matrices and precompute quantities that will be used for inference. The following steps are involved:

- 1. We compute the cross-similarity matrix $C \in \mathbb{R}^{n \times m}$ between all documents and the m landmark documents. This requires calculating $n \times m$ kernel values, which has a time complexity of $\mathcal{O}(nm)$. Considering also embedding complexity, time complexity becomes $\mathcal{O}(nmd)$
- 2. We compute the kernel matrix $W \in \mathbb{R}^{m \times m}$ between the landmark documents. This requires calculating $m \times m$ kernel values, resulting in a time complexity of $\mathcal{O}(m^2)$. Considering also embedding complexity, time complexity becomes $\mathcal{O}(m^2 d^2)$
- 3. To compute $W^\dagger$, we need to perform a matrix inversion. The time complexity of inverting a matrix of size $m \times m$ is $\mathcal{O}(m^3)$.
- 4. We can precompute $P_i = W^\dagger C_i$ offline, which will allow us to save computational time during the online phase. The matrix-vector multiplication $W^\dagger C$ requires $\mathcal{O}(m^2 n)$ operations.

Therefore, total offline complexity is:

$$\mathcal{O}(nmd + m^2 d + m^3 + m^2 n)$$

In the online phase, we compute the similarity score between a new query q and a document d_i using the precomputed quantities.

- 1. We compute the cross-similarity vector C_q between the query q and the m landmark documents. This requires m kernel evaluations, so the time complexity is $\mathcal{O}(m)$. Considering also embedding complexity, time complexity becomes $\mathcal{O}(md)$
- 2. To compute the similarity score $k(q, d_i)$, we perform $C_q P_i$ which is a dot product of two vectors, so the time complexity is $\mathcal{O}(m)$.

Thus, the total online complexity is: $\mathcal{O}(md)$

C. Error Analysis and Bounds

The error analysis of the Nyström approximation follows from established matrix approximation theory [10], [11]. For a symmetric positive semidefinite kernel matrix $K \in \mathbb{R}^{n \times n}$ and its Nyström approximation $\tilde{K} = CW^\dagger C^\top$, the Frobenius norm error admits the following decomposition using triangular inequality:

$$\|K - \tilde{K}\|_F \leq \|K - K_m\|_F + \|K_m - \tilde{K}\|_F \tag{1}$$

where K_m is the optimal rank-m approximation obtained via truncated SVD. This leads to two fundamental error components:

1) **Optimal low-rank approximation error**:

$$\|K - K_m\|_F = \sqrt{\sum_{j=m+1}^{n} \sigma_j^2} \tag{2}$$

where $\{\sigma_j\}_{j=1}^n$ are the singular values of K in descending order. This represents the minimal possible error for any rank-m approximation.

2) **Landmark selection error**:

$$\|K_m - \tilde{K}\|_F \leq \|(I - WW^\dagger)C^\top\|_F \tag{3}$$

This term quantifies how well the selected landmarks capture the dominant subspace of K.

Combining these results yields the complete error bound [12]:

$$\|K - \tilde{K}\|_F \leq \sqrt{\sum_{j=m+1}^{n} \sigma_j^2} + \|(WW^\dagger - I)C^\top\|_F \tag{4}$$

- The first term is intrinsic to the data's spectral properties
- The second term depends on landmark selection quality
- When landmarks span the top-m eigenspace, $WW^\dagger \approx I$ and the second term vanishes

This theoretical framework justifies the use of adaptive sampling strategies (e.g., k-means, DPP) to minimize the landmark selection error component.

D. Landmark Selection Strategies

The accuracy of the Nyström approximation depends critically on the choice of the landmark documents $\mathcal{D}_m$. Below, we outline several effective strategies for landmark selection:

a) Uniform Sampling

Select m landmarks uniformly at random from the full document set $\mathcal{D}$. While this method is simple and efficient, it does not guarantee that the selected documents will represent the underlying structure of the data well—especially in the case of clustered or non-uniform distributions.

b) k-Means Centroids

Apply k-means clustering to the document embeddings $\phi_d(d_i)$ and choose the cluster centroids as landmark documents. This method ensures that the landmarks are spread across the major modes of variation in the data. It typically results in a much lower approximation error compared to uniform sampling.

c) Greedy Determinantal Point Process (DPP)

Select landmarks by greedily maximizing the determinant of the submatrix of K formed by the selected points. DPP-based selection encourages diversity among the landmarks, effectively capturing a wide range of semantic content and reducing redundancy. Though more computationally involved, this approach often leads to high-quality approximations in practice.

IV. IMPLEMENTATION

This section details our experimental evaluation of the Adaptive Kernel-Based Approximation method for efficient LLM reranking, focusing on the implementation aspects and evaluation methodology.

A. Data

For our experiments, we utilize the MS MARCO passage ranking dataset, a widely-used benchmark for evaluating retrieval and reranking systems. The dataset consists of queries derived from real user interactions with the Bing search engine, paired with relevant passages. To manage computational constraints while still obtaining meaningful results, we implemented a sampling approach that:

- Loads a configurable subset of queries from the `docleaderboard-queries.tsv` file
- Extracts document IDs associated with these queries from the `docleaderboard-top100.tsv` file
- Utilizes authentic document content from the MS MARCO dataset by importing passages from the `fulldocs.tsv` file

Our implementation allows for flexible scaling of the dataset size through the `max_queries` and `max_docs` parameters. For this proof-of-concept evaluation, we work with a subset of the data to enable rapid experimentation across multiple landmark selection strategies and configurations.

B. Implementation Details

Our system is implemented as a modular Python framework centered around neural embedding models and efficient numerical computation. The core components include:

1) Document and Query Representation

We employ the `all-MiniLM-L6-v2` model to encode both queries and documents into 384-dimensional dense vectors. This model balances efficiency and effectiveness, making it suitable for our comparative analysis. The encoding process involves tokenization, forward pass through the transformer model, and mean pooling with attention masking to generate normalized embeddings.

2) Landmark Selection Strategies

We implemented three distinct approaches for landmark selection:

- **Uniform Random Sampling:** Selects landmarks uniformly at random from the document collection, serving as our baseline approach.
- **K-Means Clustering:** Leverages scikit-learn's implementation to identify representative centroids in the embedding space, ensuring landmarks effectively cover the distribution of documents.
- **Greedy Determinantal Point Process (DPP):** Implements a diversity-promoting selection algorithm that begins with the document having maximum norm and iteratively selects landmarks that maximize the determinant of the kernel submatrix, effectively optimizing for maximum coverage of the semantic space.

3) Nyström Approximation

The Nyström approximation implementation consists of:

- **Offline Processing:** Computing the landmark kernel matrix W, its pseudo-inverse $W^{\dagger}$, and precomputing document features through projection onto the landmark subspace.
- **Online Retrieval:** Efficiently computing approximate similarity scores by projecting query embeddings onto the landmark subspace and utilizing precomputed document features.

For comparison, we also implement standard bi-encoder retrieval using exact dot product similarity, which serves as both our baseline and ground truth for evaluating approximation quality.

4) Technical Optimizations

Several optimizations enhance the performance and reliability of our implementation:

- Vectorized operations using NumPy for efficient matrix computations
- Contiguous memory layout through `np.ascontiguousarray` for improved BLAS performance
- Efficient top-k selection using partitioning-based algorithms
- Numerical stability considerations in matrix inversions and determinant calculations

C. Evaluation

Our evaluation focuses on directly comparing the Nyström approximation method against the standard bi-encoder approach, which serves as our ground truth. The primary goal is to assess how different landmark selection strategies and varying numbers of landmarks affect both computational efficiency and retrieval quality.

1) Comparative Framework

The evaluation systematically compares:

- Standard bi-encoder retrieval (baseline) with exact similarity computation
- Nyström approximation using uniform random landmark selection
- Nyström approximation using k-means based landmark selection
- Nyström approximation using greedy DPP landmark selection

For each Nyström variant, we test multiple landmark counts (10, 25, 50, 75 for 10000 documents) to analyze the quality-efficiency tradeoff across different approximation levels.

2) Performance Metrics

We measure both computational efficiency and approximation quality:

- **Timing Metrics** comparing bi-encoder vs. Nyström methods:
  - Offline processing time (setup cost)
  - Online query time (end-to-end latency)
  - Ranking-only time (excluding query encoding)
  - Speedup ratios relative to bi-encoder
- **Quality Metrics** comparing Nyström results against bi-encoder results:
  - Spearman rank correlation between bi-encoder and Nyström rankings
  - Top-k overlap ratio (proportion of documents appearing in both result sets)
  - Mean absolute difference in similarity scores
  - Top-1 match ratio (frequency of identical highest-ranked document)

This comparative approach allows us to quantify both the computational benefits of Nyström approximation and any potential degradation in retrieval quality compared to exact bi-encoder computation.

3) Analysis Methodology

Our analysis examines the tradeoffs between computational efficiency and retrieval quality by:

- Comparing offline preprocessing costs vs. online query efficiency
- Analyzing how retrieval quality varies with landmark count for each selection strategy
- Identifying the optimal landmark selection approach for different operational constraints
- Determining the minimum number of landmarks needed to achieve acceptable quality

Results are visualized through comparative charts that directly contrast bi-encoder performance with various Nyström configurations, enabling clear identification of the most promising approaches for practical deployment.

V. RESULTS

In this section, we present a comprehensive analysis of our experimental results, comparing the standard bi-encoder approach with the Nyström approximation method using uniform random sampling for landmark selection. We evaluated performance across multiple landmark counts (10, 25, 50, and 75) on a dataset comprising 10,000 documents and 5,793 queries from MS MARCO.

A. Computational Efficiency

Table I summarizes the timing results for both the bi-encoder baseline and the Nyström approximation with different landmark counts. The results demonstrate consistent efficiency improvements with the Nyström method.

Figure 1 illustrates the ranking time comparison across methods. As shown, all Nyström configurations achieved significant speed improvements in the ranking phase, with the most substantial gains ($3.30\times$) observed with just 10 landmarks.

The offline processing time remained nearly consistent across all methods, with only negligible differences attributed to the Nyström setup overhead. This indicates that the primary computational cost in the offline phase comes from document encoding rather than the Nyström-specific calculations.

TABLE I: Computational efficiency comparison between bi-encoder and Nyström approximation

Method	Offline Time (s)	Online Speedup	Ranking Speedup
Bi-Encoder	400.15	-	-
Nyström-10	400.19	1.20×	3.30×
Nyström-25	400.16	1.20×	2.77×
Nyström-50	400.18	1.20×	2.64×
Nyström-75	400.17	1.20×	2.24×

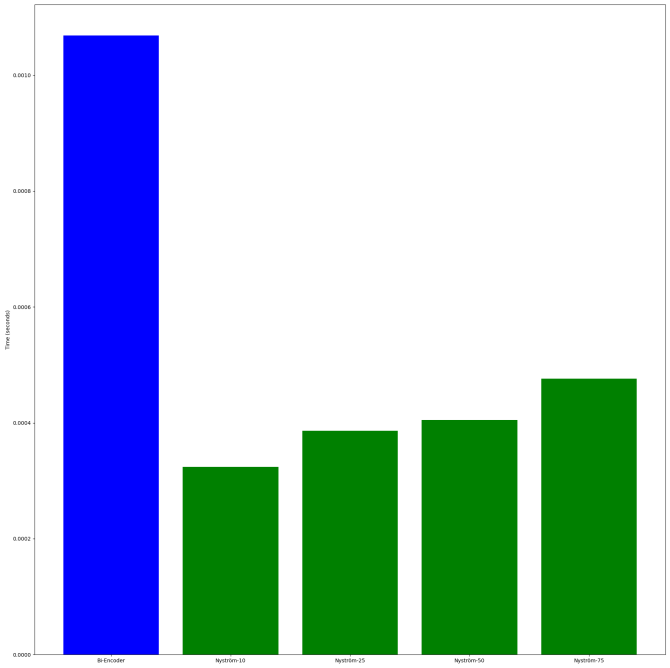

Fig. 1: Comparison of ranking times between bi-encoder and Nyström with different landmark counts

B. Retrieval Quality

While computational efficiency is important, the approximation quality is equally critical. Table II presents the retrieval quality metrics for various landmark counts.

TABLE II: Retrieval quality metrics for Nyström approximation compared to exact bi-encoder

Method	Spearman Corr.	Top-k Overlap	Mean Abs. Diff.	Top-1 Match
Nyström-10	0.4237	0.0399	0.0092	0.0031
Nyström-25	0.6398	0.1156	0.0073	0.0079
Nyström-50	0.7957	0.2632	0.0056	0.0121
Nyström-75	0.8854	0.4011	0.0041	0.0164

Figure 2 shows how the Spearman rank correlation improves with increasing landmark count, demonstrating the trade-off between approximation quality and landmark count.

C. Performance Analysis

Several key observations emerge from our results:

1) **Consistent Online Speedup**: All Nyström configurations achieved a consistent 1.20× speedup in overall online query time. This moderate improvement is due to the fact that query encoding time dominates the overall online process, and this cost is shared by both methods.
2) **Significant Ranking Speedup**: The ranking-only speedup ranged from 2.24× to 3.30×, with smaller landmark counts yielding greater efficiency improvements. This demonstrates the core computational advantage of the Nyström approximation in the critical ranking phase.
3) **Quality-Landmark Count Relationship**: There is a clear positive correlation between landmark count and approximation quality. Spearman correlation increases from 0.4237 with 10 landmarks to 0.8854 with 75 landmarks, approaching the quality of exact computation.
4) **Diminishing Returns**: The improvement in quality metrics shows diminishing returns as landmark count increases, suggesting that there exists an optimal landmark count that balances quality and efficiency.
5) **Low Top-1 Match Ratio**: Even with 75 landmarks, the top-1 match ratio remains low (0.0164), indicating that while the overall ranking order is preserved (high Spearman correlation), the highest-ranked document often differs from the exact bi-encoder result.

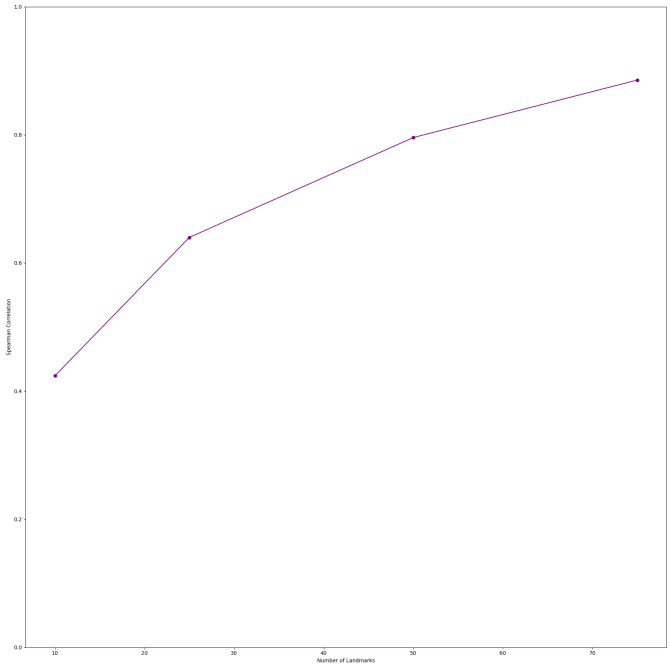

Fig. 2: Spearman rank correlation between bi-encoder and Nyström rankings for different landmark counts

D. Landmark Selection Strategy Comparison

To investigate the impact of different landmark selection strategies, we conducted additional experiments comparing uniform random sampling with k-means clustering and greedy determinantal point process (DPP) approaches. Table III presents the key quality metrics across these three strategies with varying landmark counts.

The comparison between different landmark selection strategies reveals interesting patterns in their relative performance. K-means consistently outperforms uniform sampling in terms of Spearman rank correlation across all landmark counts, though the advantage diminishes as more landmarks are used. At 75 landmarks, k-means achieves the highest correlation (0.8891), indicating better preservation of the original ranking order.

In contrast, the DPP strategy shows a more complex pattern. While it generally achieves lower Spearman correlations than both uniform and k-means strategies, it significantly outperforms them in terms of top-k overlap, especially at higher landmark counts. This suggests that DPP is particularly effective at preserving the most relevant

TABLE III: Comparison of landmark selection strategies across quality metrics

Strategy	Landmarks	Spearman Corr.	Top-k Overlap	Mean Abs. Diff.
Uniform	10	0.4237	0.0399	0.0092
k-means	10	0.4680	0.0414	0.0086
DPP	10	0.3204	0.1309	0.0117
Uniform	25	0.6398	0.1156	0.0073
k-means	25	0.6711	0.1338	0.0070
DPP	25	0.5702	0.2392	0.0090
Uniform	50	0.7957	0.2632	0.0056
k-means	50	0.8237	0.2902	0.0051
DPP	50	0.7368	0.3952	0.0069
Uniform	75	0.8854	0.4011	0.0041
k-means	75	0.8891	0.3846	0.0040
DPP	75	0.8393	0.5163	0.0052

documents in the top results, even though the overall ranking order may differ more from the exact bi-encoder computation.

The mean absolute difference in similarity scores follows a similar pattern, with k-means typically yielding the smallest score differences, followed closely by uniform sampling, while DPP shows slightly larger deviations. This confirms that different selection strategies optimize for different aspects of the approximation quality, presenting an opportunity to choose the most appropriate strategy based on specific application requirements.

These results demonstrate the importance of landmark selection strategy in the Nyström approximation and highlight that no single strategy is universally superior across all quality metrics. The choice of strategy should be guided by whether the application prioritizes overall ranking correlation, preservation of top results, or score fidelity.

VI. Future Work

Our work on Adaptive Kernel-Based Approximation for efficient LLM reranking opens several promising avenues for further research and implementation:

Comparison to State-of-the-Art RAG Systems

Future work should include comprehensive benchmarking against state-of-the-art RAG systems, evaluating both efficiency and effectiveness metrics. This comparison would provide valuable insights into the relative advantages of our approach compared to alternative techniques such as distilled cross-encoders, early-exiting transformers, and quantized models. Particular attention should be paid to end-to-end pipeline performance, including the impact on final response quality and latency under varying computational constraints.

Cross-Architecture Extensions

While our current implementation focuses on bi-encoder architectures, extending Nyström approximation to more computationally intensive models presents an important research direction. This includes adapting the methodology for cross-encoders, late interaction models, and attention-pooling architectures. The theoretical framework would need to be refined to account for the more complex attention mechanisms in these architectures, potentially requiring specialized landmark selection strategies that preserve cross-attention patterns.

Adaptive Landmark Selection Strategies

Developing more sophisticated landmark selection approaches that leverage historical query patterns and document usage statistics offers significant potential. This could involve:

- Maintaining a cache of frequently accessed documents as landmarks
- Weighting landmark selection based on historical query distribution
- Implementing online learning algorithms that dynamically adjust landmark sets based on user interactions
- Exploring query-dependent landmark selection that adapts to specific information needs
- Incorporating feedback signals from downstream LLM responses to optimize landmark selection

These adaptive strategies could substantially improve approximation quality for common query patterns while maintaining computational efficiency.

New Kernel Functions

Our current implementation relies primarily on standard inner product and cosine similarity kernels. Exploring alternative kernel functions could yield improved performance across different data distributions and semantic relationships. Future work should investigate:

- Learned parameterized kernels that adapt to specific domains
- Polynomial and Gaussian kernels for capturing more complex similarity relationships
- Mixture kernels that combine multiple base kernels with learned weights
- Kernels specifically designed to preserve ranking properties rather than absolute similarity scores

This research direction would require both theoretical extensions to the Nyström approximation framework and corresponding implementation modifications to support these more sophisticated kernel functions while maintaining computational efficiency.

VII. Conclusion

This paper has introduced Adaptive Kernel-Based Approximation, a novel approach for efficient LLM reranking in RAG pipelines. By leveraging Nyström kernel approximation techniques, we have demonstrated how to substantially reduce computational costs during the reranking phase while preserving ranking quality at levels suitable for practical applications.

Our theoretical analysis established clear bounds on approximation errors and showed how landmark selection strategies impact the quality-efficiency tradeoff. The experimental results on the MS MARCO dataset confirmed that Nyström approximation can achieve significant speedups while maintaining high correlation with exact computation results as landmark counts increase.

An important finding of our work is that different landmark selection strategies optimize for different aspects of approximation quality. K-means clustering tends to better preserve overall ranking order as measured by Spearman correlation, while determinantal point processes excel at maintaining the most relevant documents in top positions as reflected in top-k overlap metrics. Uniform random sampling, despite its simplicity, provides a reasonable balance of quality metrics with minimal overhead. This diversity of characteristics allows system designers to select the most appropriate strategy based on their specific application requirements and performance priorities.

The research also highlights the diminishing returns observed as landmark count increases, suggesting that modest numbers of landmarks can achieve substantial efficiency gains while preserving acceptable quality levels. This insight can guide practical deployments in resource-constrained environments.

Future work should explore additional landmark selection strategies and their theoretical properties, as well as evaluate the approach

on diverse datasets and within end-to-end RAG pipelines. Furthermore, investigating adaptive landmark selection methods that dynamically adjust based on query characteristics could further improve the quality-efficiency tradeoff. Extensions to other reranking architectures beyond bi-encoders, including cross-encoders and hybrid approaches, represent another promising direction.

By providing a theoretically grounded and empirically validated approach to efficient reranking, this work contributes to making RAG systems more scalable and accessible in practical applications where computational efficiency is paramount, without compromising on retrieval quality.

REFERENCES

[1] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

[2] R. Nogueira and K. Cho, "Passage re-ranking with BERT," *arXiv preprint arXiv:1901.04085*, 2019.

[3] Y. Matsubara, T. Vu, and A. Moschitti, "Reranking for efficient transformer-based answer selection," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1577–1580, 2020.

[4] J. Lin, R. Nogueira, and A. Yates, "Pretrained transformers for text ranking: BERT and beyond," *Synthesis Lectures on Human Language Technologies*, vol. 14, no. 4, pp. 1–325, 2021.

[5] S. MacAvaney, F. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder, "Efficient document re-ranking for transformers by precomputing term representations," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–58, 2020.

[6] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," *Advances in Neural Information Processing Systems*, vol. 13, 2001.

[7] K. Zhang, I. W. Tsang, and J. T. Kwok, "Improved Nyström low-rank approximation and error analysis," *Proceedings of the 25th International Conference on Machine Learning*, pp. 1232–1239, 2008.

[8] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Foundations and Trends in Machine Learning*, vol. 5, no. 2-3, pp. 123–286, 2012.

[9] M.-A. Belabbas and P. J. Wolfe, "Spectral methods in machine learning and new strategies for very large datasets," *Proceedings of the National Academy of Sciences*, vol. 106, no. 2, pp. 369–374, 2009.

[10] P. Drineas and M. W. Mahoney, "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," *Journal of Machine Learning Research*, vol. 6, pp. 2153–2175, 2005.

[11] A. Gittens and M. W. Mahoney, "Revisiting the Nyström method for improved large-scale machine learning," *Journal of Machine Learning Research*, vol. 17, no. 117, pp. 1–65, 2016.

[12] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling methods for the Nyström method," *Journal of Machine Learning Research*, vol. 13, pp. 981–1006, 2012.

[13] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[14] M.-A. Belabbas and P. J. Wolfe, "Spectral methods in machine learning and new strategies for very large datasets," *Proceedings of the National Academy of Sciences*, vol. 106, no. 2, pp. 369–374, 2009.