Evaluating Metric Sensitivity to Offline–Online Alignment in Information Retrieval

Satsuki Udagawa[1]

Supervisor(s): Avishek Anand[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfillment of the Requirements
For the Bachelor of Computer Science and Engineering
January 24, 2026

Name of the student: Satsuki Udagawa
Final project course: CSE3000 Research Project
Thesis committee: Avishek Anand, Julian Urbano

Abstract

This study examines how effectively widely used offline information retrieval (IR) metrics reflect changes in online performance. As offline evaluation plays a central role in model development, understanding its alignment with user-oriented signals is essential. Using 52 diverse ranking pipelines and approximately 2,000 queries from the MS MARCO DL19 and DL20 benchmarks, we analyze the sensitivity of five offline metrics: Precision@10, Recall@10, MAP, MRR, and NDCG@10, to five simulated online metrics: CTR, SSR, ZRR, ADT, and SAR. Sensitivity is quantified through slope-based analysis, and alignment is assessed using the Pearson correlation coefficient. Our results show that NDCG@10 and Recall@10 are the most sensitive offline metrics across multiple online behaviors, while Precision@10 consistently exhibits low sensitivity. Furthermore, we demonstrate that sensitivity and alignment capture complementary aspects of offline–online relationships: some metric pairs show strong responsiveness but weak linear consistency. Overall, this study provides a detailed and reproducible evaluation of how offline metrics behave in relation to simulated online performance, offering practical guidance for selecting offline metrics that better reflect user-centric outcomes.

Keywords: Information retrieval, offline evaluation, online evaluation, alignment, metric, sensitivity

1 Introduction

The goal of information retrieval (IR)[1] systems is to find the most relevant results to a search query. Relevance, in this context, refers to how well a search result matches the user's information needs. Relevance can be measured by, for example, how often users click on a specific result or how long they stay on the page shown as the result. Depending on the perspective, relevance may also be inferred from ad clicks on pages where ads are displayed, or from purchase behaviour on shopping platforms, where frequent clicks without purchases may not hold much value.

IR systems are typically evaluated in two ways: offline and online. Offline evaluation tests a system using pre-labelled datasets without involving real users. These labels specify which documents are relevant to each query, allowing the system's output to be compared against known judgements. Online evaluation[2], on the other hand, measures performance in real time with real users, often through A/B testing[3], by observing behaviours such as clicks or other engagement signals. Because offline testing is cheaper and faster, it is widely used during development, but its usefulness depends on how well it predicts real user behaviour.

For offline evaluation to be reliable, its results must align with what happens when real users interact with the system. Strong alignment means that improvements observed during offline evaluation correspond to improvements experienced by real users. When this alignment is weak, a system may appear promising offline but fail to improve, or even degrade, user experience when deployed. Ensuring good alignment is essential because offline evaluation often guides model selection and is frequently used without seeing how the system performs with real users.

To evaluate the performance of IR systems, offline evaluation relies on metrics that quantify ranking quality. A metric is considered *sensitive* if it reflects changes in offline–online alignment. When the relation between offline and online performance changes, the metric's scores should change accordingly. Different metrics can lead to different conclusions about system quality, but a sensitive metric will show clear differences between ranking systems when those systems perform differently with real users, while an insensitive metric may fail to reflect these differences. If an offline metric does not correspond to what users actually value in practice, it may give a misleading signal during development. Understanding which metrics best reflect real user behaviour is therefore essential for making the best choice of offline metric to evaluate ranking systems with.

In this study, we aim to answer the research question: How sensitive are commonly used offline information retrieval evaluation metrics to changes in offline-online alignment? To help us answer this question, we consider the following sub-questions to guide us in our research:

- How do different offline evaluation metrics (Precision@k, Recall@k, MAP, nDCG@k, and MRR) vary in their sensitivity to differences in online performance signals?
- Do offline evaluation metrics respond consistently to changes in offline–online alignment across different ranking models?

Answering these questions provides insight into which offline metrics most reliably reflect online performance.

The structure of the rest of this paper is as follows: Section 2 evaluates the prior work and explains the research gap. Section 4 describes the methodology used to evaluate metric sensitivity. Section 3 outlines the contributions of this study. Section 5 presents the experimental setup. Section 6 reports the results of the sensitivity and correlation analyses. Section 8 discusses responsible research considerations. Section 7 provides a broader discussion of the findings. Section 9 concludes the thesis and highlights directions for future work.

2 Related Work

A substantial amount of research has examined how IR systems should be evaluated and to what extent offline metrics reflect real user experience. While prior work such as Kutlu et al.[4] has examined how IR metrics correlate with one another and how metric choice affects system ranking, these studies do not address how sensitive metrics are to changes in offline–online alignment.

Early research had already raised doubts about whether conventional relevance-based metrics adequately reflect user behaviour and expectations in realistic, interactive settings. Su (1992)[5] showed that commonly used metrics such as precision (Precision@k) are only weakly correlated with real users' perceptions of success. Broader notions, such as the overall value of search results, better explain user satisfaction. This highlights that metrics optimized in offline settings may fail to capture what users truly consider successful, especially in online environments where real user behavior is involved.

More recent work has investigated the ability of offline evaluation metrics to predict online performance. Meta-evaluation research on the performance of offline metrics (Chen et al., 2017)[6] showed that offline metrics often disagree on system rankings and that small changes in relevance assumptions or user models can substantially alter conclusions. Together, these findings highlight the importance of offline–online alignment and show that the choice of offline metrics strongly influences how system quality is measured.

Despite these contributions, previous studies have primarily examined how offline metrics correlate with online performance. Less attention has been given to the question of metric *sensitivity*, that is, how much a metric changes in score when offline–online alignment improves or deteriorates. This distinction is important because offline metrics are often used to guide model selection, and insensitive metrics may fail to capture meaningful differences between systems, potentially leading to misleading development decisions.

This study addresses this gap by investigating the sensitivity of offline metrics with respect to offline–online alignment. By identifying which metrics respond most reliably to changes in alignment, we provide clearer guidance for metric selection and improve the reliability of offline evaluation in predicting real user outcomes.

3 Contributions

This study makes several contributions to the evaluation of information retrieval systems.

First, we investigate how strongly five commonly used offline metrics respond to changes in offline–online alignment. While prior work has focused primarily on correlations between offline and online metrics, sensitivity has received little attention. Our analysis fills this gap by quantifying how offline metric scores change in relation to multiple online performance signals.

Secondly, using 52 ranking systems evaluated on the MS MARCO dataset, we examine sensitivity across a broad set of offline and online metrics. This large-scale comparison provides a detailed view of how different metrics behave under varying conditions and offers practical insight into their robustness.

Finally, by combining slope-based sensitivity analysis with Pearson correlation, we show that high sensitivity does not necessarily imply strong offline-online alignment. Several metric pairs exhibit steep slopes but weak correlations, indicating that offline metrics may react strongly to changes in online performance without being consistent in detecting relative changes. The results of this analysis reveal the limitations of relying solely on sensitivity when selecting offline metrics.

4 Methodology

Our study aims to investigate the sensitivity of offline metrics with respect to offline–online alignment in information retrieval. To achieve this, we adopt a systematic approach that allows us to analyze different system behaviors and observe how metrics respond to changes in performance.

We first use several ranking models, including BM25, DFIC, DirichletLM, PL2, and TF-IDF, to rank the set of queries in our dataset. The inclusion of multiple rankers allows us to evaluate each offline metric on several retrieval systems separately. This reflects common practice in information retrieval, where offline metrics are used to compare the performance of different systems to determine which one performs best. Using multiple rankers also ensures a sufficient number of data points for meaningful analysis.

For each ranker, we compute a score using the offline evaluation metrics Precision@k, Recall@k, MAP, nDCG@k, and MRR. To estimate online performance, we simulate user behavior through click-through rate (CTR), average dwell time (ADT), session abandonment rate (SAR), session success rate (SSR), and zero result rate (ZRR). The calculated values are used directly in our analysis.

Finally, we analyze the relationship between the offline and online scores by plotting the scores of each offline metric against the corresponding online measure. The slope of the least-squares fit to the resulting scatter plots will tell us how much the offline metric's scores change in relation to differences in online metric scores. In this experiment, we consider an offline metric with a steeper slope to be more sensitive to the corresponding online measure. To account for both positive and negative relationships, we take the absolute value of the slope, where a larger absolute value corresponds to a steeper slope and thus indicates greater sensitivity.

In addition to this slope-based analysis, we compute the correlation coefficient[7] for each offline–online metric pair to determine how strong the linear association is between each pair. In contrast to the slope-based analysis, correlation reflects consistency across systems, not magnitude. This allows us to evaluate to how consistently offline metrics track the behaviour of online metrics, supporting the reliability of our slope-based sensitivity analysis.

In the textbook *Statistics*[7], the correlation coefficient r is defined as:

$$r = \frac{\operatorname{cov}(x, y)}{\mathrm{SD}(x)\,\mathrm{SD}(y)},$$

where x and y are the two variables on which the correlation coefficient is calculated on, and $\operatorname{cov}(x, y)$ and $\mathrm{SD}(X)$ are the covariance of the two variables and the standard deviation of variable X.

This formula is mathematically equivalent to the definition of the Pearson's r correlation coefficient implemented by the `pearsonr` function in the SciPy library. According to the SciPy documentation, Pearson's r is computed as:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2}\,\sqrt{\sum(y - m_y)^2}},$$

where m_x and m_y are the means of x and y, respectively.The documentation for SciPy's Pearson correlation function is available in the online docs[1].

This methodology directly addresses the research questions introduced in Section 1. By computing offline metric scores for a wide range of ranking systems and comparing them against simulated online performance signals, we can quantify how strongly each offline metric responds to changes in online behaviour, thereby answering Subquestion 1. The slope-based sensitivity analysis captures the magnitude of these responses, while the correlation analysis evaluates their consistency across systems, jointly addressing Subquestion 2. Together, these analyses allow us to determine which offline metrics most reliably reflect online performance and thus provide a complete answer to the main research question.

4.1 Offline Metrics

Precision@k[1] measures the fraction of relevant documents among the top k retrieved items for a given query. This metric measures how well a ranking system places relevant results near the top of the list, and it is computed as:

$$\text{Precision@k} = \frac{\text{Number of relevant documents in top } k}{k}.$$

Recall@k[1] measures the fraction of all relevant documents that are retrieved in the top k results. This is similar to Pricision@k, but is also different in the sense that, while Precision@k divides by the number k, Recall@k divides by the total number of relevant documents for the given query. Recall@k is defined as:

$$\text{Recall@k} = \frac{\text{Number of relevant documents in top } k}{\text{Total number of relevant documents}}.$$

MAP (Mean Average Precision)[1] is defined as the mean of the average precision values computed over all queries. For a single query, the average precision is obtained by averaging the precision scores at all ranks where relevant documents occur. Formally, for a query q with R_q relevant documents, the average precision is given by:

$$\text{AveragePrecision}(q) = \frac{1}{R_q} \sum_{k=1}^{n} P(k) \cdot \text{rel}(k),$$

where n denotes the total number of retrieved documents, $P(k)$ is the precision at cutoff k, and $\text{rel}(k)$ is an indicator function that equals 1 if the document at rank k is relevant and 0 otherwise. The MAP score is then defined as:

$$\text{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AveragePrecision}(q),$$

where Q is the number of queries in the evaluation set.

MRR (Mean Reciprocal Rank)[8] focuses on the rank of the first relevant document for each query. For a single query, the reciprocal rank is the inverse of the rank position of the first relevant result. MRR is then the average of these reciprocal ranks over all queries. This metric emphasizes how quickly a system retrieves the first relevant result:

$$\text{MRR} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{\text{Rank of first relevant document for } q}.$$

nDCG@k (Normalized Discounted Cumulative Gain at k)[9] measures ranking quality by taking into account graded relevance and the position of each relevant document in the top k results. Highly relevant documents appearing earlier in the ranking contribute more to the score. The DCG for a query is defined as:

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)},$$

where rel_i is the graded relevance of the document at rank i. The nDCG@k score is obtained by dividing DCG@k by the ideal DCG@k (IDCG@k), which represents the best possible ordering of results:

$$\text{nDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}}.$$

4.2 Online Metrics

Click-Through Rate (CTR)[10] measures the proportion of presented documents that receive a click. It reflects how often users interact with the retrieved results and is defined as:

$$\text{CTR} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} c_{ij}}{\sum_{i=1}^{N} M_i},$$

where N is the number of queries, M_i is the number of documents shown for query i, and $c_{ij} \in \{0, 1\}$ indicates whether document j was clicked.

Session Success Rate (SSR)[11] represents the proportion of sessions in which the user clicked at least one result. It is defined as:

$$\text{SSR} = \frac{\sum_{i=1}^{N} s_i}{N} = 1 - \text{SAR},$$

where $s_i = 1$ if $\sum_j c_{ij} > 0$. A higher SSR indicates that users frequently find at least one relevant document.

Zero Result Rate (ZRR)[12; 13] measures the proportion of sessions in which no clicks occur. It captures cases where the user does not click any document, whether due to dissatisfaction, lack of relevant results, or simulation fallback behavior. ZRR is defined as:

$$\text{ZRR} = \frac{\sum_{i=1}^{N} z_i}{N},$$

where $z_i = 1$ if query i resulted in zero clicks ($\sum_j c_{ij} = 0$). High ZRR values indicate that users frequently fail to find any document worth clicking.

Average Dwell Time (ADT)[14; 15] captures the average time users spend on clicked documents. Longer dwell times typically indicate higher user satisfaction. It is computed as:

$$\text{ADT} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} d_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{M_i} c_{ij}},$$

where d_{ij} denotes the dwell time for document j under query i. The denominator ensures that the average is taken only over clicked items.

[1] https://docs.scipy.org/doc/scipy-1.16.2/reference/generated/scipy.stats.pearsonr.html

Session Abandonment Rate (SAR)[11; 16] measures the fraction of sessions in which the user abandons the search without finding a satisfactory result. In our simulation, abandonment is explicitly indicated by the LLM through an `abandoned` flag[17]:

$$\mathrm{SAR} = \frac{\sum_{i=1}^{N} a_i}{N},$$

where $a_i = 1$ if the session for query i was marked as abandoned, and 0 otherwise.

5 Experimental Setup

This section describes the experimental setup used in our study. To ensure full reproducibility, the complete code and analysis notebooks are available on GitHub[2], including the Python notebook in the `metric-sensitivity` folder that contains all analysis steps.

For our experiments, we used the MS MARCO passage dataset[3], which contains approximately 300,000 real-world queries sampled from Bing search logs. From this dataset, we specifically used the queries from the TREC Deep Learning tracks DL19 and DL20, comprising roughly 2,000 queries with high-quality relevance judgments. These tracks are part of the TREC evaluation campaign and provide a curated subset of MS MARCO with more detailed and reliable annotations, making them well suited for comparing offline and online metrics in a controlled setting.

A diverse set of 52 ranking pipelines was used in this study. These systems were originally constructed using a combination of first-stage retrievers from PyTerrier [18] and re-ranking cascades submitted to TIRA/TIREx [19; 20]. The collection spans multiple retrieval paradigms, including lexical retrieval (via PyTerrier), dense bi-encoders from BEIR [21] and SentenceTransformers [22], multi-vector representations such as ColBERT [23], and learned sparse approaches such as Splade [24]. Using this broad set of pipelines ensures that our sensitivity analysis captures a wide range of ranking behaviours.

To compute both offline and online metrics, we follow a reproducible processing pipeline implemented in a Jupyter notebook. The input to our analysis consists of (1) TREC-formatted run files produced by the 52 ranking pipelines and (2) simulated click logs. The run files follow the standard six-column TREC format:

```
query-id  Q0  doc-id  rank  score  system-
name
```

These files are parsed to obtain the ranked lists for each query, from which all offline metrics are computed using the official MS MARCO relevance judgments for DL19 and DL20. The simulated click logs provide the click information required to derive the click-based online metrics CTR, SSR, and ZRR for each query–document pair. All intermediate data structures—parsed runs, relevance labels, and simulated click logs—are stored as Pandas dataframes, and the notebook produces a unified table containing all offline metrics and the available online metrics for each system. This explicit pipeline ensures that the analysis remains reproducible even if external tools or repositories become unavailable.

For each system variant and ranker combination, we computed offline evaluation metrics. Online performance was then simulated for each ranker using a click model to obtain the click-based metrics CTR, SSR, and ZRR. The remaining online metrics involving dwell time (ADT) and session abandonment (SAR) were obtained from the results of experiments by Ziliang Zhang (2026)[17], communicated personally. The corresponding online metric scores are included in the `metric-sensitivity` folder of the GitHub repository. The notebook outputs the final results as a CSV file containing one row per ranking pipeline and columns for all computed offline and online metrics, enabling a direct comparison between the two.

6 Results

In this section, we report the results of our experiments and analyze the sensitivity of offline metrics with respect to offline–online alignment in IR evaluation. We first present an example scatter plot comparing the scores of the offline metric NDCG@10 against the online metric CTR, after which we provide a table showing the performances of the offline metrics against the values computed by the online metrics. Finally, we summarize how these results relate to the research question and its subquestions, without yet interpreting *why* particular metrics behave as they do. That discussion is reserved for Section 7.

6.1 Metric Sensitivity Analysis

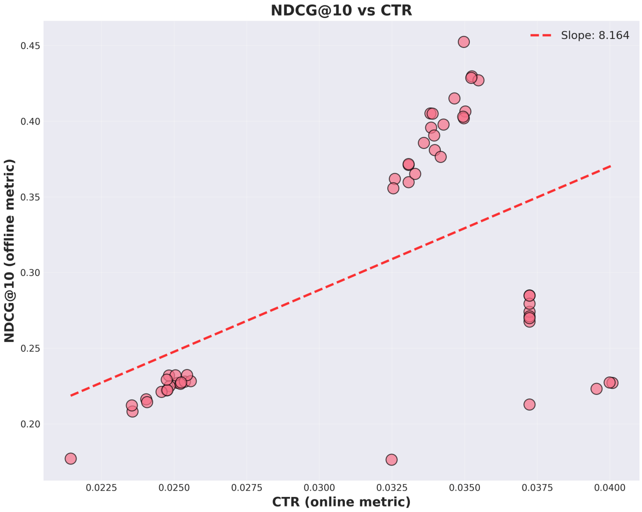

Figure 1: Example scatter plot comparing the offline metric NDCG@10 with the online metric CTR.

As shown in Figure 1, each ranking system is represented by a dot, with the x-axis corresponding to the online metric CTR and the y-axis to the offline metric NDCG@10. The red

[2] https://github.com/AinzOoalGown123/Metric-Sensitivity-Analysis

[3] https://microsoft.github.io/msmarco/

dotted line represents the least-squares fit of the points. The line has a positive slope of 8.164, indicating that, in general, NDCG@10 tends to increase as CTR increases. However, the points are spread across the plot rather than closely following the line: some systems cluster in the upper right and lower left regions, while others appear below the line in the lower right. This spread suggests that while NDCG@10 is generally responsive to online performance, it does not consistently capture the relative ordering of all systems. The spread of points around the line indicates that the connection between NDCG@10 and CTR is not perfectly consistent and this may suggest that the offline metric captures some, but not all, aspects of online performance.

Next, we extend this analysis by comparing all offline metrics against the online metrics, covering all combinations. The resulting scatter plots for each combination are provided in Appendix A, illustrating the relationships between offline and online performance measures for all the combinations considered in this study.

	CTR	**SSR**	**ZRR**	**ADT**	**SAR**
MAP	7.40	0.525	-0.525	-0.00109	0.117
MRR	7.48	0.529	-0.529	-0.00110	0.118
NDCG@10	8.16	0.595	-0.595	-0.00116	0.107
Precision@10	0.813	0.0886	-0.0886	-0.000108	0.00149
Recall@10	7.76	0.845	-0.845	-0.00103	0.0130

Table 1: Comparison of offline metrics against online metric scores. Values are rounded to 3 significant figures.

Table 1 shows the slope of each offline metric against each of the considered online metrics. A higher slope indicates that an offline metric is more sensitive to an online metric. However, one must take into account that the online metrics vary widely in scale: CTR values are below 0.05 for all systems, SAR, SSR, and ZRR values are roughly between 0 and 1, while ADT values can reach up to 120. Comparisons of slope steepness should therefore remain within the same column, where the offline metrics are each compared to the same online metric. the original scatter plots used to compute the slopes can be found in Appendix A.

For the online metric CTR, NDCG@10 is the most sensitive offline metric, with a slope of 8.16. MAP, MRR and Recall@10 have similar values for their slopes, ranging between 7 and 8. However, the offline metric Precision@10 has a slope that is much smaller, with a value of 0.813. This indicates that Precision@10 has low sensitivity to changes in CTR scores. It can also be noticed that this insensitivity is not limited to the online metric CTR: the slopes for Precision@10 are roughly an order of magnitude smaller than those of the other offline metrics across all online measures. These observations show that different offline metrics vary substantially in how strongly they respond to changes in online performance, directly contributing to answering Subquestion 1 about differences in sensitivity.

The slopes for the online metrics SSR and ZRR follow a notable trend. The slopes for SSR are the positive counterparts of the slopes for ZRR. For both metrics, Recall@10 is the most sensitive offline metric, with its absolute slope value being 0.845. MAP, MRR and NDCG@10 have similar values, ranging between 0.5 and 0.6, suggesting that these offline metrics are nearly equal in their sensitivity to SSR and ZRR. This mirrored behavior arises naturally from the formulas defining SSR and ZRR, which are mathematically related such that SSR = 1 - ZRR. As a result, the two online metrics can be considered essentially equivalent for the purpose of analyzing offline metric sensitivity.

The slopes of ADT show a different pattern. Except for Precision@10, whose slope is roughly one-tenth of the other metrics, the slopes have very close values, compressed in the narrow range between 0.00103 and 0.00116. This low magnitude results from the large scale of ADT scores. The scores computed by ADT reach up to 120, reducing the relative variance of the offline metrics. Despite this, NDCG@10 has the highest slope at 0.00116, which is approximately 5.45% higher than the second-highest slope, MRR at 0.00110. This demonstrates that NDCG@10 is the most sensitive offline metric for ADT. Interestingly, this relative difference is slightly larger than the difference observed for CTR, where the top two slopes differed by 5.15%.

Finally, for the online metric SAR, two offline metrics have very similar slopes: MAP and MRR. MAP has a slope of 0.117, while MRR has a slightly higher slope of 0.118, making MRR the most sensitive offline metric for SAR. When rounded to four significant figures, the difference is very small (MAP: 0.1170, MRR: 0.1175), which introduces some ambiguity as to which metric is actually more sensitive. For the purposes of this analysis, we will consider MRR as the most sensitive to SAR.

Based on the slope analysis, NDCG@10 and Recall@10 emerge as the offline metrics that are most frequently the most sensitive to changes in online metrics. Each of them is the most sensitive for two online metrics, although the two cases for Recall@10 correspond to SSR and ZRR, which are mathematically linked and effectively represent the same behaviour. MRR is the most sensitive for one online metric, while MAP and Precision@10 are never the most sensitive. Taken together, these results indicate that sensitivity to offline–online alignment is not uniform across offline metrics and that NDCG@10 and Recall@10 tend to be the most responsive in our setting, providing a partial answer to the main research question.

6.2 Correlation Analysis

	CTR	**SSR**	**ZRR**	**ADT**	**SAR**
MAP	0.521	0.846	-0.846	-0.296	0.169
MRR	0.523	0.845	-0.845	-0.297	0.168
NDCG@10	0.523	0.872	-0.872	-0.286	0.141
Precision@10	0.385	0.958	-0.958	-0.197	0.014
Recall@10	0.385	0.957	-0.957	-0.196	0.013

Table 2: Pearson correlation coefficients between offline metrics and online metrics across ranking systems.

Table 2 reports the Pearson correlation coefficients between each offline and online metric. Unlike the slope-based sensitivity analysis, which captures the magnitude of change of an offline metric in an online metric, Pearson correlation measures the strength of their linear relationship. A higher absolute correlation indicates that an offline metric more consistently follows trends in changes of online metric score, suggesting that there is good offline-online alignment between the metrics in question.

For the online metric CTR, the offline metrics MAP, MRR, and NDCG@10 display similar correlation values, with coefficients slightly above 0.52. Precision@10 and Recall@10 show noticeably lower correlations, both at 0.385, which indicates that these metrics are less effective at reflecting relative differences in CTR across systems. Although Recall@10 was relatively more sensitive to CTR in terms of slope than Precision@10, the identical correlation coefficients indicate that both offline metrics are similarly effective at consistently reflecting variations in CTR scores.

The correlation patterns for SSR and ZRR mirror each other, with equal magnitudes and opposite signs. This result is expected, given the strong mathematical relationship between the two online metrics. Precision@10 and Recall@10 show the strongest correlations with SSR and ZRR, with coefficients close to 0.96 in absolute values. Combined with the steep slopes from our previous results, this shows that these metrics both respond strongly to and consistently track changes in SSR and ZRR. The other offline metrics, MAP, MRR, and NDCG@10, also present high correlation coefficients, indicating that they capture the score changes in SSR and ZRR scores in a reasonably linear way.

For ADT, all offline metrics exhibit weak negative correlations, ranging approximately from -0.20 to -0.30. This suggests that none of the considered offline metrics reliably capture relative differences in ADT across ranking systems. Among them, NDCG@10 shows the weakest negative correlation, while MAP and MRR are slightly more negatively correlated. These findings contrast with the sensitivity analysis, where we found NDCG@10 to be the most sensitive offline metric to ADT, and thus illustrate that sensitivity and alignment can diverge.

Finally, correlations with SAR are generally low across all offline metrics. MAP and MRR exhibit the highest correlations at approximately 0.17, while Precision@10 and Recall@10 show almost no correlation with SAR. Figure 2 illustrates this pattern using Recall@10 as an example to represent this situation. The data points are widely scattered and symmetrically distributed around the regression line, which has a very shallow slope. This visual pattern confirms that Recall@10 does not consistently reflect variations in SAR scores. A similar distribution is observed for Precision@10, reinforcing the conclusion that SAR captures aspects of user behavior that are largely unrepresented by these offline metrics.

The radar charts in Figure 3 summarize the relative behaviour of the offline metrics across all online metrics by jointly visualizing normalized slopes and absolute correlations. This representation highlights differences that are less apparent in the tables: for example, metrics such as Recall@10 show high sensitivity to SSR and ZRR but comparatively weaker alignment for CTR and SAR. Conversely, NDCG@10 maintains relatively strong sensitivity across most online metrics, even when its correlation is moderate. These charts therefore provide an intuitive overview of how sensitivity and alignment jointly characterize the relationship between offline and online metrics.

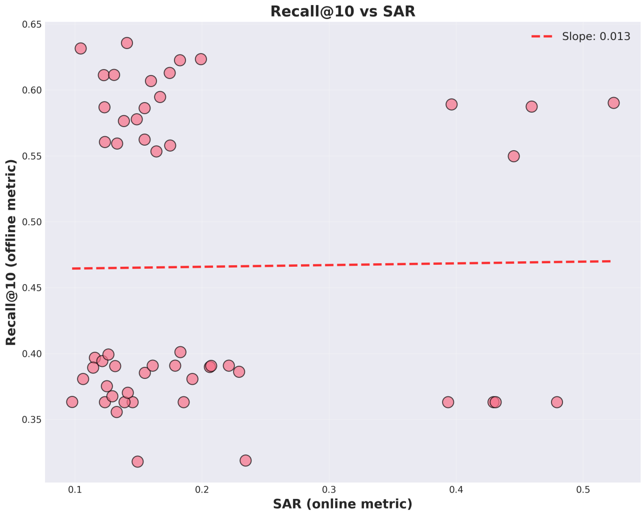

Figure 2: Example scatter plot comparing the offline metric Recall@10 with the online metric SAR.

Overall, the correlation analysis complements the sensitivity results by highlighting that high sensitivity does not necessarily imply strong alignment in system ranking. The scores of offline metrics may respond strongly to changes in certain online metrics, yet still fail to consistently reflect the relative differences in their online counterparts. These correlations reveal how consistently offline metrics track online behaviour, addressing Subquestion 2 and, together with the slope-based analysis provide a complete empirical basis for answering the research question.

7 Discussion

In this section, we discuss how to interpret the results obtained through our experiment, while considering possible limitations of our findings. We then relate these results to prior work to place them within the broader landscape of IR research.

7.1 Interpretation of Results

Our results show that high sensitivity of an offline metric to an online metric does not necessarily imply strong linear alignment, as several metric pairs exhibit steep slopes but weak linear correlations. This confirms that sensitivity and alignment capture fundamentally different properties: sensitivity reflects how strongly an offline metric responds to changes in online performance, whereas alignment reflects how consistently it preserves the relative ordering of systems. When metrics with steep slopes have weak correlations, the offline metric reacts to changes in online behaviour but is inconsistent in doing so across ranking systems. This distinction is essential for interpreting our findings and answering the main research question.

The only offline–online metric pairs for which we can be confident that the relationship is close to linear are Pre-

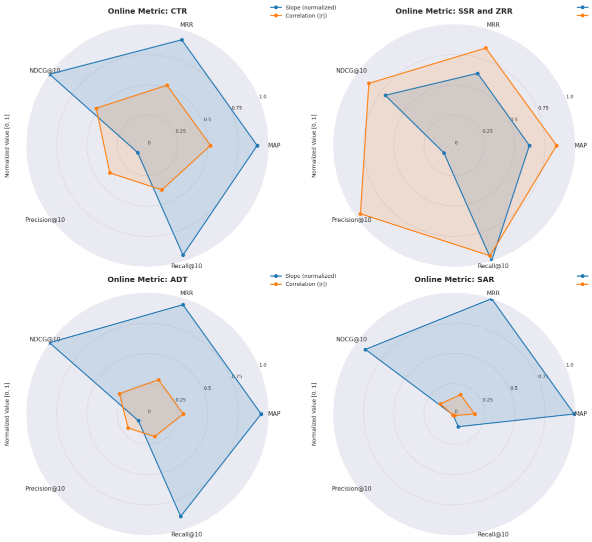

Figure 3: Radar charts comparing the relative sensitivity (normalized slope) and linear alignment (absolute Pearson correlation r) of offline metrics with respect to each online metric. Slopes are normalized separately for each online metric.

cision@10–SSR/ZRR and Recall@10–SSR/ZRR, for which the Pearson correlation coefficients exceed 0.9. This strong alignment is unsurprising given the conceptual similarity between these metrics: SSR and ZRR measure whether users clicked anything at all, while Precision@10 and Recall@10 measure whether relevant documents appear in the top ranks. In contrast, several other offline–online metric combinations exhibit high slope values but weak correlations. This reveals an important limitation of interpreting sensitivity in isolation: high sensitivity does not guarantee that an offline metric reliably captures changes in the online metric scores of systems observed under online evaluation.

Some of the observed behaviours can be further understood by considering how the metrics are defined. For example, NDCG@10 is particularly sensitive to small changes in ranking because it applies a logarithmic discount that heavily rewards placing highly relevant documents earlier in the list. Even minor adjustments in the top positions can therefore produce noticeable changes in NDCG@10, which helps explain why it often shows relatively high sensitivity across online metrics. In contrast, ADT does not align well with any of the offline metrics because it reflects dwell time rather than relevance. A user may click a highly relevant document, find the needed information immediately, and leave the page quickly, resulting in a short dwell time despite the document being genuinely useful. Since offline metrics are based on static relevance judgments rather than behavioural engagement, they are not designed to capture this type of user interaction.

These findings also have practical implications for metric choice. Metrics that exhibit high sensitivity but weak alignment—such as NDCG@10 with respect to CTR or ADT—may give the impression of being informative because they respond strongly to changes in online behaviour, yet they may not reliably distinguish between systems. Conversely, metrics that show both high sensitivity and high alignment for specific online behaviours, such as Recall@10 and Precision@10 for SSR/ZRR, offer more reliable guidance for model selection in those contexts. The results therefore suggest that offline evaluation should not rely on sensitivity alone: both sensitivity and alignment must be considered jointly when selecting metrics intended to reflect user-centric outcomes.

Overall, this analysis clarifies how offline metrics behave in relation to online performance signals and underscores the importance of understanding both responsiveness and consistency.

7.2 Limitations

While our analysis provides insight into how offline metrics relate to simulated online performance, several limitations should be acknowledged. First, the slope-based sensitivity analysis implicitly assumes a linear relationship between offline and online metrics. As the results show, this assumption does not always hold: many metric pairs exhibit steep slopes but weak correlations. Only a small number of offline–online combinations—most notably Precision@10–SSR/ZRR and Recall@10–SSR/ZRR—display correlations above 0.9, indicating a relationship that is close to linear. For the remaining pairs, high sensitivity alone does not guarantee that an offline metric reliably captures changes in online performance.

Second, the complexity of offline–online alignment limits the extent to which strong linear relationships can be expected. Offline metrics rely on pre-labeled datasets with static and often incomplete relevance judgments, whereas online metrics arise from dynamic user interactions in which users independently decide whether a document is useful and may revise their judgments over time[25]. Because offline relevance is fixed while online usefulness is context-dependent and behavioural, perfect correspondence between offline and online metrics is unlikely.

Finally, our online metrics are derived from simulated user behaviour, including click models and LLM-based simulations. Although these simulations provide controlled and reproducible conditions, they cannot fully replicate the complexity of real user behaviour. As a result, the alignment patterns observed in this study may differ from those obtained with real user interactions. These limitations should be kept in mind when interpreting the results and when generalizing them to real-world IR systems.

7.3 Relation to Prior Work

The Pearson correlation coefficients observed in our study are generally lower than those reported in the prior study that analyzes correlations between offline metrics[4]. This difference is expected, as offline-only comparisons use the same set of pre-labeled relevance judgements, while our experiment involves simulating real user behavior through click models and LLM-based simulations, where the LLM acts as the user[17]. Since LLMs do not perfectly mirror human decision-making, the simulated online metrics inevitably differ from those derived from real user interactions. Therefore, numerical comparisons between offline-only correlations and offline-online correlations should be made with caution.

When compared to the results of Chen et al., 2017[6], the correlation coefficients are comparable or even higher. However, Chen's study calculated the correlation between offline metrics and real user satisfaction, obtained from real users. While both approaches aim to capture real user experiences, this difference in signal type complicates direct comparisons. Nonetheless, the presence of similar correlation magnitudes suggests that the online metrics we considered in this study can serve as meaningful intermediates between offline evaluation and users' perception of success.

Our correlation analysis further shows that different offline metrics vary in how consistently they reflect changes in online performance. This observation aligns with conclusions from prior research, which argue that offline metrics capture different aspects of system effectiveness and user satisfaction. Metrics such as NDCG@10 may be more responsive to certain user behaviors, while others fail to reflect those behaviors reliably. These findings reinforce the view that no single offline metric universally captures online performance, highlighting the importance of careful metric selection depending on the evaluation goal.

8 Responsible Research

In this section, we critically reflect on the ethical aspects of our research. First, we discuss the reliability of our work. We have been faithful in presenting our research data. The results come directly from the experiment, without being manipulated or edited in any way. The experiments and results are publicly available in the GitHub[4] repository for anyone to inspect and verify. Together with Section 4 and Section 5 of this paper, the public availability of our repository also ensures reproducibility. That is, any computer scientist with the appropriate knowledge and coding skills will be able to replicate the experiment and confirm our results. By faithfully presenting our findings, we ensure that any research building on this study is not misled by modified or fabricated data.

We also made ethical considerations by simulating online user behaviour instead of involving real users. The results of this study therefore do not pose any risk of harm to participants, unlike experiments that rely on human subjects. Chen et al., 2017[6] discusses involving actual users to measure satisfaction with ranking systems, but we avoid any potential harm to participants by refraining from including concrete values from that study. We only reference the range of correlations they report, and these values do not reveal any personal information about participants.

8.1 Use of LLMs

In this work, we used LLMs in two ways. First, we used LLMs to simulate online user behaviour for the online metrics ADT and SAR[17]. The LLM used in this simulation was assigned a persona and instructed to act as a user while examining the results of a query. The LLM was therefore not used as a source of information or critical reasoning, even though acting as a user may appear to involve some form of reasoning. We emphasize that the LLMs used in this study were not involved in analyzing our results in any way.

The second use of LLMs was for proofreading our written text to correct language mistakes and improve clarity. The LLMs involved in this process were ChatGPT and Copilot. However, we made a conscious effort to paraphrase the output returned by these tools and to replace words with synonyms that better matched the meaning and nuance we intended to convey. In doing so, we sometimes used LLMs to generate multiple paraphrased versions of a sentence and combined parts of these versions to best achieve our writing goals.

9 Conclusions and Future Work

9.1 Conclusions

This study set out to determine how sensitive commonly used offline IR evaluation metrics are to changes in offline–online alignment, and whether these metrics respond consistently across different ranking systems. Our findings show that sensitivity varies substantially across offline metrics and depends strongly on the type of online behaviour being measured. No offline metric was universally sensitive, and high sensitivity did not always coincide with strong offline–online alignment. These results demonstrate that sensitivity and alignment capture distinct aspects of offline–online relationships and must be considered jointly when evaluating the reliability of offline metrics.

Only a subset of offline–online metric pairs exhibited strong linear relationships. Precision@10 and Recall@10 showed high alignment with SSR and ZRR, while many other combinations displayed weak correlations despite having steep slopes. This confirms that sensitivity and alignment do not necessarily coincide and should be treated as distinct properties.

Overall, the results indicate that offline metrics differ both in how strongly they react to changes in online performance and in how consistently they reflect differences between systems. These findings suggest that offline evaluation should be used with care: metrics that appear highly responsive may still fail to track user-centric outcomes reliably, while metrics with strong alignment may only do so for specific types of online behaviour. As offline and online metrics capture different aspects of relevance and user engagement, perfect correspondence should not be expected. These insights highlight the need for evaluation frameworks that consider both sensitivity and alignment when selecting offline metrics.

9.2 Future Work

There are several limitations in our study that point toward promising directions for future work. First, the relationship between offline metrics and online performance should be examined in greater depth. Many offline–online metric pairs appear to follow non-linear patterns, suggesting that linear sensitivity measures may capture only part of the underlying relationship. Another direction worth considering is adding more offline and online metrics. Expanding the current set of metrics could provide a more complete view of offline-online

[4] https://github.com/AinzOoalGown123/Metric-Sensitivity-Analysis

alignment and sensitivity. Since the relationship between offline and online metrics is often not linear, a different way to measure the sensitivity of metrics might also be worthwhile. Finally, it might be advantageous to involve real users instead of using simulations for online metrics. While simulated online metrics provide controlled conditions, validating these findings with real user interactions would provide more realistic results.

References

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Evaluation in information retrieval. In *An Introduction to Information Retrieval*, chapter 8, pages 151–175. Cambridge University Press, Cambridge, England, 2009.

[2] Katja Hofmann, Lihong Li, and Filip Radlinski. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 10(1):1–117, 2016.

[3] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020.

[4] Mucahid Kutlu, Vivek Khetan, and Matthew Lease. Correlation and prediction of evaluation metrics in information retrieval. *arXiv preprint arXiv:1802.00323*, 2018.

[5] Louise T. Su. Evaluation measures for interactive information retrieval. *Information Processing & Management*, 28(4):503–516, 1992.

[6] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 15–24, 2017.

[7] David Freedman, Robert Pisani, and Roger Purves. *Statistics*. W. W. Norton, 4 edition, 2007.

[8] Ellen Voorhees and Dawn Tice. The trec-8 question answering track evaluation. In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, 2000.

[9] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[10] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, 2005.

[11] Murat Turan and Ben Carterette. Predicting user satisfaction with search results. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1229–1230, 2011.

[12] Andrew Turpin, Falk Scholer, and Mark Sanderson. Intents and beliefs in the search engine result page. *Journal of the American Society for Information Science and Technology*, 62(12):2376–2387, 2011.

[13] Mark D. Smucker and Chandan P. Jethani. Time pressure and system delays in information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 789–790, 2008.

[14] Chang Liu, Jing Liu, Yinglong Wang, Nicholas J. Belkin, and Xiaojun Zhang. Understanding and predicting user behavior in interactive information retrieval. *Journal of the Association for Information Science and Technology*, 71(9):1045–1062, 2020.

[15] Susan Dumais and Thorsten Joachims. Implicit measures of user interests and preferences. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2001.

[16] Ben Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. *ACM SIGIR Forum*, 45(1):17–22, 2011.

[17] Ziliang Zhang. Evaluating prompting strategies for reliable llm-based user simulation in information retrieval. 2026.

[18] Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 4526–4533. ACM, 2021.

[19] Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. Continuous integration for reproducible shared tasks with tira.io. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, pages 236–241, Berlin Heidelberg New York, April 2023. Springer.

[20] Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. The information retrieval experiment platform. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, pages 2826–2836. ACM, July 2023.

[21] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A hetero-

geneous benchmark for zero-shot evaluation of information retrieval models. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[22] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019.

[23] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM, 2020.

[24] Thibault Formal, C. Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval. *CoRR*, abs/2109.10086, 2021.

[25] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.

A Additional Figures

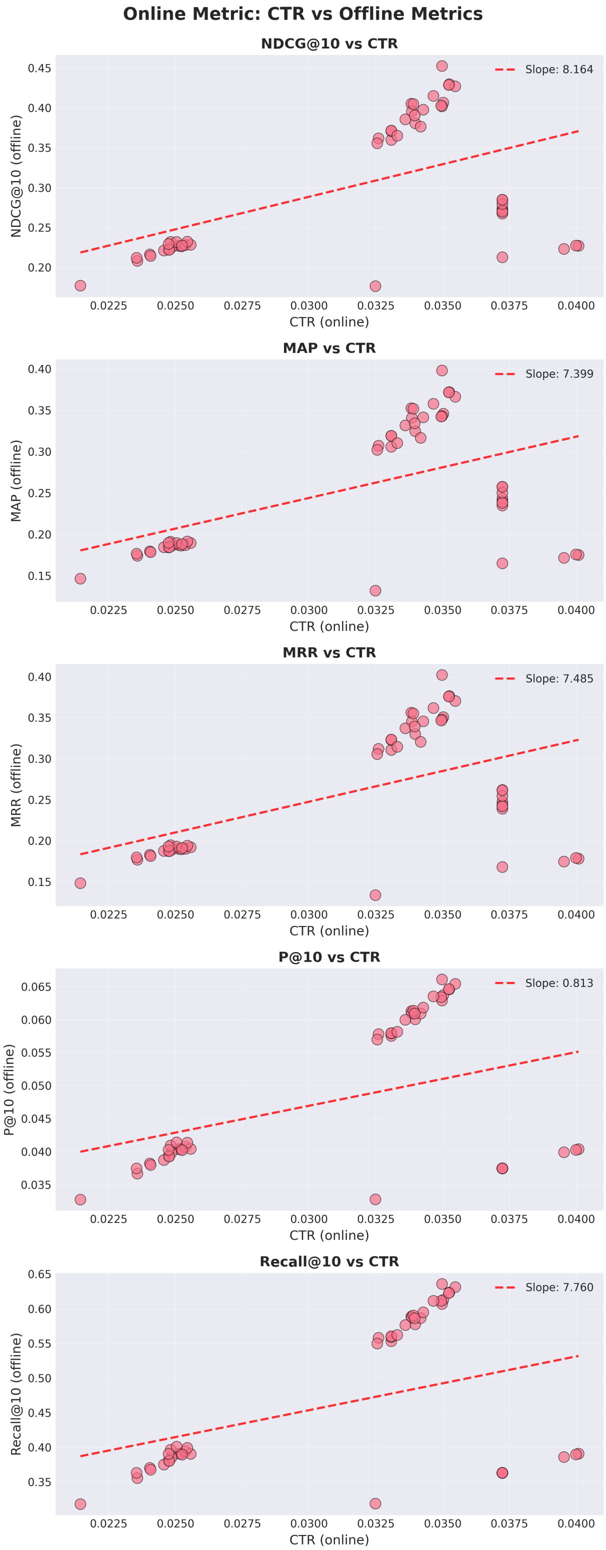

Figure A1: Offline metrics vs. CTR.

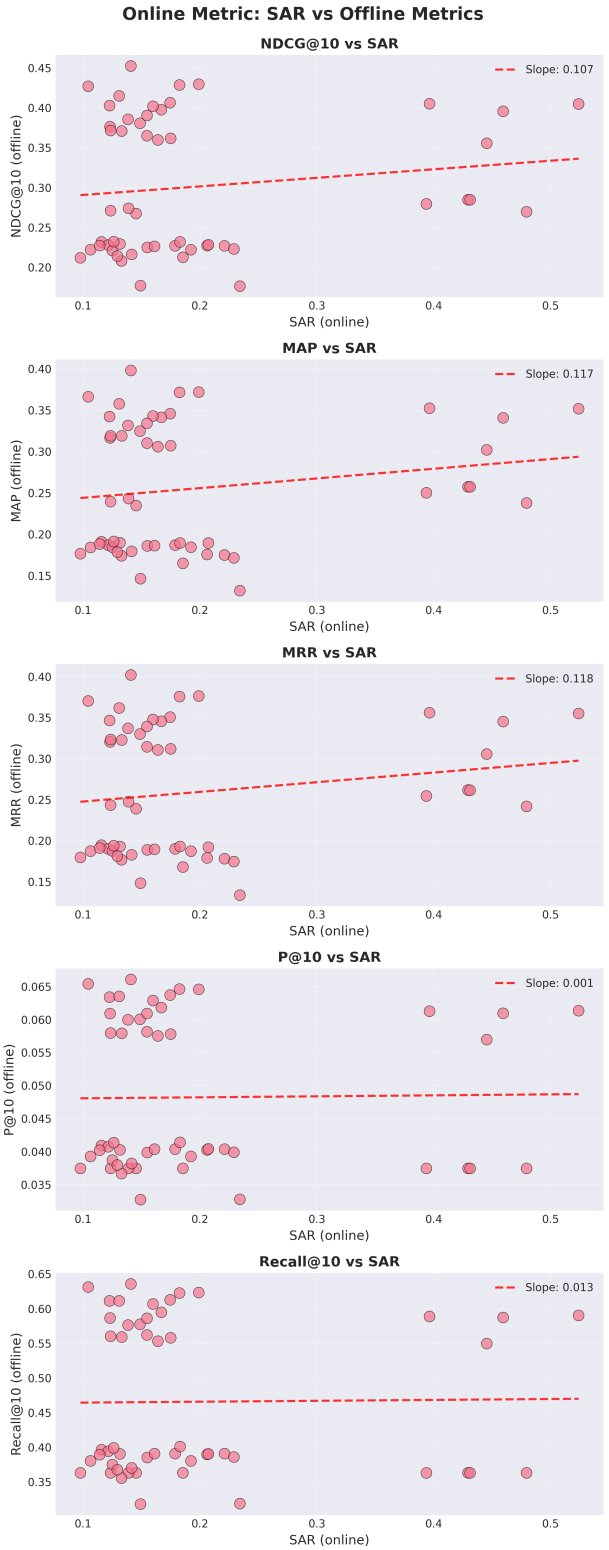

Figure A2: Offline metrics vs. SAR.

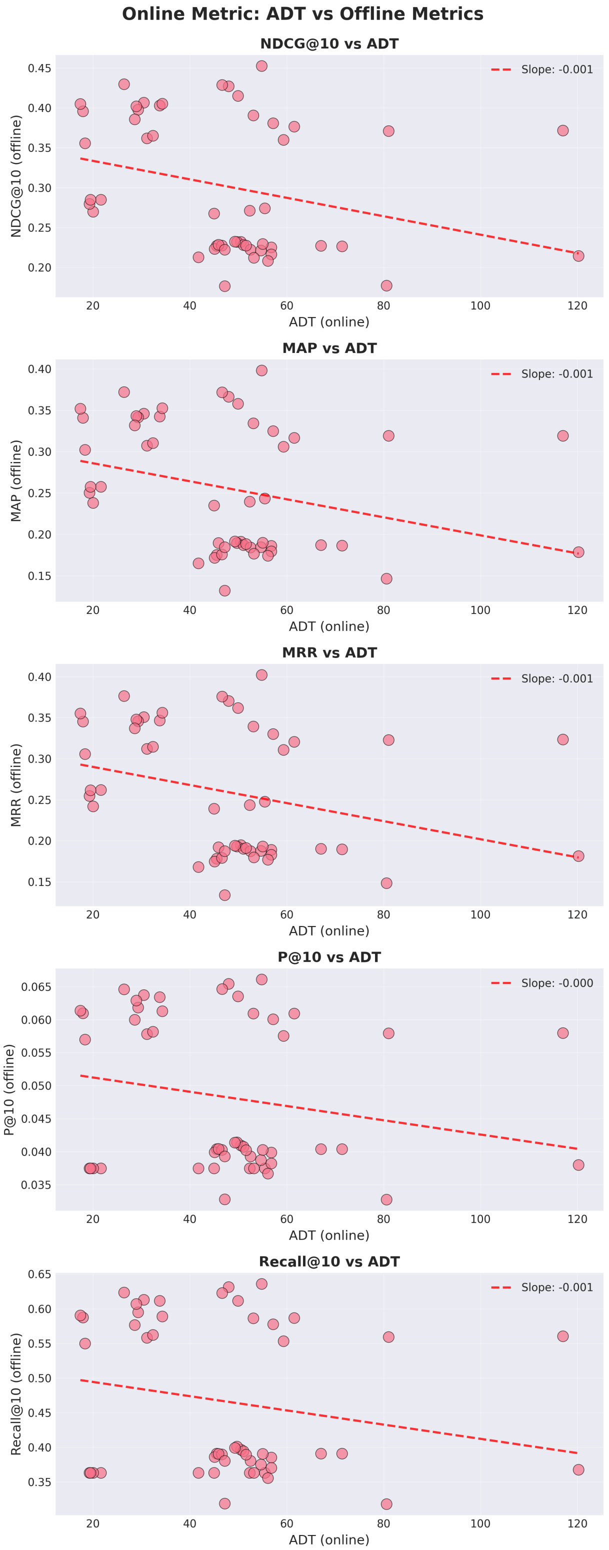

Figure A3: Offline metrics vs. ADT.

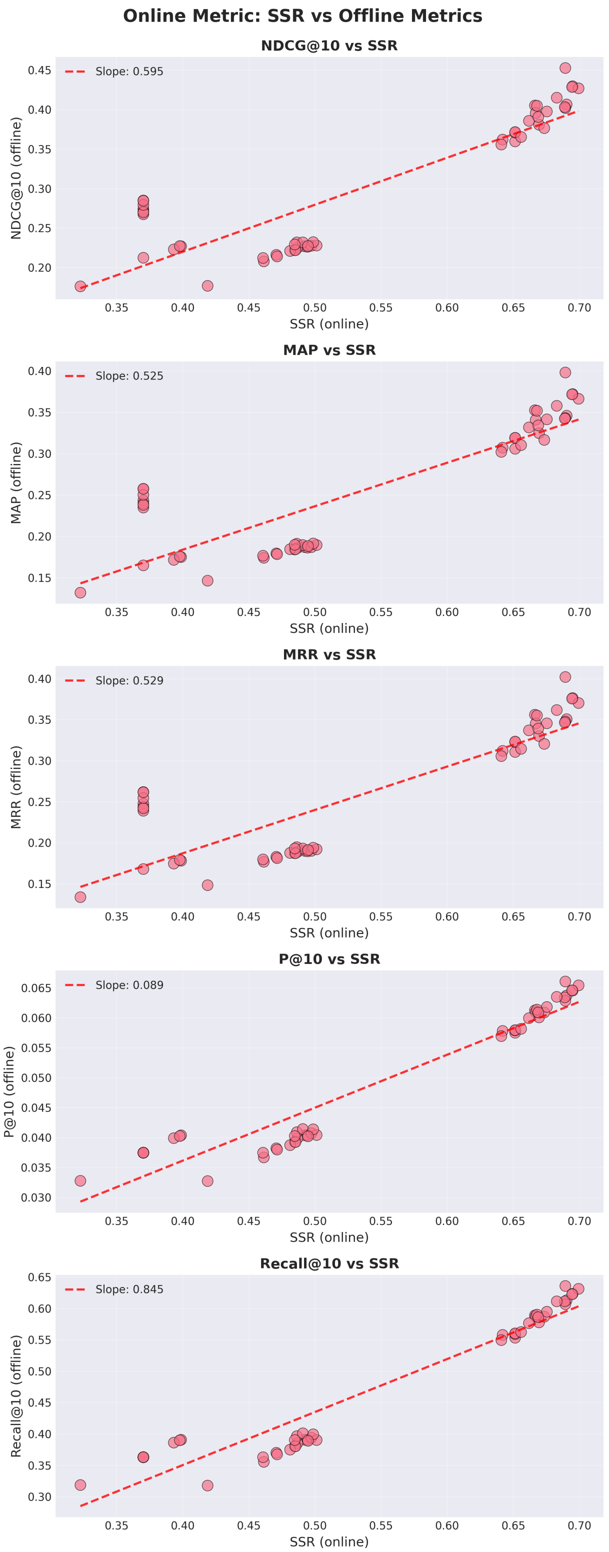

Figure A4: Offline metrics vs. SSR.

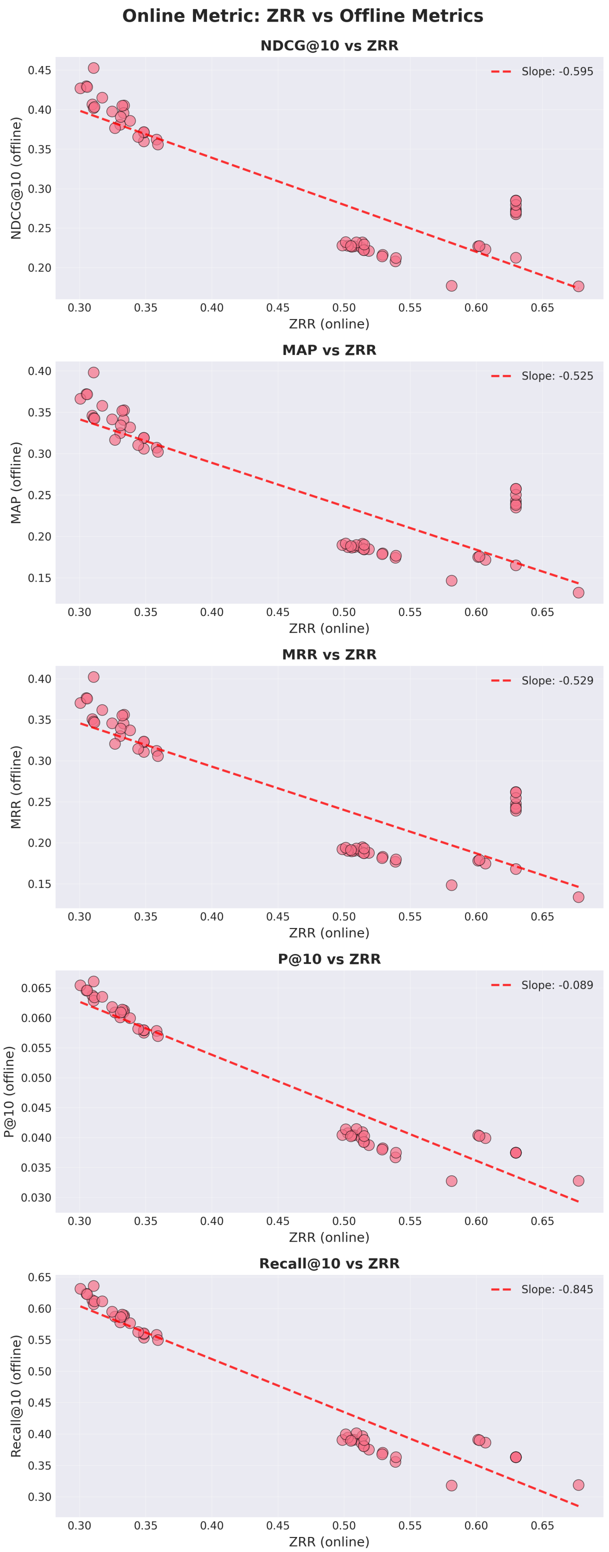

Figure A5: Offline metrics vs. ZRR.