A Robustness Assessment of Query Performance Prediction (QPP) Methods Based on Risk-Sensitive Analysis

Ricardo Marçal de Andrade Nascimento
ricardomarcal02@hotmail.com
Instituto Federal de Educação, Ciência e Tecnologia de Goiás
Anápolis, Brazil

Daniel Xavier de Sousa
daniel.sousa@ifg.edu.br
Instituto Federal de Educação, Ciência e Tecnologia de Goiás
Anápolis, Brazil

Guglielmo Faggioli
guglielmo.faggioli@phd.unipd.it
Università degli Studi di Padova
Padova, Italy

Paulo José Lage Alvarenga
pjlalvarenga@ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Nicola Ferro
ferro@dei.unipd.it
Università degli Studi di Padova
Padova, Italy

Marcos André Gonçalves
mgoncalv@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Abstract

Query Performance Prediction (QPP) estimates Information Retrieval (IR) systems' effectiveness without relying on manual relevance judgments. A central challenge in QPP lies in its unstable performance, which may vary significantly across queries. In parallel, the concept of *risk-sensitive evaluation* in IR seeks to enhance robustness by minimizing performance variance and mitigating poor retrieval outcomes. Despite commonalities and complementarities, existing research has failed to integrate these two perspectives, specifically by attempting to apply risk-sensitive metrics to enhance QPP evaluation robustness. Indeed, current QPP assessments, typically based on correlation measures and the sMARE framework, insufficiently address robustness, potentially incurring into misleading conclusions. This paper proposes a novel risk-sensitive evaluation methodology to assess QPP robustness. Through empirical analysis on the Deep Learning'19, Deep Learning'20, and Robust'04 datasets, we demonstrate that high correlation does not necessarily imply robustness. Risk-aware metrics such as U_{RISK}, T_{RISK}, and *GeoRisk* uncover critical variations in QPP performance, offering statistically sound insights with reduced variability. Our findings underscore the value of incorporating risk-sensitive evaluation into QPP, ultimately contributing to the development of more reliable and robust IR systems.
Code: https://github.com/RicardoMarcal/qpp-risk-evaluator

CCS Concepts

• **Information systems → Evaluation of retrieval results**; **Retrieval effectiveness**.

Keywords

Query Performance Prediction; Risk-Sensitiveness; IR Evaluation

ICTIR '25, Padua, Italy.

ACM ISBN 979-8-4007-1861-8/25/07
https://doi.org/10.1145/3731120.3744611

ACM Reference Format:
Ricardo Marçal de Andrade Nascimento, Daniel Xavier de Sousa, Guglielmo Faggioli, Paulo José Lage Alvarenga, Nicola Ferro, and Marcos André Gonçalves. 2025. A Robustness Assessment of Query Performance Prediction (QPP) Methods Based on Risk-Sensitive Analysis. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR '25), July 18, 2025, Padua, Italy*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3731120.3744611

1 Introduction

Robustness is a key property of Information Retrieval (IR) models, referring to their ability to mitigate performance variability across different user queries. Its primary objective is to *reduce the likelihood of poor performance on individual queries—thus improving user experience—while maintaining strong overall retrieval effectiveness [44].* Query failures and inconsistent response quality have been shown to significantly impact user satisfaction with IR systems [23].

Although supervised ranking models are often highly effective, they frequently overlook the critical aspect of robustness. Beyond achieving strong average performance, a robust model should also minimize the risk of poor outcomes on individual queries. However, effectiveness and robustness are often competing objectives—models optimized solely for (average) effectiveness may fail to ensure consistent performance across diverse queries. In this context, Risk-sensitiveness evaluates retrieval models based on their performance variability, where higher variability signals lower robustness [14, 27, 37, 44]. Risk-sensitive metrics—such as U_{RISK} [44], T_{RISK} [13], and *GeoRisk* [14]—favor IR methods that are less prone to failure relative to (multiple) baseline approaches.

In a related line of research, Query Performance Prediction (QPP) methods aim to predict the performance of an IR system without relying on manual relevance judgments [4]. A major limitation of these methods is their high variability, influenced by factors such as the IR system, the corpus, and the query topics. As a result, a QPP method may perform well on certain queries while failing—sometimes severely—on others. Although this issue is well-documented, **robust evaluation methodologies on existing QPP methods remain lacking**. Existing approaches, such as scaled Mean Absolute Rank Error (sMARE) [19], typically rely on correlation metrics or topic-wise effectiveness, without directly accounting for prediction robustness.

We tackle these issues in this paper. In particular, we argue that correlations and sMARE metrics fall short of delivering a comprehensive QPP robustness evaluation. These metrics may yield misleading evaluations, as high correlation or sMARE values for specific queries do not guarantee strong prediction performance across all queries. Specifically, we defend that *a detailed evaluation considering the distribution of correlations and their variances relative to other QPP methods is essential to avoid incomplete conclusions.*

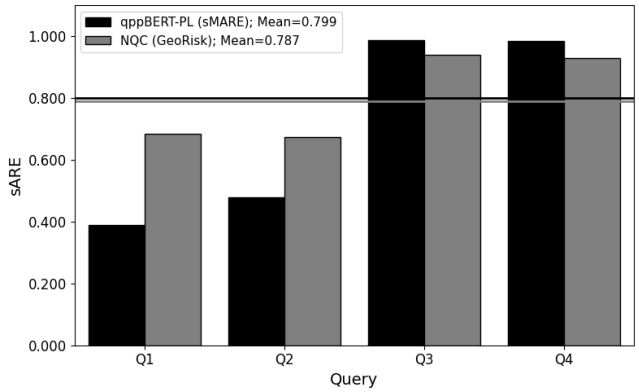


Figure 1: Real-world use case comparing qppBERT-PL(sMARE) and NQC(GeoRisk) on Robust'04 Dataset using BM25 as retrieval system.

Thus, as a *main contribution* of this paper, we demonstrate how Risk-Sensitive metrics (**RSMs**) may help to deal with current QPP evaluation limitations. We advocate that, as done for RSMs that consider multiple baselines, QPP methods should also be evaluated by considering the variance among one or several (other) methods. Accordingly, we propose a new evaluation measure— not a new QPP— aimed at assessing the robustness of predictors relative to one another. Our goal is not to replace existing evaluation measures with robustness-based ones, but to complement them. Just as IR models are evaluated across multiple dimensions (e.g., precision, efficiency), we argue that QPPs should also be assessed on multiple fronts, including risk-sensitiveness.

To clarify this issue, Figure 1 illustrates a real-world downstream scenario comparing two QPP methods, qppBERT-PL [11] and NQC [36], on the Robust'04 dataset[1]. According to sMARE [19], qppBERT-PL outperforms NQC overall, achieving a higher average sARE (0.799 vs. 0.787). However, qppBERT-PL fails on queries Q1 and Q2, where NQC shows superior performance, indicating greater robustness. Traditional QPP metrics, which weigh deviations uniformly across queries, may favor methods that perform well on average but overlook critical failures. In contrast, GeoRisk highlights overall precision while penalizing poor individual outcomes, providing a more balanced and robust evaluation.looseness=-1

More specifically, our main hypotheses are that: (i) *current QPP evaluation strategies, e.g., correlations and sMARE, lack robustness (risk-sensitive) objective criteria, usher to misleading conclusions when considering robustness as a critical requirement*; and (ii) while risk-sensitive metrics are capable of evaluating specific robustness properties that current QPP evaluation cannot, *they can also capture properties that these evaluation strategies can, such as the positive correlations between QPP method's scores and effectiveness, e.g. Average Precision (AP).* To provide evidence for these hypotheses, we break them into the following research questions, empirically answered:

- **RQ1**: Considering a straightforward application of correlation and sMARE evaluations, are the most effective QPP methods also the most robust (less risky) ones?

To answer this question, we apply well-known RSMs (U_{RISK}, T_{RISK}, and *GeoRisk*) to evaluate several QPP methods. We compare these QPP methods using a range of evaluation metrics, including correlation (Kendall's τ, Pearson's r, and Spearman's ρ), sMARE and risk-sensitiveness. Our experimental results are based on the publicly available and well-known Deep Learning'19, Deep Learning'20 and Robust'04 datasets, revealing that the answer to RQ1 is **NO**. Specifically, some QPP methods that achieve higher effectiveness correlations do not necessarily demonstrate high robustness. Conversely, methods with slightly lower correlation scores may exhibit superior risk-sensitive performance.

- **RQ2**: What other relationships do exist between risk-sensitiveness and current QPP evaluation strategies (correlations and sMARE), especially regarding query difficulty and variability?

This question investigates additional relationships between Query Performance Prediction (QPP) and risk-sensitive evaluation. Notably, our experiments indicate that RSMs are more responsive to instances where a QPP method underperforms relative to the baselines. In contrast, traditional metrics such as correlation and sMARE tend to favour overall effectiveness, often overlooking sensitivity to poor predictions. Among the evaluated RSMs, *GeoRisk* demonstrates the most consistent performance across the three datasets, likely due to its incorporation of multiple baselines. Moreover, our empirical findings reveal that RSMs offer statistically more robust evaluations, characterized by narrower confidence intervals compared to conventional QPP evaluation methodologies.

In summary, this paper´s main contributions are: (i) a novel risk-sensitive evaluation methodology for QPP methods leveraging the complementarity between risk-sensitiveness and QPP and offering new insights into their robustness and (ii) a thorough evaluation of such methodology using three datasets, 21 QPP methods, three risk-sensitive metrics (U_{RISK}, T_{RISK}, and *GeoRisk*), and 4 QPP evaluation approaches (Kendall's τ, Pearson's r, Spearman's ρ and sMARE), totalizing more than 750 analysed results.

2 Related Work

As mentioned, robustness in Information Retrieval (IR) has been previously analyzed from two distinct perspectives, which exploit different but potentially complementary strategies: (i) one focusing on predicting query difficulty [48] and (ii) the second aimed at analyzing the *risk* of retrieval methods to produce poor query results [37]. Methods that predict query difficulty analyze factors such as unexpected vocabulary, ambiguous content, or missing information to suggest that some queries may have a high potential to produce poor results [43]. This set of techniques is known as *Query Performance Prediction (QPP in short)* and is employed to enable alternative methods to handle more challenging queries [12, 20, 28, 32, 40]. By design, QPP aims to estimate the retrieval quality of a query's results in the absence of true relevance judgments for each (query, document) pair.

[1]Both, methods and datasets are described in Section 4.

In this second perspective, IR methods are assessed by their performance *variability*, considering strong results for some queries and poor results for others. High variability indicates low robustness. The IR subfield addressing robustness through performance variance is known as *Risk-Sensitiveness*, a established area in IR [44]. RSMs seek to favor methods less prone to failure compared to baselines.

Key RSMs such as U_{RISK} [44], T_{RISK} [13], and *GeoRisk* [14] are not only applied to evaluate robustness but are also exploited as objective functions in learning-to-rank methods to improve model effectiveness [37]. According to these metrics, models that achieve strong average effectiveness while minimizing negative deviations compared to other ranking models (baseline performance) are more risk-sensitive (less risky) or robust.

QPP methods are often evaluated by analyzing the relationship between a query-assigned score of difficulty and the actual system effectiveness measured by standard IR metrics, typically AP or NDCG. In QPP studies, this relationship is usually assessed using a correlation coefficient, such as Kendall's τ, Pearson's r, and Spearman's ρ correlation metrics[4, 17]. A QPP method that achieves a higher correlation value is generally considered a superior approach. This comparison summarizes the evaluation between methods considering only a single (value), with a limited (per-query) perspective and lack of statistical analyses.

With these limitations in mind, the authors of [19] introduced a framework called sMARE, which estimates QPP performance for each query based on the distance between the QPP-predicted rank and the expected effectiveness rank. sMARE [19] provides a distributional evaluation of QPP performance across multiple queries, enabling more comprehensive statistical assessments.

To the best of our knowledge, this is the first work to evaluate RSMs in the context of QPP, based on the hypothesis that existing QPP evaluation methods inadequately address robustness. We analyze robustness by comparing QPP methods and RSMs, highlighting their differences and similarities in handling strong and weak results.

3 Background and QPP Evaluation Methodology

Risk-sensitive metrics have been increasingly utilized to evaluate various IR-related tasks, including Learning to Rank[14, 37] and Recommender Systems [23]. These evaluations generally focus on two key perspectives: (i) the variance in predictive performance across different retrieval models for a single query and (ii) the variance across multiple queries for a specific retrieval model. Risk-sensitive methods often calculate the performance difference between a given model M and baseline models B, along with the associated variance in several queries. RSMs place particular emphasis on scenarios where the difference is negative, highlighting instances where model M underperforms compared to B.

We argue that traditional QPP evaluation methods fail to account for these two perspectives (i and ii). As such, these evaluation methods lack a rigorous risk-sensitive assessment. They neither measure the magnitude of errors relative to baselines nor consider variances across queries and baselines, limiting their ability to provide a comprehensive robustness evaluation. This shortcoming in existing literature motivates the evaluation framework proposed in this work.

To address this gap, we propose leveraging the variance in performance among queries and QPP methods, as well as the error differences relative to QPP baselines, to enable a more robust QPP evaluation. Specifically, we assess the magnitude of improvement or degradation of a QPP method M relative to one (or many) QPP baseline B. For this, we adopt the sMARE metric as an evaluation strategy (described in 3.2), incorporating into it both, (i) a distributional assessment and (ii) risk-sensitive metrics. These metrics are applied to scenarios involving a single baseline (detailed in Section 3.3) and multiple baselines (explored in Section 3.4).

Before introducing our methodology, we describe the correlation metrics used in QPP and their limitations to assess robustness.

3.1 Correlation-based QPP evaluation

Correlations metrics have been employed in numerous QPP studies to evaluate the relationship between a difficulty ranking score of a query q and the effective ranking produced for q, measured by AP or NDCG[4]. Following [18, 19], the correlation metrics more commonly used in QPP are: Spearman's ρ, Pearson's r, and Kendall's τ.

Spearman's ρ evaluates the monotonic ranking values relationship, emphasizing overall ranking consistency and robustness to outliers. Pearson's r measures the linear correlation between predicted and actual query performance, reflecting the strength and direction of their relationship without relying on ranks. And Kendall's τ focuses on ranking concordance, quantifying the swaps needed for alignment, and offering an intuitive metric of rank agreement. We suggest referring to [4] for a more detailed review.

One limitation of these metrics for the sake of QPP evaluation is their inability to capture the magnitude of errors for individual queries. Additionally, because these metrics produce a single correlation value for the entire set of queries, resulting in one single-point evaluation for QPP methods, they may overlook finer-grained evaluation aspects. To address these shortcomings, [19] introduced the sMARE function, incorporating additional error information.

3.2 Scaled Absolute Rank Error and Scaled Mean Absolute Rank Error

Motivated by the need to evaluate QPP methods using a more fine-grained approach, Faggioli et al. [19] proposed a new function – scaled Absolute Rank Error (sARE) – to move from point-wise QPP metrics to a distributional performance. sARE is a function designed to assess each (QPP method, query) combination individually and summarize the overall performance.

sARE computes the difference in rank positions for each query as assigned by a QPP method versus the ground truth rank position determined by an effectiveness metric (such as AP or NDCG) assigned by a retrieval method. Ties in rankings are resolved using the average of the ranks spanned, following the approach described in [21]. More in detail, sARE is computed as:

$$sARE(q) = \frac{|Q| - |r_q^i - r_q^e|}{|Q|} \tag{1}$$

Here, r_q^i and r_q^e denote the ranks assigned by the QPP method i and the effectiveness evaluation for query q, respectively, and Q represents the set of queries. In the original paper [19], the sARE follows a "lower-is-better" approach, where smaller values of $r_q^i - r_q^e$ indicate smaller difference (error) between the QPP score and the effectiveness evaluation. To adapt this for RSMs, Eq. 1 presents an inverted version of sARE.

To summarize the sARE values across all queries, the sMARE is defined in [19] as:

$$sMARE(Q) = \frac{1}{|Q|} \sum_{q \in Q} sARE(q) \tag{2}$$

The primary contribution of Eq. 2 lies in its ability to focus on the distribution of rank differences for each query. However, it does not account for the variance across different baselines or among multiple queries. To address this limitation, our work leverages the query-error feature of sARE to introduce the first QPP evaluation with risk-sensitive capabilities. Specifically, we utilize Equation 1 to assess the performance of each (QPP method, query) combination.

As detailed next, we integrate sARE as a core component of our risk-sensitive evaluation framework, utilizing it for conducting a comprehensive comparative analysis across multiple QPP methods.

3.3 Risk-Sensitiveness - One Baseline

In the context of this work, a QPP method is deemed risk-sensitive if it effectively predicts most queries without producing significantly worse predictions for others, when compared to a QPP baseline. Accordingly, we perform a multi-objective evaluation of QPP, considering the quality of the predictions and the QPP's ability to minimize the risk of poor predictions. To assess the robustness of a system, it is common to employ a Risk-Sensitive metric.

A seminal work in Risk-Sensitive is [44]. It divides robustness into two components: *degradation* and *reward*. Degradation (reward) of a model M represents the negative (positive) change in query performance relative to a specific baseline IR system B. More formally, considering a set of training queries Q and two ranking models (or QPP methods in case of our evaluation): a baseline B and a proposed model M. The degradation of model M is defined as the average difference (or gain) in effectiveness between the baseline B and M across all queries in Q. Wang et al. [44] define the concept of degradation using the F_{RISK} function, as outlined in Eq. 3.

$$F_{RISK}(Q,M) = \frac{1}{|Q|} \sum_{q \in Q} max[0, B(q) - M(q)] \tag{3}$$

In this context, $B(q)$ and $M(q)$ represent the effectiveness values of the baseline and the proposed model for a given query q, respectively. The F_{RISK} function utilizes the effectiveness scores of each query in Q. Here we use the sARE metric as defined in Eq. 1, which follows a higher-is-better approach, as normally occurs in IR metrics. Thus, the primary purpose of F_{RISK} function is to compare the performance of two models (or QPP methods) based on the sARE metrics. A lower value of the F_{RISK} function indicates a more robust model, as it reflects a smaller degree of degradation for model M compared to baseline B.

In contrast to degradation, *reward* of a proposed method M relative to a baseline model B is defined as the average improvement in effectiveness of M over B across all queries in Q. In [44], the reward is formally defined as in Eq.4:

$$F_{REWARD}(Q,M) = \frac{1}{|Q|} \sum_{q \in Q} max[0, M(q) - B(q)] \tag{4}$$

Reward and degradation can be combined in various ways to assess the extent to which a method M is sensitive to risk. In [44], both functions F_{RISK} and F_{REWARD} are combined, introducing the U_{RISK} function. This function is defined as:

$$U_{RISK}(Q,M) = F_{REWARD}(Q,M) - (1+\alpha) F_{RISK}(Q,M) \tag{5}$$

The parameter α is the weight given to the degradation. Different values of α can significantly impact the risk-sensitive evaluation of the method [13, 44]. In our case, $\alpha = 0$ provided results that are more similar to existent QPP metrics, without considering the higher weight to the degradation.

Another risk-sensitive metric was introduced in [13]. In that work, the authors extend [44] by proposing a generalization of the U_{RISK} function, which is referred to as T_{RISK}.

$$T_{RISK}(Q,M) = \frac{U_{RISK}(Q,M)}{SE(U_{RISK}(Q,M))} \tag{6}$$

where SE is the estimation of the U_{RISK} standard error. The original papers suggest using the regular standard error of the mean to U_{RISK}, that it is, $\sigma(U_{RISK})/\sqrt{|Q|}$, where $|Q|$ means the cardinality of Q_T and σ the variance of values in U_{RISK}.

T_{RISK} leverages inferential hypothesis testing to provide a risk-sensitive metric. Considering our application in QPP methods, we use the same inferential methods in [13] to assess if the observed level of risk for a QPP method is statistically significant and identify specific queries that contribute to a substantial level of risk individually.

One important issue of functions T_{RISK} and U_{RISK} is that both use only one baseline to evaluate the variance of a specific query. Considering this, [15] examines how the selection of the baseline impacts the risk-sensitive evaluation. The authors demonstrate that choosing an appropriate baseline is crucial for achieving an unbiased assessment of the risk-sensitive performance of individual systems. Specifically, they find that the higher the correlation between a given system M and the baseline across queries, the greater the average risk-sensitive scores of M. This indicates a bias in the estimation of risks. To address this, the paper proposes unbiased baselines that use the mean ranking performance across multiple ranking methods. Alternatively, the authors of [14] suggest that not only one method should be used as a baseline but several.

3.4 Risk-Sensitiveness - Many Baselines

Dinçer et al. [14] suggest using multiple baselines systems for risk-sensitive evaluation. They consider not only the mean and variance of observed losses and gains against a baseline method but also the shape of the score distribution when employing a set of different methods as risk baselines. Dinçer et al. argue that utilizing a group of systems as baselines provides a more accurate understanding of query difficulty, mitigating the influence of queries poorly handled by a system but not by others. To achieve this, Dinçer et al. employ χ^2 test statistics to estimate the expected ranking effectiveness for each query based on the combined performance of the evaluated and the baseline systems. Dinçer et al. [14] define Z_{RISK} as:

$$Z_{RISK}(i) = \left[\sum_{q \in Q_+} z_{iq} + (1+\alpha) \sum_{q \in Q_-} z_{iq} \right] \tag{7}$$

where

$$z_{iq} = \frac{x_{iq} - e_{iq}}{\sqrt{e_{iq}}}, e_{iq} = S_i \times \frac{T_q}{N}, \tag{8}$$

For sake of QPP evaluation, we propose x_{iq} be the sARE metric (defined in Eq. 1) performed for a query q corresponding to a QPP method i. Element i is defined as $i \in \{1,2,...,r\}$ for each QPP method, where r is the number of methods, and $|Q|$ is the max size of queries. Q_+ (Q_-) stands for the sum of positive (negative) values of z_{iq}. Let S_i be the performance (sARE) sum for all queries in QPP method i, T_q the performance (sARE) sum for all QPP methods for a specific query q, and $N = \sum_{i=1}^{r} \sum_{q=1}^{Q} x_{iq}$ the sum of all elements.

Accordingly to [14], Z_{RISK} assesses the risk sensitiveness regarding the variance of the model i, therefore it does not provide a comparative risk-sensitive evaluation between QPP methods. Thus, following [14], we also use the Geometric Mean in the *GeoRisk* formula:

$$GeoRisk(S_i) = \sqrt{S_i/|Q| \times \Phi(Z_{Risk}(i)/|Q|)} \quad (9)$$

where $\Phi()$ represents the cumulative distribution function of the Standard Normal Distribution. Essentially in our work, *GeoRisk* offers a comparison of QPP methods from a robustness standpoint, assessing each query based on its sARE performance expectation. This expectation is derived from the distribution of observed query difficulty across QPP methods for a specific query.

We claim that sARE can be an appropriate function for evaluating individual query-level performance in QPP outcomes. As described in [19], sARE and sMARE show strong correlation with other traditional QPP evaluation metrics. In extensive evaluations reported in [2, 24, 25], sARE emerges as the most well-grounded choice.

In sum, as a major contribution of this paper, and unlike other risk-sensitive studies in the literature, in here we present a novel adaptation of risk-sensitiveness for evaluating QPP methods from a robustness perspective. This involves incorporating the concepts of sARE into Z_{RISK}, enabling a thorough and robust QPP evaluation.

4 Experimental Results and Discussion

This section presents the experimental results and evaluation when robustness is incorporated into QPPs. The primary objective is to analyze and compare the results of applying the risk-sensitiveness metrics U_{RISK}, T_{RISK}, and *GeoRisk* to QPP methods with those produced by conventional QPP evaluation strategies, such as Pearson's r, Spearman's ρ, Kendall's τ, and sMARE[2].

4.1 Experimental Setup

Experiments are conducted on three well-known benchmark collections: Robust'04 [42], Deep Learning'19 passages [5] and Deep Learning'20 [6]. These collections differ in terms of the number of topics, the underlying corpus, and the availability of training queries. Robust'04 comprises 249 topics and is commonly used for ad-hoc document Text Information Retrieval. Deep Learning'19 includes 43 annotated topics from the MS-MARCO collection, making it particularly relevant for neural retrieval scenarios. Deep Learning'20 uses the same passages of DL 2019, providing 54 queries. Using these datasets ensures diversity, supporting a robust evaluation of the proposed framework across different retrieval paradigms.

To comprehensively evaluate our robustness experiments, we analyze the results of the main categories of QPP methods: (i) pre-retrieval, (ii) post-retrieval, and (iii) deep learning [29, 35, 36]. For pre-retrieval, we consider SCSsum, SCQavg, SCQmax, ICTFavg, ICTFmax, IDFavg, IDFmax, VARavg, and VARmax [4]. As post-retrieval methods, we include: Clarity [7], Weighted Information Gain (WIG) [48], Normalized Query Commitment (NQC) [36], Score Magnitude and Variance (SMV) [39], and their Utility Estimation Framework (UEF) versions [34]. Lastly, we consider the following supervised QPPs: neural-QPP [45], BERT-QPP [1], qppBERT-PL[11], deepQPP [9]. In the Appendix, the reader can find a brief description of these methods.

As the target IR system, we use BM25 considering Average Precision (AP) as the main metric, following a large part of the previous literature in the QPP domain[26].

We fine-tuned post-retrieval unsupervised methods by considering different cutoffs of the ranked list: 5, 10, 50, 100, and 500. This fine-tuning was conducted using a two-fold partitioning procedure, as detailed in [10, 34, 45, 46]. To enhance reliability, this approach was repeated 30 times. Concerning supervised QPPs we employ the scores precomputed by Saha et al. [31] and publicly available[3].

Statistical significance is tested using one-way ANOVA [30] with Tukey's Honestly Significant Differences (HSD) post-hoc comparison [41] to correct for multiple comparisons.

To evaluate the RSMs, we consider specific baseline values. For U_{RISK} and T_{RISK}, we use the average of sARE scores for all evaluated QPPs, as the average is regarded as an unbiased strategy [15]. For *GeoRisk* evaluation, we consider sARE for all QPP methods except the one under evaluation, following the approach outlined in [38]. For the α parameter, we adhere to the guidelines in [38], evaluating the outcomes for $\alpha = 1,5,10,20$. For clarity, we present the results for $\alpha = 5$ and $\alpha = 10$, as the other values have similar interpretations.

4.2 Results for RQ1

To answer RQ1, we propose two experiments. In section 4.2.1, we analyze traditional QPP evaluations (i.e., Spearman's ρ, Pearson's r, Kendall's τ and sMARE) and the proposed evaluation framework using RSMs across various QPP methods. In Section 4.2.2, we assess the win-loss performance of the best QPP methods selected by the traditional QPP evaluation metrics, distilling the robustness differences.

4.2.1 ***Overall QPP Evaluation.*** We begin by confirming in Tables 1a and 1b that the rankings of the most effective QPP methods identified by all correlation metrics and sMARE are not necessarily the same as the most robust ones, according to the RSMs, mainly considering the larger Robust'04. Tables 1a and 1b summarize the evaluation of the Robust'04 and Deep Learning'20 datasets, respectively, using multiple metrics, including Spearman's ρ, Pearson's r, Kendall's τ, sMARE, *GeoRisk* (with $\alpha = 5$ and $\alpha = 10$), U_{RISK} (with $\alpha = 5$ and $\alpha = 10$), and T_{RISK} (with $\alpha = 5$ and $\alpha = 10$). For Deep Learning'19 dataset, the results are described in the Appendix. To enhance the clarity, the tables present the rank position for each QPP method according to each metric — in the Appendix, the reader can find the absolute values corresponding to these tables. The rank position is based on the average of 1,000 bootstrap samples across all queries, along with its statistical significance (using the Wilcoxon signed-rank test [22]) against the

[2]All correlation metrics formulas are defined in [4].

[3]https://github.com/souravsaha/qpp-comb

QPP	τ	ρ	r	sMARE	*GeoRisk*		U_{RISK}		T_{RISK}	
					$\alpha=5$	$\alpha=10$	$\alpha=5$	$\alpha=10$	$\alpha=5$	$\alpha=10$
SMV	1.5	1	1.5	5.5	3	3	2.5	2.5	4.5	6
UEFSMV	1.5	6	5	5.5	4	4.5	4	4	4.5	5
UEFNQC	3.5	3.5	3	3.5	2	2	2.5	2.5	3	3.5
qppBERT-PL	3.5	3.5	1.5	1	15	17	10.5	13	2	2
NQC	5.5	2	5	7	1	1	1	1	7	8
BERT-QPP	5.5	5	5	2	17	20	14	15	6	3.5
deepQPP	7	7.5	8	3.5	18	21	16	16.5	1	1

(a) Robust'04

QPP	τ	ρ	r	sMARE	*GeoRisk*		U_{RISK}		T_{RISK}	
					$\alpha=5$	$\alpha=10$	$\alpha=5$	$\alpha=10$	$\alpha=5$	$\alpha=10$
NQC	1	2.5	1	1	1	1	1	1	1	1
neural-QPP	2	1	3	3	9	12	6	8.5	3.5	3.5
BERT-QPP	3.5	4.5	4	3	13	13	10	10.5	2	2
qppBERT-PL	3.5	2.5	2	3	14	14.5	10	10.5	8.5	12
VARavg	5	10	5	6	3.5	4	3	3	8.5	10.5
IDFavg	6	11	7	5	2	2	2	2	3.5	3.5
ICTFavg	7.5	12	8	7.5	3.5	3	4	4	5	5

(b) Deep Learning'20

Table 1: QPP methods ranking induced by different measures. In case of statistical ties, we apply the 'average' tie-break approach.

methods ranked immediately above and below. To avoid discrepancies in the minimum rank among the methods, we apply average (fractional) ranking. For reference and comparison, we use the ordering of the QPP methods given by Kendall's τ correlations. For clarity, Figures 1a and 1b show the ranking order of only the most representative QPP methods – for all results, the reader can refer to the Appendix.

From Table 1a and following Kendall's τ ranking, SMV and UEFSMV (tied up in first place) are the most effective QPP methods in Robust'04, followed by UEFSMV, and qppBERT-PL. All correlation metrics suggested SMV as the best method. Spearman's ρ, Kendall's τ and Pearson's r defined a similar third place, UEFNQC. And differently of Kendall's τ which places UEFSMV as first position, Spearman's ρ and Pearson's r place UEFSMV as sixth and fifth position. Considering sMARE as a criterion, the most effective QPP method is qppBERT-PL, followed by BERT-QPP and UEFNQC.

However, when using RSMs, specifically *GeoRisk* and U_{RISK} (both with $\alpha=5$ and $\alpha=10$), NQC emerges as the most robust method. In fact, there is a convergence across these two RSMs, defining NQC as the most robust method, followed by UEFNQC, SMV, and UEFSMV. T_{RISK} produces less consistent results with regard to the other RSMs, with a different method in the second position: qppBERT-PL.

We perform a similar analysis on Deep Learning'20 in Table 1b. Similarly as before, in Table 1b, *GeoRisk* ($\alpha=5$ and $\alpha=10$) ranks IDFavg and ICTFavg in second and third place, respectively. However, for correlation-based metrics, neural-QPP, BERT-QPP, and qppBERT-PL are more frequently ranked in second and third positions. Considering Deep Learning'19, the results are described in our Appendix.

Considering a macro-view rank analysis of Tables 1a, 1b, and for Deep Learning'19, we use Table 2 to report a pairwise Kendall's τ correlation between the ranks produced by classic QPP evaluation strategies and RSMs, averaged across the datasets. In general, the correlation is high, suggesting that the RSMs generally agree with the QPP evaluation approaches. At the same time, we highlight differences between rankings, especially for RSMs $\alpha=10$, confirming our previous observations that RSMs bring novel perspectives for QPP evaluation. As a general trend, U_{RISK} is the metric that better aligns with classic QPP evaluation, followed by *GeoRisk* and T_{RISK}. Specially for T_{RISK}, there is a strong correlation with sMARE, up to 0.86 when considering $\alpha=5$.

The literature highlights *GeoRisk* as the most consistent and robust risk-sensitive metric [14, 38], as it evaluates variance across a set of baselines rather than relying on a single one, as is the case with U_{RISK} [14] and T_{RISK} [15]. In our experiments (Table 2), *GeoRisk* and U_{RISK} exhibit strong correlations, with values ranging from 0.72 to 0.89. In contrast, T_{RISK} produces a ranking more closely aligned with sMARE, showing correlation values between 0.82 and 0.86 (Table 2), while *GeoRisk* presents a weaker alignment, ranging from 0.42 to 0.60. A plausible explanation for the divergence of T_{RISK} from U_{RISK} and *GeoRisk* is that T_{RISK} applies a linear transformation to U_{RISK} (Equation 6) [13], which may reduce its robustness to outliers.

We also provide an in-depth evaluation of the alpha parameters (Eqs. 8 and 5) in the context of QPP evaluation, varying $\alpha=1,5,10,20$. As shown in both Figure 6 for Robust'04 in the Appendix and Table 2 (for $\alpha=5,10$), we observe that as the alpha values increase, GeoRisk and URisk become less correlated with traditional correlation-based metrics and place greater emphasis on robustness. While this behavior helps minimize poor results, excessively high alpha values can unfairly penalize methods that achieve high average effectiveness, potentially compromising overall performance. We consider values around 5 and 10 to be well-suited for QPP evaluation, accepting a very similar overall average and reducing the risk of bad solutions.

In summary, we find greater divergence between traditional QPP evaluations and risk-sensitive assessments in the larger datasets (Robust'04 and Deep Learning'20) than in the smaller dataset (Deep Learning'19), although notable differences remain in the latter. Among RSMs, *GeoRisk* and U_{RISK} offer a more innovative perspective than T_{RISK}. While traditional QPP metrics and RSMs show some correlation, each provides a distinct and complementary evaluative view.

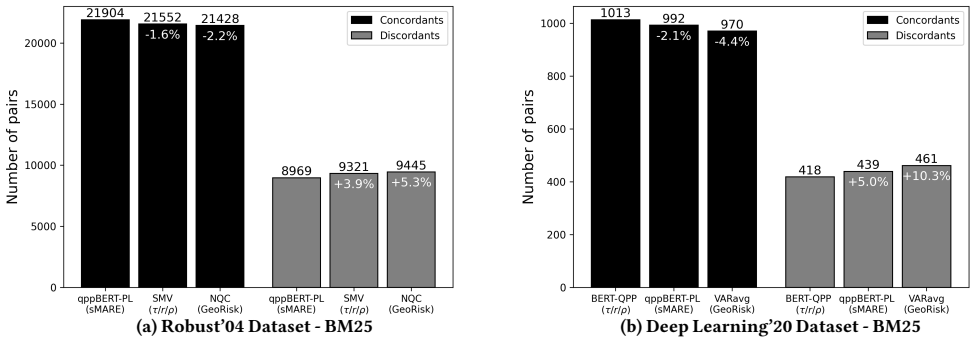
4.2.2 ***Breaking Down the Robustness Evaluation.*** Considering the traditional correlation metrics and sMARE (Eq. 2), it is evident that the evaluation for each query is aggregated and summarized over the total number of queries (or pairs, in the case of certain correlation metrics). This aggregation can obscure poor individual performances by averaging them with stronger results. Therefore, to ensure a more comprehensive and robust evaluation, it is essential to examine whether a given method effectively mitigates poor outcomes and, consequently, minimizes user dissatisfaction.

To this end, we select the most effective QPP methods as determined by Pearson's r, Spearman's ρ, Kendall's τ, sMARE, and *GeoRisk* (Tables 1a and 1b). We then proceed to quantify both favourable and unfavourable outcomes. The evaluation is decomposed along two complementary dimensions: (i) enumerating the number of concordant and discordant query pairs, and (ii) assessing the magnitude of the associated errors. The core objective is to demonstrate there is no significant difference in the frequency of correct versus incorrect orderings between query pairs. However, when evaluating the magnitude (or impact) of the error, the difference between robust and non-robust results becomes much clearer.

This analysis begins with Figures 2a and 2b, which depict the number of concordant and discordant query pairs for Robust'04 and

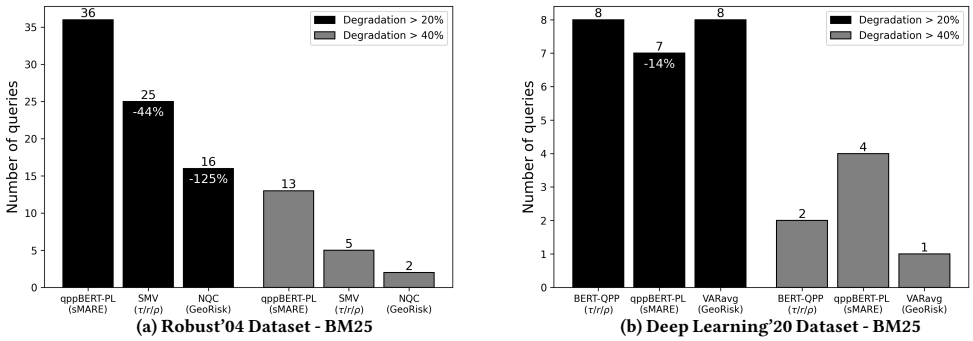
Table 2: Correlation between different evaluation approaches. The good, but not pathological, correlation between the risk-sensitive evaluation approaches ($GeoRisk$, T_{RISK} and U_{RISK}) and classical correlation metrics suggests that the metrics capture similar patterns—each with its peculiarities.

	Kendall's τ	Pearson's r	Spearman's ρ	sMARE	$GeoRisk$ ($\alpha=5$)	$GeoRisk$ ($\alpha=10$)	T_{RISK} ($\alpha=5$)	T_{RISK} ($\alpha=10$)	U_{RISK} ($\alpha=5$)	U_{RISK} ($\alpha=10$)
Kendall's τ	1.000	0.761	0.933	0.868	0.653	0.506	0.809	0.755	0.761	0.729
Pearson's r	0.761	1.000	0.783	0.708	0.542	0.440	0.672	0.650	0.643	0.612
Spearman's ρ	0.933	0.783	1.000	0.884	0.624	0.478	0.806	0.752	0.733	0.701
sMARE	0.868	0.708	0.884	1.000	0.597	0.450	0.861	0.826	0.718	0.679
$GeoRisk(\alpha=5)$	0.653	0.542	0.624	0.597	1.000	0.853	0.602	0.567	0.860	0.898
$GeoRisk(\alpha=10)$	0.506	0.440	0.478	0.450	0.853	1.000	0.455	0.426	0.720	0.758
$T_{Risk}(\alpha=5)$	0.809	0.672	0.806	0.861	0.602	0.455	1.000	0.940	0.710	0.672
$T_{Risk}(\alpha=10)$	0.755	0.650	0.752	0.826	0.567	0.426	0.940	1.000	0.662	0.630
$U_{Risk}(\alpha=5)$	0.761	0.643	0.733	0.718	0.860	0.720	0.710	0.662	1.000	0.962
$U_{Risk}(\alpha=10)$	0.729	0.612	0.701	0.679	0.898	0.758	0.672	0.630	0.962	1.000



(a) Robust'04 Dataset - BM25 **(b) Deep Learning'20 Dataset - BM25**

Figure 2: Concordance between QPP and AP ranking scores



(a) Robust'04 Dataset - BM25 **(b) Deep Learning'20 Dataset - BM25**

Figure 3: Queries with high sARE degradation

Deep Learning'20, respectively.[4] These comparisons adopt BM25 and AP as the ground truth, following standard practices in QPP evaluations [16]. In the figures, the y-axis denotes the number of query pairs, while the x-axis represents the QPP methods selected by each evaluation metric. For Robust'04 (as summarized in Table 1a), SMV emerges as the top method according to Kendall's τ, Spearman's ρ, and Pearson's r (indicated in the table as τ, ρ, and r, respectively); qppBERT-PL is selected by sMARE, and NQC by $GeoRisk$. In the case of Deep Learning'20 (Figure 2b), due to a convergence in top-performing models, the third-best methods are considered: BERT-QPP is selected by Kendall's τ, Pearson's r and Spearman's ρ, qppBERT-PL is selected by sMARE and VARavg selected by $GeoRisk$ and Kendall's τ, respectively. In Figure 2a, qppBERT-PL is highlighted as having a greater number of concordant pairs (and fewer discordant ones) than BM25, and therefore the reported percentages correspond to its performance. The same rationale applies to BERT-QPP in Figure 2b.

As anticipated, Figures 2a and 2b reveal relatively minor differences in the number of concordant and discordant query pairs among the evaluated QPP methods. For instance, in Figure 2a, qppBERT-PL (selected by sMARE) exhibits only 1.6% and 2.2% more concordant pairs than SMV (selected by correlation-based metrics) and NQC (selected by $GeoRisk$), respectively. In contrast, SMV and NQC yield approximately 3.9% and 5.3% more discordant pairs than qppBERT-PL. For Deep Learning'20, the disparities in concordant and discordant pairs are slightly more pronounced, surpassing 4.4% and 10.3%, respectively. Nonetheless, these differences remain relatively modest across the evaluated methods.

We turn our attention to the magnitude of prediction error, specifically examining poor outcomes or severe degradations relative to

[4]Due to space constraints, this evaluation is presented for Robust'04 and Deep Learning'20 only.

an expected value for each query. Figures 3a and 3b present the number of queries for which the degradation—computed using Equation 3—exceeds thresholds of 20% and 40% when compared to a baseline method B. In this context, baseline B denotes the average effectiveness (sARE) per query, calculated across all evaluated methods excluding the one under analysis. As seen in Figure 3a, the method selected by *GeoRisk* (NQC) predicts 125% fewer queries with more than 20% degradation than the methods identified by sMARE and the correlation-based metrics—namely, qppBERT-PL and SMV, respectively. When considering the set of queries with degradation exceeding 40%, NQC continues to yield fewer predictions with large errors. Comparable patterns are observed for Deep Learning'20 (Figure 3b).

Although the QPP methods selected by correlation metrics and sMARE, qppBERT-PL and SMV, may exhibit slightly higher concordance for overall effectiveness, they are also associated with a greater likelihood of producing poor outcomes. These poor outcomes refer to instances where a query is incorrectly predicted as either more difficult or easier than it truly is. Specifically, qppBERT-PL and SMV have more instances of prediction error — defined as degradation exceeding 20% — compared to NQC, the method selected by *GeoRisk*. Conversely, while the methods identified by *GeoRisk* may demonstrate marginally lower concordance in effectiveness (by approximately 1.6% to 10.3%), they result in markedly fewer severe prediction errors.

4.3 Results for RQ2

To answer RQ2, we focus our analysis on the performance of the best QPP methods identified by Kendall's τ[5], sMARE, and RSMs in the previous section, providing a more sound statistical evaluation.

Evaluating the most difficulty queries. The performance of QPP methods is influenced by various factors, including data content, document-query similarity, and vocabulary specificity [4]. As a result, certain queries may receive suboptimal effectiveness predictions. Relying predominantly on metrics such as Kendall's τ and sMARE may compromise the evaluation quality, particularly for more challenging queries. To investigate whether RSMs are more attentive to such difficult cases, Figure 4a presents a comparison of the best-performing QPP methods according to Kendall's τ (SMV) and *GeoRisk* (NQC), based on their Kendall's τ correlation scores on Robust'04. Similarly, Figure 4b examines the sMARE performance of the top methods identified by sMARE (qppBERT-PL) and *GeoRisk* (NQC) on Deep Learning'20, focusing on subsets of queries identified as the most difficult.

To identify the most challenging queries, we grouped all queries according to their lowest average effectiveness, following the methodology established in [4]. The most difficult queries—those exhibiting the lowest average effectiveness across all QPP methods—constitute the first group of columns, with subsequent groups representing progressively easier queries. In both figures, each group comprises 20% of the most difficult queries, as correlation measures tend to exhibit greater stability when calculated over larger query groups [3].

Figures 4a and 4b illustrate that the method selected by *GeoRisk* provides a more effective evaluation for the most challenging queries. In Figure 4a, for the top 20% most difficult queries, NQC (identified by *GeoRisk*) achieves a correlation score that is 92% higher than that of SMV, the method selected by Kendall's τ. For the subsequent group, representing the 20%–40% most difficult queries, NQC demonstrates an 11.6% improvement in correlation. Comparable results are observed when evaluating sMARE on Deep Learning'20.

Statistical Evaluation. We now assess the effectiveness of RSMs in identifying Statistically Significantly Different (s.s.d.) QPPs pairs. Determining whether the performance difference between two QPPs is statistically significant enables the exclusion of ineffective methods. The ability of an evaluation protocol—comprising a performance metric and a statistical test—to identify s.s.d. pairs is referred to as its "power" or "test sensitivity." If the objective is to filter out the least effective QPPs, prioritizing the most promising ones, an evaluation procedure with higher statistical power is preferable. This is typically desirable, especially in late stages of the evaluation, where a single "system" must be chosen to be put in production.

Correlation-based metrics provide a list-wise evaluation—yielding a single value for the entire query set—which precludes the application of statistical significance testing[6]. The same limitation applies to RSMs. Thus, following [19], we adopt a bootstrap-based procedure to construct confidence intervals for our evaluation metrics.

For each dataset, we perform 1,000 bootstrap resampling iterations, drawing with replacement subsets of queries equal in size to the original query set. For each QPP, retrieval method, and sampled subset, we compute the list-wise evaluation metrics, including correlation and risk-sensitive measures. The 95% confidence interval around the performance estimate is derived by identifying the 0.025 and 0.975 percentiles of the performance distribution across the resampled subsets. For a given collection and retrieval method, a QPP is considered statistically superior to another if the lower bound of its confidence interval exceeds the upper bound of the other method's interval. In the case of sMARE, where confidence intervals can be computed directly, we employ Studentized t-based confidence intervals at the 0.05 significance level.

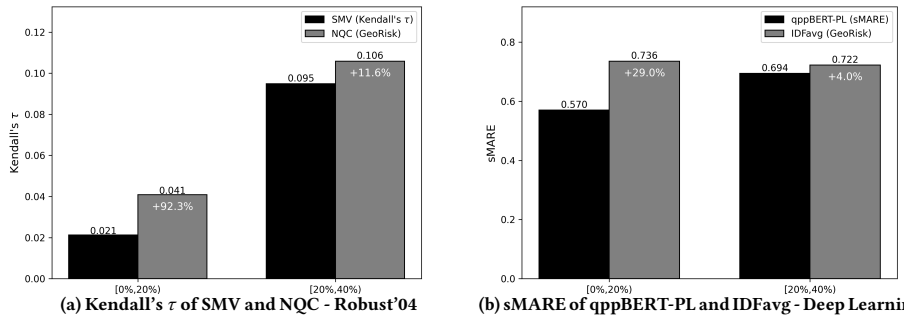
Figure 5 shows the distribution of confidence intervals for the QPP methods across different evaluation measures. **Smaller intervals indicate more accurate estimates**. We report only RSMs with $\alpha = 5$ for clarity. As shown, *GeoRisk* yields consistently small confidence intervals, indicating higher accuracy compared to other measures. In contrast, U_{RISK} produces larger intervals than both sMARE and *GeoRisk*, with its intervals spanning the full range for Deep Learning'19 and Deep Learning'20.

We now investigate quantitatively which metric provides the most powerful evaluation. Table 3 reports the number of s.s.d. pairs of QPP methods identified by the various metrics, divided by collection. Notice that, since we considered 21 different QPPs, in total there are 210 $(21 \times (21-1)/2)$ pairs of QPPs to be compared. Besides the average pairs of s.s.d. systems, we report in Table 3 also the pairs of systems considered not s.s.d. ($\overline{\text{s.s.d.}}$), as well as the ratio over the total of pairs.

Regarding Deep Learning'19, U_{RISK} is the most powerful metric. On average, depending on the value of α, it recognizes 7 to 10 more pairs than Kendall's τ and 10 to 13 more s.s.d. pairs than sMARE. *GeoRisk* has a similar power, with more power for $\alpha = 1$ and $\alpha = 5$. The same pattern repeats almost unchanged for Deep Learning'20.

[5]We do not evaluate Spearman's ρ and Pearson's r in this part because they are correlated with the selection of Kendall's τ.

[6]Importantly, we are comparing pairs of QPPs and are interested in determining whether their performance differs significantly. This is distinct from assessing whether the correlation of individual QPPs is greater than zero.



(a) Kendall's τ of SMV and NQC - Robust'04 (b) sMARE of qppBERT-PL and IDFavg - Deep Learning'20.

Figure 4: Evaluation of query intervals ordered by difficulty (i.e., ascending by AP) utilizing Kendall's τ and sMARE for their respective top-performing QPP methods, alongside *GeoRisk*'s leading QPP method.

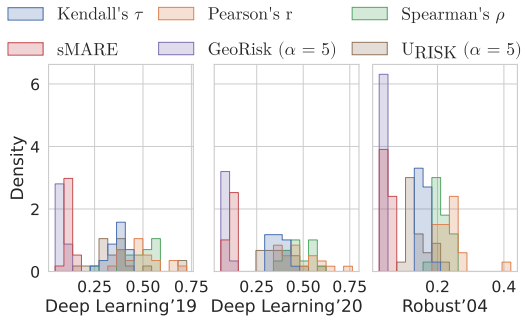


Figure 5: Distribution of the confidence interval sizes. GeoRisk show smaller confidence intervals (purple bars towards the left side), indicating greater discriminative power. The more the curves are shifted to the left (indicating smaller confidence intervals), the more powerful the test is.

In general, the dataset allows us to discriminate less between QPPs, but U_{RISK} and *GeoRisk* remain the most discriminative measures. For Robust'04, classical approaches tend to be more powerful. In particular, Kendall's τ and sMARE are the most discriminating solutions, followed by Spearman's ρ. T_{RISK} is always the most conservative metric. This aligns with the fact that, compared to other solutions, T_{RISK} has much wider confidence intervals.

Based on these findings, we recommend using TRisk in the early stages of developing a new QPP method—when it is desirable to explore a wide range of hypotheses and retain flexibility in evaluating various factors beyond effectiveness and robustness (e.g., efficiency, latency, energy consumption, cost). In contrast, URisk and GeoRisk are likely more suitable for later stages, when the goal is to make a definitive selection for deployment. Their higher statistical power supports making more confident decisions about which predictor performs best.

Overall, we observe that all RSMs represent promising strategies for evaluating QPP methods, as they offer a complementary perspective—enabling practitioners to assess these techniques not only in terms of predictive accuracy, but also with respect to their stability across diverse scenarios.

5 Conclusion

This work presents the first comprehensive analysis of Query Performance Prediction (QPP) methods through the lens of Risk-Sensitive metrics (RSMs). Although QPP and risk-sensitive approaches aim to enhance robustness, existing evaluations overlook key aspects. We address this gap by enabling the assessment of: (i) variability in predictive performance across QPP methods for a single query, and (ii) variability across queries for a specific QPP method. By integrating the sARE into RSMs, we overcome limitations inherent in traditional evaluation strategies, which often yield incomplete or misleading conclusions regarding robustness.

Table 3: Number of s.s.d. pairs of QPP methods found by different QPP evaluation strategies. Proportion (prop.) indicates the number of s.s.d. pairs found over the total (210 possible pairs with 21 considered QPPs). The larger number of s.s.d. pairs identified by the Risk-based approaches, *GeoRisk* and U_{RISK} in particular, confirms their stronger statistical power.

collection	Deep Learning'19			Deep Learning'20			Robust'04		
	s.s.d.	$\overline{\text{s.s.d.}}$	ratio	s.s.d.	$\overline{\text{s.s.d.}}$	ratio	s.s.d.	$\overline{\text{s.s.d.}}$	ratio
Kendall's τ	71	139	0.34	30	180	0.14	**110**	**100**	**0.52**
Pearson's r	75	135	0.36	32	178	0.15	89	121	0.42
Spearman's ρ	76	134	0.36	29	181	0.14	109	101	0.52
sMARE	68	142	0.32	26	184	0.12	**110**	**100**	**0.52**
GeoRisk (α=1)	78	132	0.37	34	176	0.16	103	107	0.49
GeoRisk (α=5)	78	132	0.37	36	174	0.17	97	113	0.46
GeoRisk (α=10)	70	140	0.33	39	171	0.19	105	105	0.50
T_{RISK} (α=1)	64	146	0.30	27	183	0.13	103	107	0.49
T_{RISK} (α=5)	60	150	0.29	26	184	0.12	83	127	0.40
T_{RISK} (α=10)	52	158	0.25	22	188	0.10	72	138	0.34
U_{RISK} (α=1)	78	132	0.37	34	176	0.16	101	109	0.48
U_{RISK} (α=5)	**81**	**129**	**0.39**	**43**	**167**	**0.20**	102	108	0.49
U_{RISK} (α=10)	80	130	0.38	**43**	**167**	**0.20**	103	107	0.49

Our extensive evaluation on Robust'04, Deep Learning'19 and '20, encompassing 21 QPP methods, four traditional metrics (Kendall's τ, Pearson's r, sMARE, and Spearman's ρ), and three RSMs (*GeoRisk*, U_{RISK}, and T_{RISK}), demonstrates the advantages of Risk-Sensitive evaluation. In particular, *GeoRisk* consistently identified methods with significantly fewer poor predictions—e.g., 40% fewer queries with degradation exceeding 20%—and proves to be the most effective metric for capturing robustness across diverse conditions. Our results emphasize the need to integrate RSMs into QPP evaluations to better comprehend the trade-offs between effectiveness and robustness. Future work will extend the proposed framework to other retrieval tasks and datasets and refine metrics to enhance the reliability of IR systems even further. Finally, we intend to provide a more in-depth query-feature analysis to explain the reasons why certain QPPs fare better under GeoRisk.

Acknowledgments

This work is supported by CNPq (443011/2023-0 and 403184/2021-5), FAPEMIG (APQ-00262-22), AWS (421773/2022-7), the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (INCT-TILD-IAR 408490/2024-1) and UNIFEI.

References

[1] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. BERT-QPP: Contextualized Pre-trained transformers for Query Performance Prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 2857–2861. doi:10.1145/3459637.3482063

[2] Negar Arabzadeh, Chuan Meng, Mohammad Aliannejadi, and Ebrahim Bagheri. 2024. Query Performance Prediction: Techniques and Applications in Modern Information Retrieval. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Tokyo, Japan) *(SIGIR-AP 2024)*. Association for Computing Machinery, New York, NY, USA, 291–294. doi:10.1145/3673791.3698438

[3] Chris Buckley and Ellen M. Voorhees. 2017. Evaluating Evaluation Measure Stability. *SIGIR Forum* 51, 2 (Aug. 2017), 235–242. doi:10.1145/3130348.3130373

[4] David Carmel and Elad Yom-Tov. 2010. *Estimating the query difficulty for information retrieval.* Morgan & Claypool Publishers, 1537 Fourth St, Ste 228. San Rafael, CA 94901. United States of America.

[5] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* 2020-March (3 2020), 1–22. https://arxiv.org/abs/2003.07820v2

[6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. arXiv:2003.07820 [cs.IR] https://arxiv.org/abs/2003.07820

[7] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland) *(SIGIR '02)*. Association for Computing Machinery, New York, NY, USA, 299–306. doi:10.1145/564376.564429

[8] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2004. A framework for selective query expansion. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (Washington, D.C., USA) *(CIKM '04)*. Association for Computing Machinery, New York, NY, USA, 236–237. doi:10.1145/1031171.1031220

[9] Suchana Datta, Debasis Ganguly, Derek Greene, and Mandar Mitra. 2022. Deep-QPP: A Pairwise Interaction-based Deep Learning Model for Supervised Query Performance Prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 201–209. doi:10.1145/3488560.3498491

[10] Suchana Datta, Debasis Ganguly, Mandar Mitra, and Derek Greene. 2022. A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants. *ACM Trans. Inf. Syst.* 41, 2, Article 38 (Dec. 2022), 31 pages. doi:10.1145/3545112

[11] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. A 'Pointwise-Query, Listwise-Document' based Query Performance Prediction Approach. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2148–2153. doi:10.1145/3477495.3531821

[12] Giorgio Maria Di Nunzio, Guglielmo Faggioli, Alessandro Micarelli, Giuseppe Sansonetti, and Giuseppe D ' Aniello. 2021. A Study of a Gain Based Approach for Query Aspects in Recall Oriented Tasks. *Applied Sciences 2021, Vol. 11, Page 9075* 11, 19 (9 2021), 9075. doi:10.3390/app11199075

[13] B. Taner Dinçer, Craig Macdonald, and Iadh Ounis. 2014. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 23–32. doi:10.1145/2600428.2609625

[14] B. Taner Dinçer, Craig Macdonald, and Iadh Ounis. 2016. Risk-sensitive evaluation and learning to rank using multiple baselines. In *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(Sigir '16)*. Association for Computing Machinery, Inc, New York, USA, 483–492. doi:10.1145/2911451.2911511

[15] B. Taner Dinçer, Iadh Ounis, and Craig MacDonald. 2014. Tackling Biased Baselines in the Risk-Sensitive Evaluation of Retrieval Systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8416 Lncs (2014), 26–38. doi:10.1007/978-3-319-06028-6{_}3

[16] Guglielmo Faggioli, Thibault Formal, Simon Lupart, Stefano Marchesin, Stephane Clinchant, Nicola Ferro, and Benjamin Piwowarski. 2023. Towards Query Performance Prediction for Neural Information Retrieval: Challenges and Opportunities. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval* (Taipei, Taiwan) *(ICTIR '23)*. Association for Computing Machinery, New York, NY, USA, 51–63. doi:10.1145/3578337.3605142

[17] Guglielmo Faggioli, Thibault Formal, Stefano Marchesin, Stéphane Clinchant, Nicola Ferro, and Benjamin Piwowarski. 2023. Query Performance Prediction for Neural IR: Are We There Yet?. In *European Conference on Information Retrieval*. Springer, Springer, Dublin, Ireland, 232–248. doi:10.1007/978-3-031-28244-7{_}15

[18] Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. 2021. An Enhanced Evaluation Framework for Query Performance Prediction. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, Cham, 115–129.

[19] Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. 2022. sMARE: a new paradigm to evaluate and understand query performance prediction methods. *Information Retrieval Journal* 25, 2 (6 2022), 94–122. doi:10.1007/s10791-022-09407-w

[20] Debasis Ganguly and Emine Yilmaz. 2023. Query-specific Variable Depth Pooling via Query Performance Prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2303–2307. doi:10.1145/3539618.3592046

[21] Jean Dickinson Gibbons and Subhabrata Chakraborti. 2021. *Nonparametric Statistical Inference.* Chapman and Hall/CRC, Boca Raton.

[22] Jer Guang Hsieh, Yih Lon Lin, and Jyh Horng Jeng. 2008. Preliminary study on Wilcoxon learning machines. *Journal of IEEE Transactions on Neural Networks and Learning Systems* 19, 2 (2008), 201–211.

[23] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (10 2012), 441–504. doi:10.1007/s11257-011-9118-4/metrics

[24] Chuan Meng. 2024. Query Performance Prediction for Conversational Search and Beyond. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 3077. doi:10.1145/3626772.3657658

[25] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Query Performance Prediction using Relevance Judgments Generated by Large Language Models. arXiv:2404.01012 [cs.IR] https://arxiv.org/abs/2404.01012

[26] Josiane Mothe. 2023. On correlation to evaluate QPP. In *Query Performance Prediction and Its Evaluation in New Tasks Workshop (QPP++ 2023) co-located with 45th ECIR (CEUR Workshop Proceedings, Vol. 3366)*, Guglielmo Faggioli, Nicola Ferro, Josiane Mothe, and Fiana Raiber (Eds.). CEUR-WS. org, Springer, Aachen, 29–36. http://ceur-ws.org/Vol-3366/#paper-06

[27] Josiane Mothe and Md. Zia Ullah. 2023. Selective Query Processing: A Risk-Sensitive Selection of Search Configurations. *ACM Trans. Inf. Syst.* 42, 1, Article 31 (Aug. 2023), 35 pages. doi:10.1145/3608474

[28] Haggai Roitman. 2018. Enhanced Performance Prediction of Fusion-based Retrieval. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval* (Tianjin, China) *(ICTIR '18)*. Association for Computing Machinery, New York, NY, USA, 195–198. doi:10.1145/3234944.3234950

[29] Haggai Roitman, Shai Erera, Oren Sar-Shalom, and Bar Weiner. 2017. Enhanced Mean Retrieval Score Estimation for Query Performance Prediction. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 35–42. doi:10.1145/3121050.3121051

[30] Andrew Rutherford. 2013. *ANOVA and ANCOVA: A GLM Approach: Second Edition.* Wiley Blackwell, Hoboken, New Jersey.

[31] Sourav Saha, Suchana Datta, Dwaipayan Roy, Mandar Mitra, and Derek Greene. 2025. Combining Query Performance Predictors: A Reproducibility Study. In *Advances in Information Retrieval*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (Eds.). Springer Nature Switzerland, Cham, 112–129.

[32] Harrisen Scells, Leif Azzopardi, Guido Zuccon, and Bevan Koopman. 2018. Query Variation Performance Prediction for Systematic Reviews. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 1089–1092. doi:10.1145/3209978.3210078

[33] Falk Scholer, Hugh E. Williams, and Andrew Turpin. 2004. Query association surrogates for Web search: Research Articles. *J. Am. Soc. Inf. Sci. Technol.* 55, 7 (May 2004), 637–650.

[34] Anna Shtok, Oren Kurland, and David Carmel. 2009. Predicting Query Performance by Query-Drift Estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5766 Lncs (2009), 305–312. doi:10.1007/978-3-642-04417-5{_}30

[35] Anna Shtok, Oren Kurland, and David Carmel. 2010. Using statistical decision theory and relevance models for query-performance prediction. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 259–266. doi:10.1145/1835449.1835494

[36] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting Query Performance by Query-Drift Estimation. *ACM Trans. Inf. Syst.* 30, 2, Article 11 (May 2012), 35 pages. doi:10.1145/2180868.2180873

[37] Pedro Henrique Silva Rodrigues, Daniel Xavier Sousa, Thierson Couto Rosa, and Marcos André Gonçalves. 2022. Risk-Sensitive Deep Neural Learning to Rank.

In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 803–813. doi:10.1145/3477495.3532056

[38] Daniel Xavier Sousa, Sérgio Canuto, Marcos André Gonçalves, Thierson Couto Rosa, and Wellington Santos Martins. 2019. Risk-sensitive learning to rank with evolutionary multi-objective feature selection. *ACM Transactions on Information Systems* 37, 2, Article 24 (2 2019), 34 pages. doi:10.1145/3300196

[39] Yongquan Tao and Shengli Wu. 2014. Query Performance Prediction By Considering Score Magnitude and Variance Together. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (Shanghai, China) *(CIKM '14)*. Association for Computing Machinery, New York, NY, USA, 1891–1894. doi:10.1145/2661829.2661906

[40] Paul Thomas, Falk Scholer, Peter Bailey, and Alistair Moffat. 2017. Tasks, Queries, and Rankers in Pre-Retrieval Performance Prediction. In *Proceedings of the 22nd Australasian Document Computing Symposium* (Brisbane, QLD, Australia) *(ADCS '17)*. Association for Computing Machinery, New York, NY, USA, Article 11, 4 pages. doi:10.1145/3166072.3166079

[41] John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (6 1949), 99. doi:10.2307/3001913

[42] Ellen Voorhees. 2005. Overview of the TREC 2004 Robust Retrieval Track. doi:10.6028/NIST.SP.500-261

[43] Ellen M. Voorhees. 2005. The TREC robust retrieval track. *ACM SIGIR Forum* 39, 1 (6 2005), 11–20. doi:10.1145/1067268.1067272

[44] Lidan Wang, Paul N. Bennett, and Kevyn Collins-Thompson. 2012. Robust ranking models via risk-sensitive optimization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 761–770. doi:10.1145/2348283.2348385

[45] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural Query Performance Prediction using Weak Supervision from Multiple Signals. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 105–114. doi:10.1145/3209978.3210041

[46] Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpepper. 2019. Information Needs, Queries, and Query Performance Prediction. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 395–404. doi:10.1145/3331184.3331253

[47] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4956 Lncs (2008), 52–64. doi:10.1007/978-3-540-78646-7{_}8/cover

[48] Yun Zhou and W. Bruce Croft. 2007. Query performance prediction in web search environments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 543–550. doi:10.1145/1277741.1277835

A Extended Evaluation and Results

A.1 Extended Evaluation

The tables 4, 5 and 6 present the evaluation for 21 QPP methods, considering the average of 1,000 bootstrap samples across all queries. We highlight the rank order of each QPP method for the respective evaluation metric using **bold** numbers in parentheses. For reference and comparisons, we use the ordering of the QPP methods given by Kendall's τ correlations.

Figure 6 presents heatmaps of Kendall's τ correlations between rankings from traditional QPP metrics (Kendall's τ, Pearson's r, Spearman's ρ, sMARE) and RSMs across α values (0 to 20). Correlations generally decline as α increases, indicating greater divergence between RSMs and classical QPP evaluations. This trend is most pronounced for *GeoRisk* (Figure 6a), while U_{RISK} (Figure 6b) and T_{RISK} (Figure 6c) display more stable correlations.

B Brief QPP methods explanation

The table 7 presents a detailed breakdown of the 18 QPP predictors that were analyzed, categorized based on their operational stage (pre-retrieval, post-retrieval, and deep learning-based methods), offering different perspectives on query difficulty and retrieval quality. It provides a quick reference for their characteristics and operational context. The summarization of all QPP methods evaluated on this work is described on table 7. They are organized as Pre-retrieval, Post-retrieval and Deep Learning.

QPP's	Kendall's τ	Pearson's r	Spearman's ρ	sMARE	$GeoRisk$ ($\alpha=5$)	$GeoRisk$ ($\alpha=10$)	U_{RISK} ($\alpha=5$)	U_{RISK} ($\alpha=10$)	T_{RISK} ($\alpha=5$)	T_{RISK} ($\alpha=10$)
SMV	0.395 (**1.5**)	0.577 (**1**)	0.553 (**1.5**)	0.793 (**5.5**)	0.556 (**3**)	0.477 (**3**)	-0.160 (**2.5**)	-0.343 (**2.5**)	-5.898 (**4.5**)	-7.268 (**6**)
UEFSMV	0.394 (**1.5**)	0.528 (**6**)	0.545 (**5**)	0.793 (**5.5**)	0.551 (**4**)	0.468 (**4.5**)	-0.172 (**4**)	-0.368 (**4**)	-5.863 (**4.5**)	-7.156 (**5**)
UEFNQC	0.393 (**3.5**)	0.564 (**3.5**)	0.547 (**3**)	0.793 (**3.5**)	0.557 (**2**)	0.479 (**2**)	-0.160 (**2.5**)	-0.344 (**2.5**)	-5.802 (**3**)	-7.133 (**3.5**)
qppbertpl	0.392 (**3.5**)	0.562 (**3.5**)	0.554 (**1.5**)	0.799 (**1**)	0.526 (**15**)	0.412 (**17**)	-0.271 (**10.5**)	-0.571 (**13**)	-5.673 (**2**)	-6.785 (**2**)
NQC	0.387 (**5.5**)	0.571 (**2**)	0.545 (**5**)	0.787 (**7**)	0.562 (**1**)	0.492 (**1**)	-0.151 (**1**)	-0.321 (**1**)	-6.530 (**7**)	-7.987 (**8**)
bertqpp	0.387 (**5.5**)	0.539 (**5**)	0.543 (**5**)	0.796 (**2**)	0.521 (**17**)	0.402 (**20**)	-0.292 (**14**)	-0.611 (**15**)	-5.983 (**6**)	-7.095 (**3.5**)
deepqpp	0.379 (**7**)	0.513 (**7.5**)	0.530 (**8**)	0.794 (**3.5**)	0.519 (**18**)	0.399 (**21**)	-0.309 (**16**)	-0.644 (**16.5**)	-5.578 (**1**)	-6.538 (**1**)
neuralqpp	0.369 (**8**)	0.322 (**17**)	0.536 (**7**)	0.779 (**9**)	0.543 (**7**)	0.455 (**9**)	-0.244 (**8.5**)	-0.497 (**8.5**)	-7.286 (**10.5**)	-8.480 (**12**)
UEFWIG	0.362 (**9**)	0.513 (**7.5**)	0.510 (**9**)	0.781 (**8**)	0.545 (**6**)	0.460 (**6.5**)	-0.220 (**5**)	-0.453 (**5**)	-6.754 (**8**)	-7.875 (**7**)
UEFClarity	0.352 (**10**)	0.470 (**10**)	0.497 (**10**)	0.777 (**10**)	0.539 (**9**)	0.449 (**10**)	-0.246 (**8.5**)	-0.499 (**8.5**)	-7.106 (**9**)	-8.189 (**10**)
Clarity	0.344 (**11**)	0.446 (**11**)	0.489 (**11**)	0.776 (**11**)	0.542 (**8**)	0.456 (**8**)	-0.234 (**7**)	-0.476 (**7**)	-7.488 (**13**)	-8.625 (**14**)
WIG	0.341 (**12**)	0.482 (**9**)	0.482 (**12**)	0.773 (**12**)	0.547 (**5**)	0.468 (**4.5**)	-0.229 (**6**)	-0.461 (**6**)	-7.241 (**10.5**)	-8.228 (**10**)
VARavg	0.320 (**13**)	0.262 (**18**)	0.446 (**13**)	0.770 (**13**)	0.531 (**12**)	0.436 (**15**)	-0.277 (**12**)	-0.556 (**11.5**)	-7.775 (**14.5**)	-8.798 (**15**)
IDFmax	0.290 (**14**)	0.420 (**13**)	0.400 (**14**)	0.765 (**14**)	0.535 (**11**)	0.447 (**12**)	-0.266 (**10.5**)	-0.529 (**10**)	-7.345 (**12**)	-8.182 (**10**)
ICTFmax	0.266 (**15**)	0.395 (**15.5**)	0.368 (**15.5**)	0.756 (**15**)	0.529 (**13.5**)	0.440 (**13**)	-0.299 (**15**)	-0.586 (**14**)	-7.817 (**14.5**)	-8.557 (**13**)
IDFavg	0.258 (**16**)	0.425 (**12**)	0.367 (**15.5**)	0.748 (**16**)	0.537 (**10**)	0.460 (**6.5**)	-0.287 (**13**)	-0.553 (**11.5**)	-9.088 (**18**)	-9.788 (**18**)
ICTFavg	0.232 (**17**)	0.407 (**14**)	0.331 (**19**)	0.738 (**19**)	0.529 (**13.5**)	0.448 (**11**)	-0.339 (**17**)	-0.647 (**16.5**)	-10.125 (**20**)	-10.771 (**20**)
VARmax	0.227 (**18.5**)	0.246 (**19.5**)	0.335 (**17.5**)	0.739 (**17.5**)	0.513 (**19.5**)	0.412 (**17**)	-0.405 (**19.5**)	-0.781 (**19.5**)	-8.870 (**16.5**)	-9.539 (**16.5**)
SCQmax	0.227 (**18.5**)	0.246 (**19.5**)	0.335 (**17.5**)	0.739 (**17.5**)	0.513 (**19.5**)	0.412 (**17**)	-0.405 (**19.5**)	-0.781 (**19.5**)	-8.870 (**16.5**)	-9.539 (**16.5**)
SCSsum	0.220 (**20**)	0.397 (**15.5**)	0.317 (**20**)	0.733 (**20**)	0.523 (**16**)	0.438 (**14**)	-0.372 (**18**)	-0.708 (**18**)	-10.231 (**21**)	-10.836 (**21**)
SCQavg	0.169 (**21**)	0.222 (**21**)	0.252 (**21**)	0.721 (**21**)	0.505 (**21**)	0.406 (**19**)	-0.461 (**21**)	-0.874 (**21**)	-9.809 (**19**)	-10.303 (**19**)

Table 4: Robust'04 Dataset - Evaluation of QPP methods using correlation and risk-sensitive metrics. The QPP methods identified as the most effective by all correlation metrics do not align with the results from the risk-sensitive evaluation.

QPP's	Kendall's τ	Pearson's r	Spearman's ρ	sMARE	$GeoRisk$ ($\alpha=5$)	$GeoRisk$ ($\alpha=10$)	U_{RISK} ($\alpha=5$)	U_{RISK} ($\alpha=10$)	T_{RISK} ($\alpha=5$)	T_{RISK} ($\alpha=10$)
NQC	0.455 (**1**)	0.582 (**2.5**)	0.594 (**1**)	0.814 (**1**)	0.555 (**1**)	0.466 (**1**)	-0.110 (**1**)	-0.280 (**1**)	-1.489 (**1**)	-2.249 (**1**)
neuralqpp	0.413 (**2**)	0.599 (**1**)	0.561 (**3**)	0.792 (**3**)	0.522 (**9**)	0.405 (**12**)	-0.253 (**6**)	-0.544 (**8.5**)	-2.725 (**3.5**)	-3.374 (**3.5**)
bertqpp	0.392 (**3.5**)	0.539 (**4.5**)	0.525 (**4**)	0.791 (**3**)	0.516 (**13**)	0.395 (**13**)	-0.279 (**10**)	-0.594 (**10.5**)	-2.602 (**2**)	-3.173 (**2**)
qppbertpl	0.389 (**3.5**)	0.589 (**2.5**)	0.569 (**2**)	0.791 (**3**)	0.515 (**14**)	0.392 (**14.5**)	-0.273 (**10**)	-0.583 (**10.5**)	-3.074 (**8.5**)	-3.800 (**12**)
VARavg	0.357 (**5**)	0.408 (**10**)	0.491 (**5**)	0.777 (**6**)	0.541 (**3.5**)	0.453 (**4**)	-0.194 (**3**)	-0.411 (**3**)	-3.064 (**8.5**)	-3.725 (**10.5**)
IDFavg	0.346 (**6**)	0.403 (**11**)	0.469 (**7**)	0.779 (**5**)	0.543 (**2**)	0.456 (**2**)	-0.189 (**2**)	-0.402 (**2**)	-2.740 (**3.5**)	-3.340 (**3.5**)
ICTFavg	0.337 (**7.5**)	0.395 (**12**)	0.458 (**8**)	0.775 (**7.5**)	0.541 (**3.5**)	0.455 (**3**)	-0.199 (**4**)	-0.418 (**4**)	-2.821 (**5**)	-3.389 (**5**)
deepqpp	0.336 (**7.5**)	0.533 (**4.5**)	0.482 (**6**)	0.769 (**9.5**)	0.518 (**11**)	0.410 (**11**)	-0.311 (**13**)	-0.633 (**13**)	-3.135 (**10.5**)	-3.618 (**8.5**)
SMV	0.332 (**9**)	0.455 (**6**)	0.449 (**9**)	0.775 (**7.5**)	0.525 (**7.5**)	0.421 (**8.5**)	-0.254 (**6**)	-0.527 (**6**)	-2.898 (**6**)	-3.427 (**6.5**)
IDFmax	0.312 (**10.5**)	0.440 (**7.5**)	0.429 (**10.5**)	0.769 (**9.5**)	0.527 (**5.5**)	0.429 (**5.5**)	-0.258 (**6**)	-0.529 (**6**)	-2.965 (**7**)	-3.439 (**6.5**)
VARmax	0.309 (**10.5**)	0.443 (**7.5**)	0.431 (**10.5**)	0.762 (**13**)	0.520 (**11**)	0.416 (**10**)	-0.302 (**12**)	-0.610 (**12**)	-3.392 (**13**)	-3.881 (**13.5**)
SCSsum	0.302 (**12.5**)	0.366 (**14**)	0.404 (**12.5**)	0.765 (**11.5**)	0.526 (**5.5**)	0.427 (**5.5**)	-0.261 (**8**)	-0.534 (**6**)	-3.128 (**10.5**)	-3.641 (**8.5**)
ICTFmax	0.298 (**12.5**)	0.435 (**9**)	0.412 (**12.5**)	0.764 (**11.5**)	0.524 (**7.5**)	0.426 (**7**)	-0.276 (**10**)	-0.559 (**8.5**)	-3.232 (**12**)	-3.703 (**10.5**)
WIG	0.279 (**14**)	0.386 (**13**)	0.390 (**14**)	0.748 (**14**)	0.518 (**11**)	0.419 (**8.5**)	-0.334 (**14**)	-0.659 (**14**)	-3.732 (**14.5**)	-4.170 (**15.5**)
UEFSMV	0.227 (**15**)	0.332 (**15**)	0.317 (**15**)	0.741 (**15**)	0.502 (**15**)	0.390 (**14.5**)	-0.394 (**15**)	-0.774 (**15**)	-3.787 (**16**)	-4.181 (**15.5**)
UEFNQC	0.208 (**16**)	0.304 (**16.5**)	0.288 (**16**)	0.730 (**16**)	0.496 (**16**)	0.382 (**16**)	-0.442 (**16**)	-0.860 (**16**)	-4.185 (**17**)	-4.575 (**17.5**)
UEFClarity	0.179 (**17**)	0.222 (**19**)	0.247 (**17.5**)	0.720 (**17.5**)	0.480 (**18.5**)	0.353 (**19**)	-0.522 (**19**)	-1.008 (**19**)	-4.439 (**19**)	-4.806 (**19**)
UEFWIG	0.171 (**18.5**)	0.239 (**18**)	0.229 (**19**)	0.716 (**20**)	0.474 (**20**)	0.344 (**20**)	-0.553 (**20**)	-1.067 (**20**)	-4.592 (**20**)	-4.968 (**20**)
Clarity	0.170 (**18.5**)	0.305 (**16.5**)	0.259 (**17.5**)	0.722 (**17.5**)	0.487 (**17**)	0.370 (**17**)	-0.495 (**17.5**)	-0.953 (**17.5**)	-4.306 (**18**)	-4.625 (**17.5**)
SCQavg	0.130 (**20**)	0.164 (**20**)	0.183 (**20**)	0.719 (**19**)	0.481 (**18.5**)	0.358 (**18**)	-0.493 (**17.5**)	-0.949 (**17.5**)	-3.665 (**14.5**)	-3.922 (**13.5**)
SCQmax	-0.043 (**21**)	-0.091 (**21**)	-0.066 (**21**)	0.651 (**21**)	0.441 (**21**)	0.309 (**21**)	-0.826 (**21**)	-1.543 (**21**)	-5.562 (**21**)	-5.746 (**21**)

Table 5: Deep Learning'20 Dataset - Evaluation of QPP methods using correlation and risk-sensitive metrics. The difference between traditional QPP evaluation and risk-sensitive metrics becomes more evident from the 2nd rank positions.

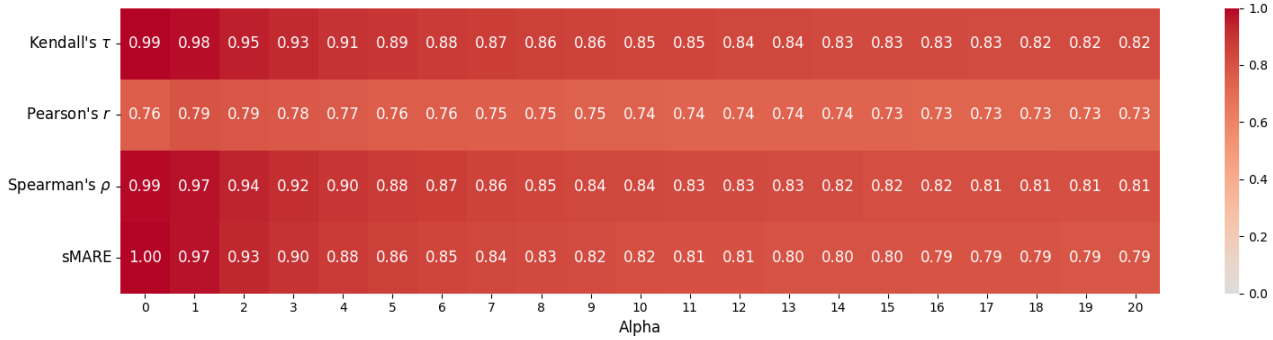
QPP's	Kendall's τ	Pearson's r	Spearman's ρ	sMARE	$GeoRisk$ ($\alpha=5$)	$GeoRisk$ ($\alpha=10$)	U_{RISK} ($\alpha=5$)	U_{RISK} ($\alpha=10$)	T_{RISK} ($\alpha=5$)	T_{RISK} ($\alpha=10$)
NQC	0.530 (**1**)	0.672 (**1**)	0.715 (**1**)	0.827 (**1**)	0.581 (**1**)	0.515 (**1**)	0.001 (**1**)	-0.073 (**1**)	0.151 (**1**)	-1.170 (**1**)
UEFNQC	0.502 (**2**)	0.607 (**2.5**)	0.664 (**2**)	0.825 (**2**)	0.570 (**2**)	0.493 (**2**)	-0.037 (**2**)	-0.149 (**2**)	-0.511 (**2**)	-1.470 (**2**)
bertqpp	0.435 (**3**)	0.563 (**4.5**)	0.609 (**3**)	0.796 (**3.5**)	0.546 (**3.5**)	0.453 (**4**)	-0.158 (**3.5**)	-0.361 (**3.5**)	-1.977 (**3.5**)	-2.664 (**3**)
UEFSMV	0.413 (**4.5**)	0.423 (**12.5**)	0.563 (**6**)	0.797 (**3.5**)	0.539 (**5**)	0.440 (**8.5**)	-0.156 (**3.5**)	-0.361 (**3.5**)	-2.036 (**3.5**)	-2.765 (**4**)
qppbertpl	0.405 (**4.5**)	0.614 (**2.5**)	0.578 (**4**)	0.791 (**5.5**)	0.534 (**8**)	0.432 (**11**)	-0.201 (**6.5**)	-0.441 (**7.5**)	-2.245 (**5.5**)	-2.876 (**5.5**)
deepqpp	0.400 (**6.5**)	0.558 (**4.5**)	0.560 (**6**)	0.791 (**5.5**)	0.523 (**12.5**)	0.407 (**17**)	-0.243 (**10**)	-0.529 (**10**)	-2.290 (**5.5**)	-2.873 (**5.5**)
neuralqpp	0.397 (**6.5**)	0.504 (**7**)	0.561 (**6**)	0.784 (**7**)	0.536 (**7**)	0.439 (**8.5**)	-0.205 (**6.5**)	-0.441 (**7.5**)	-2.486 (**7**)	-3.130 (**7.5**)
WIG	0.373 (**9**)	0.459 (**8**)	0.529 (**8**)	0.769 (**10**)	0.544 (**3.5**)	0.464 (**3**)	-0.205 (**6.5**)	-0.424 (**5.5**)	-2.724 (**10**)	-3.229 (**9.5**)
UEFWIG	0.369 (**9**)	0.539 (**6**)	0.513 (**9.5**)	0.780 (**8**)	0.538 (**6**)	0.445 (**7**)	-0.203 (**6.5**)	-0.434 (**5.5**)	-2.579 (**8.5**)	-3.198 (**9.5**)
SMV	0.366 (**9**)	0.434 (**10**)	0.511 (**9.5**)	0.777 (**9**)	0.533 (**9**)	0.435 (**10**)	-0.220 (**9**)	-0.468 (**9**)	-2.593 (**8.5**)	-3.187 (**7.5**)
UEFClarity	0.322 (**11**)	0.432 (**10**)	0.451 (**11**)	0.765 (**11**)	0.522 (**12.5**)	0.419 (**15.5**)	-0.276 (**11**)	-0.565 (**11**)	-2.925 (**11**)	-3.424 (**11**)
IDFavg	0.276 (**12**)	0.372 (**15**)	0.379 (**14.5**)	0.738 (**16**)	0.530 (**11**)	0.448 (**6**)	-0.300 (**12.5**)	-0.585 (**12.5**)	-4.132 (**17**)	-4.573 (**17**)
ICTFavg	0.274 (**13**)	0.381 (**14**)	0.384 (**12.5**)	0.738 (**14.5**)	0.531 (**10**)	0.450 (**5**)	-0.300 (**12.5**)	-0.586 (**12.5**)	-4.244 (**18**)	-4.704 (**18**)
IDFmax	0.266 (**14.5**)	0.429 (**12.5**)	0.386 (**12.5**)	0.746 (**12**)	0.520 (**14**)	0.424 (**13**)	-0.315 (**14**)	-0.624 (**14**)	-3.323 (**12**)	-3.715 (**12**)
VARavg	0.262 (**14.5**)	0.316 (**17**)	0.360 (**16**)	0.740 (**14.5**)	0.518 (**15**)	0.424 (**13**)	-0.325 (**15**)	-0.639 (**15**)	-3.426 (**13**)	-3.789 (**13**)
ICTFmax	0.251 (**16**)	0.430 (**10**)	0.373 (**14.5**)	0.743 (**13**)	0.517 (**16.5**)	0.419 (**15.5**)	-0.334 (**16**)	-0.658 (**16**)	-3.506 (**14**)	-3.894 (**14**)
SCSsum	0.236 (**17**)	0.355 (**16**)	0.329 (**17**)	0.732 (**17**)	0.517 (**16.5**)	0.423 (**13**)	-0.349 (**17**)	-0.680 (**17**)	-3.790 (**15**)	-4.173 (**15**)
Clarity	0.171 (**18**)	0.217 (**18**)	0.248 (**18**)	0.717 (**18**)	0.497 (**18**)	0.392 (**18**)	-0.446 (**18**)	-0.857 (**18**)	-4.021 (**16**)	-4.322 (**16**)
SCQavg	0.006 (**19**)	-0.093 (**19**)	0.007 (**19**)	0.663 (**19**)	0.460 (**19**)	0.342 (**19**)	-0.709 (**19**)	-1.327 (**19**)	-5.126 (**19**)	-5.313 (**19**)
SCQmax	-0.159 (**20.5**)	-0.255 (**20.5**)	-0.234 (**20.5**)	0.632 (**20.5**)	0.407 (**20.5**)	0.254 (**20.5**)	-0.928 (**20.5**)	-1.737 (**20.5**)	-5.222 (**20.5**)	-5.417 (**20.5**)
VARmax	-0.159 (**20.5**)	-0.255 (**20.5**)	-0.234 (**20.5**)	0.632 (**20.5**)	0.407 (**20.5**)	0.254 (**20.5**)	-0.928 (**20.5**)	-1.737 (**20.5**)	-5.222 (**20.5**)	-5.417 (**20.5**)

Table 6: Deep Learning'19 Dataset - Evaluation of QPP methods using correlation and risk-sensitive metrics. The difference between traditional QPP evaluation and risk-sensitive metrics becomes more evident from the 3rd to the 5th rank positions.

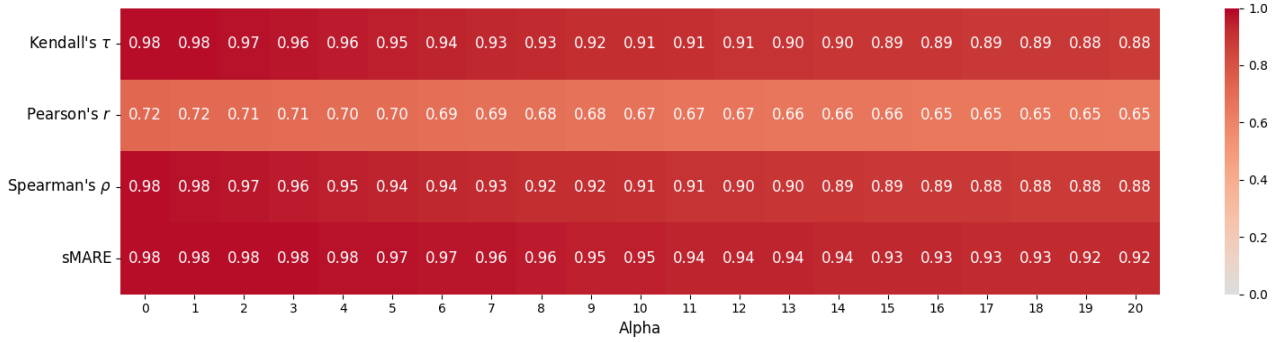
QPP Model	Description
Pre-retrieval	
SCQ [47]	Measures similarity based on *cf.idf* (collection frequency–inverse document frequency) to the corpus, summed over the query terms.
SCQavg	SCQ normalized by query length.
SCQmax	The query term with the maximal SCQ score.
SumVAR	Measures the variability of *cf.idf* values of the query terms in the corpus.
AvgVAR	Variability of *cf.idf* values normalized by query length.
MaxVAR	The query term with the maximal *cf.idf* variability.
IDFavg [8]	Mean Inverse Document Frequency (IDF) of the query terms.
IDFmax [33]	The query term with the maximal IDF value.
ICTFavg	Mean Inverse Collection Term Frequency (ICTF) of the query terms.
ICTFmax	The query term with the maximal ICTF value.
WIG [48]	Weighted Information Gain: Measures divergence between the query language model and the corpus language model.
Post-retrieval	
Clarity [7]	KL-divergence between the language model of the top retrieved documents and the corpus language model.
NQC [36]	Normalized Query Commitment: Standard deviation of the retrieval scores of the top documents.
SMV [39]	Score Magnitude and Variance: Combines the mean and variance of the top document scores.
UEF [35]	Uncertainty Estimation Framework: Compares the initial result list with a re-ranked list using a Relevance Model (RM), scaled by an estimator of RM quality.
UEFClarity	UEF-scaled Clarity score.
UEFNQC	UEF-scaled NQC score.
UEFWIG	UEF-scaled WIG score.
UEFSMV	UEF-scaled SMV score.
VARavg	Average variance of the query terms' *cf.idf* values.
VARmax	Maximal variance of the query terms' *cf.idf* values.
SCSsum [7]	Sum of Score Contributions: Sum of the retrieval scores of the top documents.
Deep Learning	
neuralqpp [45]	Neural network model trained on query-document interaction signals (e.g., term distributions, embeddings).
bertqpp [1]	BERT-based model for semantic alignment between queries and top retrieved documents.
qppbertpl	BERT-QPP enhanced with pseudo-labels for robustness in low-resource scenarios.
deepqpp [9]	Deep pairwise interaction model capturing query-document relevance patterns.

Table 7: Summary of QPP Models - Comprehensive list of pre-retrieval, post-retrieval, and deep learning methods with detailed descriptions.

(a) $GeoRisk$ **correlations**



(b) U_{RISK} **correlations**



(c) T_{RISK} **correlations**

Figure 6: Kendall's τ correlation heatmaps for $GeoRisk$, U_{RISK} and T_{RISK} with alpha values ranging from 0 to 20, and the metrics Kendall's τ, Pearson's r, Spearman's ρ, and sMARE, using BM25 on the Robust'04 dataset. As we increase alpha, the correlation with other QPP metrics decreases.