Hybrid RankingSHAP : Bridging Local and Global Explainability

Magnan Florian
Politecnico di Torino
Turin, Italy

Ouzzane Réda
Politecnico di Torino
Turin, Italy

ABSTRACT

The increasing deployment of machine learning models in high-impact domains has raised the need for transparent and interpretable decision-making, particularly in ranking systems. In this work, we propose a methodology to derive faithful global feature importance profiles for ranking models by aggregating local listwise explanations generated using RankingSHAP. Our approach applies a pre-trained listwise explainer to a LambdaMART model trained on the MQ2008 dataset, computes local feature attributions for a representative set of queries, and aggregates them using the mean to produce a coherent global profile. We validate the global profile through retraining experiments, comparing models using top globally selected features to those using random features or the full feature set. The results show that our approach provides a compact and interpretable summary of ranking model behavior, supporting reliable system-level auditing and analysis. These findings offer actionable insights for practitioners seeking to build more transparent and accountable ranking systems.

1 INTRODUCTION

In recent years, the increasing adoption of machine learning models in high-impact domains has raised essential questions regarding transparency, accountability, and trust. In particular, *Explainable Artificial Intelligence* (XAI) has emerged as a critical research area aiming to provide human-understandable insights into complex model decisions. Among the various paradigms of XAI, **feature attribution** stands out as a widely used approach, offering quantifiable measures of how input features contribute to a model's predictions. This is especially relevant in the context of ranking algorithms, where stakeholders often need to understand not only why a single item is ranked as it is, but also what consistently drives the ranking decisions across multiple queries and documents.

Case study: Feature attribution in a web search scenario

Our study focuses on the MQ2008 dataset, part of the LETOR 4.0 benchmark collection, which represents a standard resource for evaluating learning-to-rank algorithms. The dataset consists of over 800 queries, each associated with a set of candidate documents retrieved from a web search engine. Each query-document pair is characterized by 46 precomputed numerical features capturing diverse signals, such as textual relevance scores (e.g., BM25, language models), link-based metrics (e.g., PageRank), and URL structure characteristics. Labels indicate relevance on a three-point scale (0 = not relevant, 1 = relevant, 2 = highly relevant). This setting provides a realistic web search scenario for evaluating feature attribution methods [6].

Feature attribution

In this work, we focus on SHAP (SHapley Additive exPlanations), a feature attribution method rooted in cooperative game theory, which attributes importance to features by computing their average marginal contribution across all possible feature coalitions [7]. SHAP offers a theoretically sound and model-agnostic framework, making it particularly appealing for complex tasks such as ranking. When applied to ranking, feature attribution techniques can operate at different granularities: pointwise, pairwise, and listwise. But is it sufficient to understand why a single document is scored as it is (pointwise)? Or to know why document A is ranked higher than document B (pairwise)? Ultimately, should we not strive to explain the ordering of the entire list (listwise), as this is the output that truly impacts user interactions and decisions?

Ranking and the need for explainability

Ranking algorithms, such as those based on gradient-boosted decision trees (e.g., LambdaMART), play a central role in modern information retrieval systems [1]. However, their complexity and reliance on multiple interacting features often result in opaque decision processes. As ranking outcomes increasingly affect critical areas—from search engines to recommendation systems and hiring platforms—the need for explainability has become paramount. Local explainability methods, such as ShaRP, which estimates the contribution of features to ranking outputs, and RankingSHAP, which adapts Shapley values to explain entire ranked lists, have provided valuable tools to interpret individual predictions or specific query-document lists [4, 8], but they leave open the question of how to generalize these insights globally.

Research gap and our contribution

While local feature attribution methods for ranking have advanced our ability to interpret specific predictions, they fall short of addressing a fundamental question: *What consistently drives ranking decisions across the entire dataset*? Can we truly trust a model without understanding these global patterns? How can system-level stakeholders audit or debug a ranking system if explanations are fragmented across individual queries? Current approaches provide fine-grained insights but lack a systematic framework for aggregating local attributions into coherent, faithful global explanations. This gap is particularly critical in sensitive applications, where transparency, fairness, and accountability are not optional but mandatory.

To address this challenge, our work proposes a hybrid pipeline that aggregates listwise feature attributions from RankingSHAP across a representative set of queries to construct a global feature importance profile. Our goal is to bridge the gap between local interpretability and global model understanding, providing a comprehensive view of what truly drives ranking decisions at scale. We seek to demonstrate that such a global profile can generalize well and support trustworthy deployment of ranking models in real-world scenarios.

2 RELATED WORK

Explainability for ranking algorithms has received increasing attention as these models are deployed in high-impact domains where transparency and accountability are critical. Early efforts have focused on post-hoc feature attribution, providing insight into which input features drive model decisions.

Local feature attribution

A foundational concept for feature attribution is the Shapley value, drawn from cooperative game theory, which assigns credit to each feature based on its average marginal contribution across all possible coalitions [10]. This principle ensures that feature attributions satisfy properties such as fairness and symmetry, making it an appealing foundation for explainability. SHAP [7] builds on this theory to offer an additive, model-agnostic framework that quantifies feature contributions in a consistent and theoretically sound manner. SHAP has become widely used for its ability to produce explanations that are both faithful to the model and interpretable for human stakeholders. However, SHAP was originally designed for individual predictions in regression or classification, and its direct application to ranking tasks does not account for the listwise structure inherent to these problems.

Local feature attribution for ranking

To address this limitation, specialized methods have emerged that adapt local attribution principles to ranking contexts. ShaRP [8] proposes a framework that estimates feature importance by analyzing how feature perturbations influence the relative ordering of documents for a given query. It seeks to provide insight into which features most strongly affect the ranking outcome at the level of individual lists. RankingSHAP [4], in contrast, extends the SHAP framework by redefining the value function to operate on entire ranked lists rather than single predictions. It computes Shapley values that reflect each feature's contribution to the overall list order, capturing dependencies between document positions in a principled way. While these methods represent important progress for interpreting individual query rankings, they still leave unaddressed the need for coherent global explanations that summarize feature influence across many queries.

Other approaches for ranking

Beyond feature attribution, other explainability strategies have been proposed for ranking models. Contrastive explanations, such as those by Castelnovo et al. [2], aim to clarify why certain items are ranked above others by articulating key distinguishing features. Counterfactual approaches, including Credence [9], generate minimal modifications to query or document features that would change ranking outcomes. Although valuable for individual cases, these methods do not provide a systematic view of global feature importance across datasets.

Global attribution potentially used for ranking

Permutation Feature Importance (PFI) and its conditional variants are well-established global post-hoc explainability techniques that assess feature importance by measuring performance changes when feature values are randomly permuted [3]. These methods can offer insights into overall feature influence across a dataset. However, in our context, the high correlation between many features in the MQ2008 dataset reduces the reliability of PFI and conditional PFI, as permutation-based methods may conflate feature contributions and produce misleading importance scores. Furthermore, the use of PFI for ranking models remains debated in the literature, especially when applied to correlated or structured inputs [5].

3 RESEARCH GAP

While local explainability methods for ranking, such as RankingSHAP [4] and ShaRP [8], have significantly advanced our ability to interpret individual query results or document pairs, they leave a fundamental question unanswered: *What consistently drives ranking decisions across an entire dataset*? Can we truly trust ranking models that shape user experience in search engines, recommendation systems, or admissions platforms without understanding their global drivers? The literature has largely focused on fine-grained, local insights that illuminate why a particular document is ranked higher for a specific query [11], but these do not satisfy the needs of stakeholders concerned with auditing, regulatory compliance, fairness, or system-level debugging.

Moreover, existing methods lack frameworks for systematically aggregating local attributions into coherent global explanations that align with human reasoning. Isolated local explanations—however precise—fail to provide the comprehensive picture necessary to evaluate how features influence model behavior at scale. This limitation is particularly problematic in sensitive applications such as search engines, where a lack of transparency at the global level can erode user trust, obscure systemic biases, and complicate accountability efforts. How can we ensure fairness, transparency, and reliability without knowing what consistently shapes the output of ranking models?

In response to these challenges, we propose an approach that aggregates local listwise attributions from RankingSHAP across queries, building a faithful global feature importance profile that bridges the gap between local interpretability and system-level understanding. Our methodology, detailed in the next section, aims to offer stakeholders a tool for auditing and improving ranking models with a view that generalizes beyond isolated predictions.

4 METHODOLOGY

Our approach builds upon recent advances in local feature attribution for ranking models, with the goal of constructing a faithful global feature importance profile that bridges local interpretability and system-level understanding. We focus on combining the strengths of listwise explanations with a simple yet robust aggregation strategy that supports model auditing and analysis at scale.

Local explanation generation

We compute local feature attributions using RankingSHAP [4] applied to a LambdaMART model, a gradient-boosted trees algorithm for ranking tasks [1]. RankingSHAP extends the SHAP framework to ranking problems by computing Shapley values not for individual predictions, but for entire ranked lists. It quantifies the contribution of each feature by measuring the change in a listwise ranking objective—such as Kendall's tau rank correlation—when feature values are perturbed through coalition sampling. This approach allows us to assess the importance of each feature with respect to the structure of the ranking as a whole, rather than isolated document scores. RankingSHAP generates masked perturbations for document vectors, ranks the perturbed lists, and evaluates changes in the explanation objective to estimate feature contributions. This method was selected for its alignment with the listwise nature of ranking models and its ability to provide explanations that are both theoretically grounded and practically meaningful in the context of ranked outputs.

Aggregation to global importance

To derive a global view of feature importance, we aggregate the local RankingSHAP attributions obtained across a representative set of queries. For each feature, we compute the mean of its local contributions across queries. We chose the mean as our aggregation statistic because it provides a simple, interpretable, and stable estimate of overall feature influence. The mean smooths out variability inherent in individual query-level explanations, mitigates the effect of outliers, and offers a robust summary that stakeholders can readily interpret. The resulting global feature profile is visualized through a bar plot, where features are represented on the y-axis and their aggregated importance scores on the x-axis. This visual explanation serves as the cornerstone of our global interpretability analysis, enabling users to identify which features consistently drive model decisions across queries.

Evaluation of explanation validity

To validate our global aggregation, we assess whether the top-k features identified through the global RankingSHAP profile are sufficient to maintain the model's performance when used alone. This analysis helps determine if the identified features generalize well beyond specific queries. We retrain the LambdaMART model using only the top-k features and compare its performance to two baselines: (i) models trained on k randomly selected features (random-k), and (ii) the full model trained on all available features. This comparative evaluation directly addresses our research gap by testing whether our global feature selection captures the key drivers of ranking decisions at scale.

We use standard metrics for learning-to-rank tasks to assess performance: NDCG@10, MAP, and Precision@10. NDCG@10 (Normalized Discounted Cumulative Gain at rank 10) measures the quality of the top of the ranked list, emphasizing the relevance of highly ranked documents. MAP (Mean Average Precision) evaluates overall ranking quality by considering precision at all ranks where relevant documents occur. Precision@10 provides a direct measure of the proportion of relevant documents within the top 10 positions. These metrics were chosen because they reflect different aspects of ranking quality and help determine whether the selected top-k features enable the model to generalize effectively, balancing interpretability and predictive performance.

In the next section, we present the experimental results validating the effectiveness of our approach.

5 EXPERIMENT AND ANALYSIS

Dataset and pre-modeling exploration

The dataset used in our study is MQ2008, part of the LETOR 4.0 benchmark collection [6]. MQ2008 is a standard benchmark for learning-to-rank research, widely adopted for its rich set of precomputed features and realistic search scenarios. It consists of over 800 queries, each associated with candidate documents retrieved from a web search engine. For each query-document pair, a relevance label is provided on a scale of 0 (not relevant), 1 (relevant), or 2 (highly relevant). The pairs are described by 46 numerical features capturing query-document matching signals, including textual relevance scores (e.g., TF-IDF, BM25), link-based metrics (e.g., PageRank), and URL structure features (e.g., URL length, number of slashes). Importantly, the dataset does not include query text, ensuring that both learning and interpretability focus solely on the numeric feature representations.

Before modeling, we conducted an exploration of the dataset's structure. We analyzed the distribution of documents per query in the MQ2008 test set and observed substantial variability, with many queries having fewer than 20 documents while a smaller subset contained rich document lists with over 50 documents. This variation motivated a sampling strategy designed to ensure that our analysis would reflect the diversity of query complexity present in the full dataset. We selected a subset of 50 queries from the original 136 in the test set, carefully preserving the distribution of document counts. Our subset included 10 rich queries (30 or more documents), 20 medium queries (11 to 29 documents), and 20 poor queries (fewer than 11 documents). This sampling approach balanced computational feasibility with the goal of maintaining a representative and interpretable evaluation set.

Furthermore, we computed statistical correlations between features and observed that several feature groups were highly correlated. This insight led us to exclude permutation feature importance (PFI) from our comparisons, as permutation-based methods can conflate the contributions of correlated features and produce misleading importance scores in such contexts. Our methodological choices were guided by the objective of delivering explanations that align with human reasoning and support reliable model interpretation at both local and global levels.

Model training and evaluation with RankingSHAP

In this step, we trained a listwise ranking model on the MQ2008 dataset using LambdaMART and applied a pre-trained RankingSHAP explainer to compute feature attributions. The objective was to interpret the ranking decisions produced by our model and quantify the contribution of each feature to the structure of the ranked lists. RankingSHAP was particularly well-suited for our purpose as it extends SHAP to ranking tasks, providing feature attributions at the listwise level in a theoretically grounded and consistent manner. By aligning with the listwise nature of our model, this method supports meaningful interpretation of ranking decisions in the context of entire query-document lists.

Global aggregation of feature attributions

Before aggregating, we examined feature attributions at the level of individual queries to better understand the variability in local explanations. Figures 1 and 2 illustrate the feature importance profiles for two representative queries (IDs 19383 and 19954). These plots highlight substantial differences: while certain features dominate in one query, a different subset of features emerges as most influential in another. Such variability, though informative locally, poses a challenge for stakeholders seeking a unified view of what drives model behavior at scale. This observation reinforces the necessity of aggregating local attributions into a coherent global profile that can serve as a reliable tool for system-level interpretation.

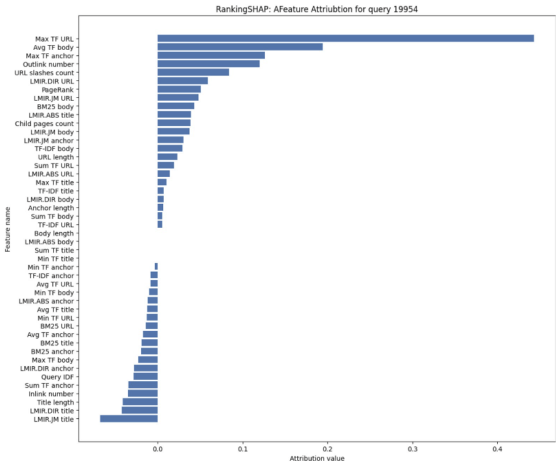

Figure 1: Feature attribution for query 19383 using RankingSHAP.

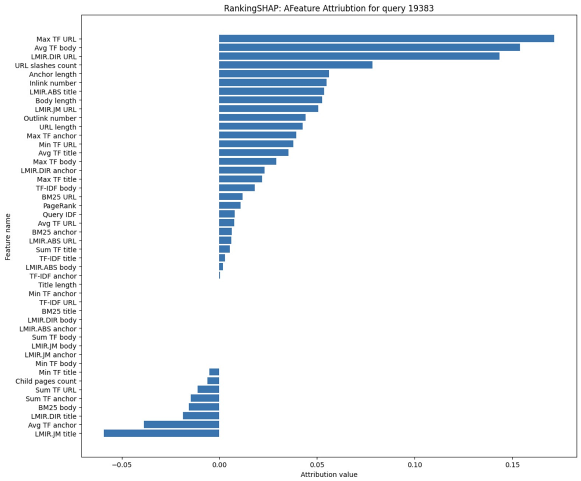

Figure 2: Feature attribution for query 19954 using RankingSHAP.

To move beyond these fragmented local explanations, we aggregated the individual RankingSHAP attributions obtained across the 50 selected queries. For each feature, we computed the mean attribution value across all queries. The choice of the mean was deliberate: it offers a simple, interpretable, and robust summary statistic that smooths out local variability while highlighting features that consistently drive ranking decisions across queries. This aggregation transforms per-query explanations into a coherent global feature profile that can be directly used for system-level interpretation and auditing.

The resulting global attributions are visualized in Figure 3, where features are shown on the y-axis and their mean importance scores on the x-axis. This bar plot constitutes the central tool supporting the global interpretability of our model, allowing stakeholders to identify at a glance which features exert the greatest influence on ranking outcomes. The aggregation not only provides stability, but also ensures that dominant patterns are faithfully captured, aligning with our objective of offering an explanation that generalizes beyond individual queries.

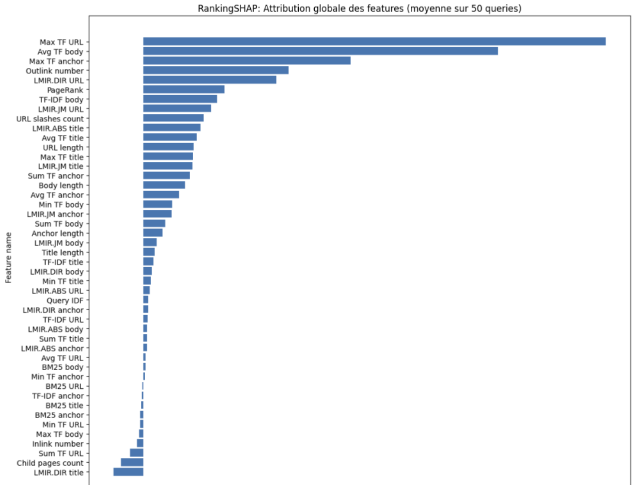

Figure 3: Global feature importance profile computed as the mean attribution over 50 queries.

Validation of top-k features through retraining and comparison

To assess whether the top features identified through our global RankingSHAP aggregation generalize beyond specific queries, we retrained ranking models using only these selected features. We chose $k = 10$ for a first experiment. The models were trained using LambdaMART implemented in LightGBM, a gradient-boosted decision tree framework tailored for ranking tasks. LambdaMART combines the boosting principle with a listwise loss function, optimizing ranking-specific objectives such as NDCG. This choice ensured that our evaluation remained consistent with our original listwise model while benefiting from LightGBM's efficiency and scalability.

We first trained a model using only the top-10 features identified by the global aggregation. We then trained additional models using randomly selected sets of 10 features, repeating this process five times to mitigate the effects of randomness and obtain stable average metrics. We compared the two approaches across standard learning-to-rank metrics: NDCG@10, MAP, and Precision@10. NDCG@10 measures ranking quality at the top of the list, aligning with user-focused relevance. MAP reflects the overall ability of the model to rank relevant documents highly across queries, while Precision@10 provides a direct measure of the proportion of relevant documents among the top results. These metrics are directly connected to our research gap, as they evaluate whether the selected top-k features retain the key drivers of ranking performance at a global level.

Table 1 summarizes the results of this comparison. The top-10 feature model consistently outperformed the random-10 feature models across all metrics, demonstrating that the selected features capture meaningful global patterns rather than noise. This validates the ability of our approach to identify features that generalize effectively and contribute to reliable model explanations.

Metric	Top-10	Random-10 (mean ± std)	Rel
NDCG@10	0.4684	0.4471 ± 0.0194	4.76%
MAP	0.1514	0.1428 ± 0.0086	6.02%
Precision@10	0.3179	0.3099 ± 0.0114	2.58%

Table 1: Comparison of ranking performance between top-10 globally selected features and random-10 features (mean over 5 runs). Rel refers to Relative gain which quantifies the improvement of top-10 over random-10 mean.

Validation of top-k features through comparison with the full model

To further evaluate the faithfulness of our global feature importance profile, we compared the performance of a model trained using only the top-10 globally selected features to that of a full model trained on all available features. The objective was to assess whether the global feature ranking captures genuinely useful information that generalizes across queries and supports reliable model predictions.

The comparison was conducted using NDCG@10, MAP, and Precision@10. These metrics, as discussed previously, measure the model's ability to produce high-quality ranked lists and are directly aligned with our research gap: they help determine whether a compact global explanation can faithfully reflect the key drivers of model behavior. Table 2 presents the results of this comparison. The top-10 model retained a high percentage of the full model's performance across all metrics, suggesting that the global feature selection provides a faithful and interpretable summary of the ranking model's behavior. While no explainability approach can claim to capture the full complexity of a model, these results indicate that our global aggregation offers a robust approximation that supports system-level understanding.

Validation of top-k features with varying k

To further assess the faithfulness and generalizability of our global feature importance profile, we conducted additional experiments in which we varied the number of top features k used to retrain the ranking model. We explored values of k ranging from 5 to 20. This range was chosen to balance two objectives: (i) avoid models that are too sparse and risk omitting critical features when k is very small, and (ii) examine whether increasing k beyond a certain point yields diminishing returns in performance relative to model complexity.

Metric	Top-10	Full model	Retention (%)
NDCG@10	0.468	0.476	98.42
MAP	0.151	0.155	97.52
Precision@10	0.318	0.324	98.02

Table 2: Comparison of ranking performance between the top-10 globally selected features and the full model using all features. Retention quantifies the top-10 score as a percentage of the full model performance.

For each value of k, we retrained models using the top-k globally selected features and compared their performance to that of models trained on k randomly selected features. We also evaluated the retention of full model performance as k increased. In this context, we define a model as generalizing sufficiently well if it retains at least 98% of the full model's performance across the chosen metrics, while also outperforming the random-k baselines in a consistent and meaningful way.

Figures 4 and 5 summarize the results of these experiments. In Figure 4, we observe that the top-k models progressively approach the performance of the full model as k increases, with retention exceeding 98% of full model performance for NDCG@10, MAP, and Precision@10 starting from around $k = 10–12$. This indicates that the selected features successfully capture the core drivers of model behavior, supporting a compact yet faithful explanation of the ranking decisions.

In Figure 5, we see that the top-k models consistently outperform the random-k models across all k values, particularly at smaller k, where the choice of features has a stronger impact. The gain over random-k diminishes as k increases, as both models converge towards using a larger subset of the available features. This trend highlights the importance of feature selection in compact models and underscores the value of our global aggregation approach in identifying the most influential features.

These results illustrate the trade-off between model simplicity and performance: while larger k values reduce variability and close the gap with the full model, they offer limited interpretability gains beyond a certain point. Our findings suggest that a range of k between 10 and 12 achieves an effective balance, providing strong generalization while maintaining a compact and interpretable feature set. Notably, $k = 10$—our initial reference point—falls within this optimal range, further validating our approach.

This analysis directly addresses our research gap by demonstrating that it is possible to derive a global feature profile that faithfully summarizes model behavior at scale. The selected features not only generalize effectively but also provide a practical basis for auditing, debugging, and understanding ranking models in high-stakes applications.

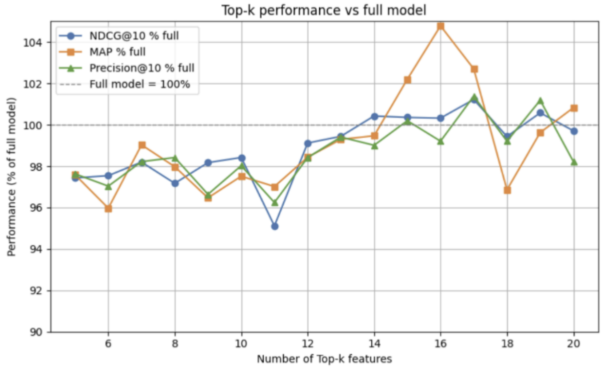

Figure 4: Top-k model performance as a percentage of full model performance across NDCG@10, MAP, and Precision@10, as k varies from 5 to 20.

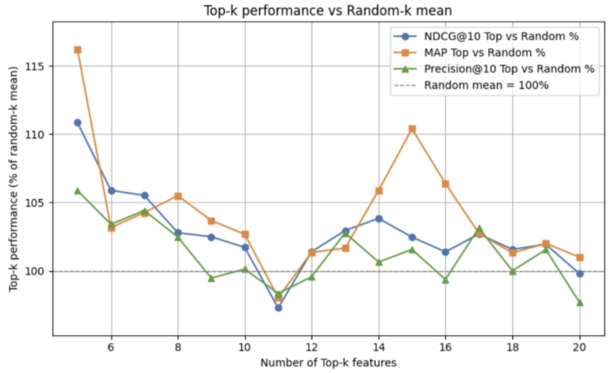

Figure 5: Top-k model performance as a percentage of random-k mean performance across NDCG@10, MAP, and Precision@10, as k varies from 5 to 20.

6 CONCLUSIONS

In this work, we proposed a methodology for deriving global feature importance profiles for ranking models by aggregating local listwise attributions generated using RankingSHAP. This approach addresses a key research gap by bridging local interpretability and system-level understanding, enabling a coherent global view of what drives model decisions across queries. We validated our methodology through retraining experiments, demonstrating that models based on top globally selected features can achieve performance comparable to full-feature models while offering a more compact and interpretable representation.

Despite these contributions, our approach presents certain limitations. The aggregation was performed using simple averaging, which may overlook complex interactions between features. The methodology was evaluated on a single dataset, limiting the generalizability of the findings across domains. Furthermore, the work focused on feature importance and did not address other critical aspects of ranking model behavior, such as fairness, robustness to adversarial inputs, or temporal dynamics in evolving query distributions.

Future work could extend this study by applying the methodology to diverse datasets and ranking tasks, exploring alternative aggregation strategies, and integrating fairness-aware or user-centered evaluation frameworks. These directions would help further strengthen the role of global explanations in supporting transparent, accountable, and trustworthy ranking systems.

REFERENCES

[1] Chris Burges. 2010. From RankNet to LambdaRank to LambdaMART: An overview. In *Learning*, Vol. 11. 81.

[2] Alessandro Castelnovo et al. 2024. Evaluative item-contrastive explanations in rankings. *Cognitive Computation* 16, 6 (2024), 3035–3050.

[3] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20, 177 (2019), 1–81.

[4] Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. RankingSHAP–Listwise Feature Attribution Explanations for Ranking Models. *arXiv preprint arXiv:2403.16085* (2024).

[5] Giles Hooker, Lucas Mentch, and Xinran Zhou. 2021. The dangers of underestimating the importance of correlation in feature importance estimates. *arXiv preprint arXiv:2103.00681* (2021).

[6] Tie-Yan Liu. 2009. *Learning to rank for information retrieval*. Vol. 3. Now Publishers. 225–331 pages.

[7] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[8] Venetia Pliatsika et al. 2025. ShaRP: A Novel Feature Importance Framework for Ranking. *arXiv preprint arXiv:2401.16744* (2025).

[9] Joel Rorseth et al. 2023. Credence: Counterfactual explanations for document ranking. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 3631–3634.

[10] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[11] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000* (2021).