Diversification as Risk Minimization

Anonymous Author(s)

Abstract

Users tend to remember failures of a search session more than its many successes. This observation has led to work on search robustness, where systems are penalized if they perform very poorly on some queries. However, this principle of robustness has been overlooked within a single query. An ambiguous or underspecified query (e.g., "jaguar") can have several user intents, where popular intents often dominate the ranking, leaving users with minority intents unsatisfied. Although the diversification literature has long recognized this issue, existing metrics only model the *average* relevance across intents and provide no robustness guarantees. More surprisingly, we show theoretically and empirically that many well-known diversification algorithms are no more robust than a naive, non-diversified algorithm. To address this critical gap, we propose to frame diversification as a risk-minimization problem. We introduce **VRisk**, which measures the expected risk faced by the least-served fraction of intents in a query. Optimizing VRisk produces a robust ranking, reducing the likelihood of poor user experiences. We then propose **VRisker**, a fast greedy re-ranker with provable approximation guarantees. Finally, experiments on NTCIR INTENT-2, TREC Web 2012, and MovieLens show the vulnerability of existing methods. VRisker reduces worst-case intent failures by up to 33% with a minimal 2% drop in average performance.

Code and appendix (w/ additional experiments and proofs) are provided in https://github.com/anonymousIR26/wsdm-sup. For ease of cross-referencing from the appendix, the repository also contains the main text.

ACM Reference Format:

Anonymous Author(s). 2025. Diversification as Risk Minimization. In . ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Users remember the worst search sessions far more than the average ones [12, 23, 25, 49, 55]. This insight has motivated extensive research on search robustness [12, 14, 16, 17, 32, 46, 49]. These prior works typically consider the downside risk, with the goal of minimizing the chances that users are unsatisfied with the search results. The fundamental principle is that search systems that perform poorly on some queries should be penalized, even if they perform well on average. However, these studies have only focused on the risks of rankings at the *query level* and have not considered the potential downside risks *within a query*.

This limitation becomes particularly problematic when queries are ambiguous or underspecified, leading to multiple possible user intents [10]. Consider a search for "jaguar." While many users may seek information about the car brand, those interested in the animal may find search results entirely dominated by the car brand. Naive ranking algorithms, designed to maximize a single relevance score, naturally produce such unbalanced results, where users with minority intents end up failing their search sessions.

The intent-based search diversification literature has long attempted to address this robustness-within-query challenge, aiming to generate a ranking to cover multiple intents [1, 9, 31, 36]. Yet, although preventing intent failures is a core motivation for diversity, existing approaches typically measure the average relevance across intents rather than explicitly addressing the worst-case outcomes. Importantly, we show mathematically that, in many cases, the family of intent-weighted diversity metrics [1, 9, 36] offers *no protection* in terms of robustness. For example, in the toy example of Table 1, we show that NDCG-IA [1] favors a ranking that completely ignores 49% of user intents. Consequently, as our experiments reveal, many existing diversification algorithms are no more robust than a naive, non-diversified ranking.

In light of this, we convert diversification into a principled risk-minimization problem. We argue that a search result must be robust within a query, in a way that minimizes the chances that users are unsatisfied. Specifically, we introduce the first framework for intent-aware risk minimization within a ranking, providing an evaluation metric **VRisk**. Drawing inspiration from finance, VRisk analyzes risk by adapting Conditional Value at Risk (CVaR), evaluating the expected relevance loss for the least-addressed fraction of user intents within each query. VRisk answers the question: "*How bad is the ranking for the worst β-fraction of users.*" VRisk can be baseline-free, measuring absolute risk, but can also be computed relative to a baseline system, following the common practice in existing IR risk metrics like URisk [49] and GeoRisk [17]. VRisk is tunable and intuitive, giving both risk and guarantee in its evaluation. Additionally, we offer an efficient optimization algorithm called **VRisker** with a strong optimality guarantee, making our approach practical and theoretically robust for real-world problems.

Finally, we demonstrate that the robustness problem is mitigated through experiments on NTCIR INTENT-2 [34] and TREC Web 2012 [11] datasets. Additionally, this problem extends beyond search, as intent-based diversity has become a key concern in modern recommender systems as well [5, 22, 26, 41, 45, 51, 52]. Thus, we also test on the MovieLens 32M [20] to verify that our risk-aware framework addresses the same issues in recommender systems. Our results show that the algorithm reduces worst-case intent failures by up to 33% while sacrificing as little as 2% in average performance.

Our contributions are summarized as follows:

- We present the first framework for intent-aware risk minimization in search, introducing a novel CVaR-based metric for both absolute and baseline-relative risk assessment, formally bridging diversification and robustness.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- We demonstrate mathematically and empirically that many mainstream diversification methods are fundamentally not robust, often performing no better than a naive baseline.
- We develop an efficient greedy algorithm with a strong approximation guarantee.
- We demonstrate through extensive experiments on three datasets that our algorithm consistently reduces worst-case intent failures by up to 33% at the cost of only a 2% drop in average relevance.

2 Related Work

2.1 Risk-Aware Search and Recommendation

Most influential works on retrieval robustness treat the entire query, rather than individual intents, as the unit of risk [12, 16, 32, 49, 54]. The proposed risk-aware frameworks generally penalize systems that perform poorly on some queries, instead of looking at the average performance. By favoring such robust algorithms, they minimize the probability of users being unsatisfied. However, while considering risks at the query level, they have not considered the risk that intents may be poorly addressed within a single query.

Portfolio-theoretic approaches [44, 48] assess the variance of a single ranking by treating each document's relevance score as a risky asset and balancing expected relevance with variance. This perspective differs fundamentally from ours, as we focus on risk at the intent level rather than at the document level.

Other studies consider risks in decision-making, search, and recommendation [18, 19, 21, 24, 43], but model the random variable at the user, session, or policy levels. Our method complements these approaches by applying the tail-risk logic directly to the diversification step, guaranteeing that even the users with minority intents in the query are satisfied.

2.2 Diversification in Rankings

The naive ranking algorithm, which ranks documents in the order of relevance, often only shows the majority intents and ignores the minority intents of a user. The literature of diversification in search has long aimed at providing reasonable satisfaction across intent groups [1, 9, 31, 36]. However, despite such clear connections with the motivation for robustness, most diversification frameworks generally consider the average rather than risks or guarantees. In fact, we find that IA metrics [1] and D-measure [36], two of the most common diversification metrics, provide identical results to a naive, non-diversified metric in many cases. Since many diversification studies are built under their logic, we show empirically that many diversification algorithms perform very poorly in terms of robustness.

Some studies provide diversification methods and metrics aiming to minimize the probability of a user not clicking on any document [1, 31, 51], which is fundamentally similar to our goal of robustness. However, typically assume a cascade behavior assumption *and* associate only click probabilities, which cases that are rare in practice. As we will show in the experiments, these methods perform poorly with graded relevance (e.g., nDCG) and explicit ratings. Some other studies consider the coverage of information without explicitly considering intents, thus conceptually different from what we aim [4, 30].

Fairness in rankings is also a related field that ensures that each candidate group gets a fair amount of exposure in a ranking [35, 39, 57]. It does not ensure that the *user* is given a diversity of exposure, thus conceptually different from our goal.

3 Problem Formulation

3.1 Preliminaries

We consider the evaluation of a single top-k ranking, denoted R_k. Let q denote a query with intent set $C(q) = \{c_1, \ldots, c_m\}$, where each intent c_i represents a distinct information need associated with q.

We denote $\Pr(c \mid q)$ as the probability that a user issuing query q has intent c. Following common prior work [1, 8, 31, 36], we assume the intents are mutually exclusive, i.e., $\sum_{c \in C(q)} \Pr(c \mid q) = 1$.

Let $\mathrm{rel}(d \mid q, c) \in [0, \mathrm{rel}_{\max}]$ be the relevance of document $d \in \mathcal{D}$ for query q given a user intent c. We suppose that $\mathrm{rel}(d \mid q, c)$ and $\Pr(c \mid q)$ are known or estimated beforehand, following the common setting in prior work on intent-based diversity [1, 9, 10, 36, 45, 51]. Their estimation is outside the scope of this work.

Naturally, the *raw* relevance of the document (that ignores intents) is calculated by the expected relevance

$$\begin{aligned} \mathrm{rel}(d \mid q) &= \mathop{\mathbb{E}}_{c \sim \Pr(c \mid q)} [\mathrm{rel}(d \mid q, c)] \\ &= \sum_{c \in C(q)} \Pr(c \mid q) \mathrm{rel}(d \mid q, c). \end{aligned} \tag{1}$$

This formulation is widely adopted in prior work [3, 6, 27, 36, 50, 53].

3.2 Standard Metric

To measure the value of a ranking, we take the average relevance

$$V_{\mathrm{std}}(R_k \mid q) = \frac{1}{k} \sum_{d \in R_k} \mathrm{rel}(d \mid q). \tag{2}$$

While we use average relevance as a default metric throughout the main text, the base metric V can be replaced with other standard metrics (e.g., nDCG, DCG, ERR, Precision@k, RBP), which we will also discuss in our experiments.

The optimal ranking for Eq. (2) and most standard metrics like nDCG would be the naive ranking algorithm that selects documents in the descending order of relevance

$$\mathrm{Naive}(q) = \operatorname*{argsort}_d \mathrm{rel}(d \mid q). \tag{3}$$

However, since the relevance of a document is heavily dependent on the intent probability (see Eq. (1)), such naive rankings may contain only documents of the majority intents, and some minority intents may be completely ignored. This is not robust, as it may result in a poor search session when the user had a minority intent [1, 9, 36, 40].

3.3 Intent-Weighted Metric

We now present another common evaluation method of a ranking *given some user intents*. We denote the value of a ranking *given intent* c as the average relevance

$$V(R_k \mid q, c) = \frac{1}{k} \sum_{d \in R_k} \mathrm{rel}(d \mid q, c). \tag{4}$$

Table 1: Toy example of how nDCG_{IW} (= NDCG-IA [1]) can neglect diversification in a top-2 ranking. (a) is the setting, and (b) shows the evaluation scores of each ranking. Although the intent probability of c_1 is only 0.02 more than c_2, nDCG_{IW} favors a non-diversified ranking. As a result, 49% of the users can be unsatisfied with the ranking.

(a) Setting.

	c_1	c_2
$\Pr(c \mid q)$	0.51	0.49
$\text{rel}(d_1 \mid q, c)$	1	0
$\text{rel}(d_2 \mid q, c)$	1	0
$\text{rel}(d_3 \mid q, c)$	0	1
$\text{rel}(d_4 \mid q, c)$	0	1

(b) Evaluation scores.

$R_{k=2}$	nDCG_{IW}	nDCG_{std}
$[d_1, d_2]$	**0.600**	**1.000**
$[d_1, d_3]$	0.502	0.871
$[d_3, d_4]$	0.400	0.667

Similar to Eq. (2), this denotes the average relevance of a ranking, but for a specific user intent c.

Given this per-intent value $V(R_k \mid q, c)$, another straightforward evaluation method of the ranking would be to compute the intent-weighted average (IW metric), denoted

$$V_{\text{IW}}(R_k \mid q) = \sum_{c \in C(q)} \Pr(c \mid q)\, V(R_k \mid q, c). \tag{5}$$

This evaluation method also demonstrates the family of "IA" metrics by Agrawal et al. [1] (e.g., NDCG-IA, MRR-IA), which are diversification metrics aiming to value minority intents more. For example, if the base metric V is replaced with nDCG (i.e., nDCG_{IW}), this matches NDCG-IA [1]. This is arguably the most common metric for diversity, as the same intent-weighted paradigm also underlies widely used diversification metrics and algorithms such as xQuAD [37], ERR-IA [6], D-measure [36], and α-nDCG [9].

3.4 IW Metrics are Surprisingly not Robust

While $V_{\text{IW}}(R_k \mid q)$ builds the ground of most diversification methods, we find that, surprisingly, when the base metric V is linear (e.g., DCG, average relevance, Precision@k), the intent-weighted metric in Eq. (5) is identical to the standard metric in Eq. (2). Formally, using average relevance as the base metric V, we have

$$\begin{aligned} V_{\text{IW}}(R_k \mid q) &= \sum_{c \in C(q)} \Pr(c \mid q) \left(\frac{1}{k} \sum_{d \in R_k} \text{rel}(d \mid q, c) \right) \\ &= \frac{1}{k} \sum_{d \in R_k} \left(\sum_{c \in C(q)} \Pr(c \mid q)\, \text{rel}(d \mid q, c) \right) = V_{\text{std}}(R_k \mid q). \end{aligned} \tag{6}$$

This means an algorithm that maximizes V_{IW} will favour the majority intent just as the V_{std} would. **Thus, for linear base metrics V, the intent-weighted metric *offers no protection* to minority intents.** Interestingly, D-measure [36], another common diversification method, also reduces to the standard metric. Consequently, the spectrum of intent-based diversification metrics and algorithms [1, 9, 36, 37] are *ineffective at providing diversity or robustness* with linear base metrics like DCG, Precision@k, and average relevance.

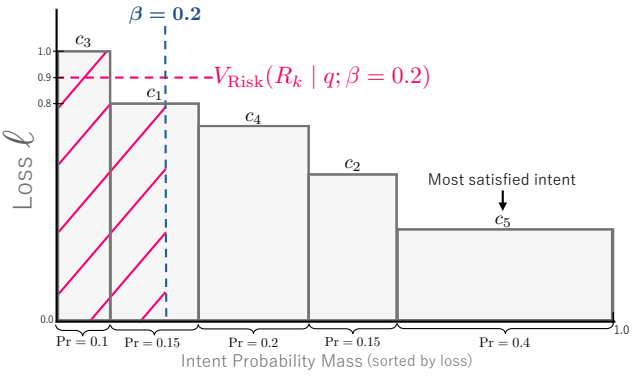

Figure 1: Illustration of VRisk. Bar height shows per-intent loss $\ell(R_k, q, c)$ and the bar width shows the intent probability $\Pr(c|q)$. 5 intents are sorted by loss (worst → best). VRisk takes the average of ▨, the worst-β intent probability mass.

For non-linear base metrics like nDCG, Expected Reciprocal Rank (ERR) [7] and Rank Biased Precision (RBP) [29], the equality like Eq. (6) no longer holds. Yet, these metrics can still fail to measure the diversity of a ranking. Table 1 shows a toy example of how nDCG_{IW} favors a ranking dominated by one intent, despite the two intent probabilities being almost equal. This is not at all robust, as **49% of the intents are completely ignored**.

Furthermore, empirical results in Section 6 reveal that intent-weighted diversification algorithms for such non-linear metrics are also barely more robust than naive algorithms.

4 Risk-Sensitive Metric: VRisk

To address the limitation that existing diversification methods fail to account for robustness, we convert diversification to a risk-minimization problem. Specifically, we propose a metric, **VRisk**, that measures the expected loss of the least-addressed β-fraction of intents, bringing the *Conditional Value at Risk* (CVaR) concept into IR. VRisk is intuitive, baseline-relative, and easy to tune.

For each possible intent, we define a loss function as the loss of intent-level value against the target level

$$\ell(R_k, q, c) = \left[V_{\text{tgt}}(q, c) - V(R_k \mid q, c) \right]_+, \tag{7}$$

where $[\cdot]_+ = \max(0, \cdot)$ and $V_{\text{tgt}}(q, c)$ is a *target level* for performance on intent c. $V_{\text{tgt}}(q, c)$ is a baseline function capturing the satisfaction threshold for intent c. Unless stated otherwise, we use the **oracle target**, which is the best possible ranking for the intent

$$V_{\text{tgt}}(q, c) = \max_{R'_k} V(R'_k \mid q, c). \tag{8}$$

By setting loss against the oracle, Eq. (7) answers the question: "How well does the ranking satisfy the intent compared to the ideal case?" We use loss instead of the raw value, as it is fair even when there exist more relevant documents for one intent than the others. Note, however, that when the target level is set to the oracle, the minimization of loss matches exactly the maximization of raw value.

Following previous query-level risk studies [17, 49], $V_{\text{tgt}}(q, c)$ may also be a value of any baseline algorithm or some heuristic threshold that we want to ensure, which we investigate in Section 6.

Now, to derive VRisk, let ζ be the β-fraction of the loss distribution, i.e. $\Pr\big(\ell(R_k, q, c) > \zeta\big) \le \beta$. Smaller β places more weight on rarer, worse-served intents. Then, VRisk is the expected loss of the worst β-fraction of intents, denoted

$$V_{\text{Risk}}(R_k \mid q; \beta) = \mathbb{E}\big[\ell(R_k, q, c) \;\big|\; \ell(R_k, q, c) \ge \zeta\big]. \tag{9}$$

Smaller VRisk is better because it reduces tail loss. Figure 1 illustrates VRisk when $\beta = 0.20$.

Since Eq. (9) is difficult to optimize directly, following the Rockafellar-Uryasev formulation [33], we express our VRisk metric as

$$V_{\text{Risk}}(R_k \mid q; \beta) = \min_{\zeta \in \mathbb{R}} \left[\zeta + \frac{1}{\beta} \sum_{c \in C(q)} \Pr(c \mid q)\ [\ell(R_k, q, c) - \zeta]_+ \right]. \tag{10}$$

VRisk can tell us both the risk of the ranking and the guarantees, as explained in the example below.

Example 1. *If $V_{Risk}(R_k \mid q; \beta) = 1.0$ at $\beta = 0.05$, we obtain*

- *Risk: The worst 5% of intents have an expected loss of 1.0,*
- *Guarantee[1]: 95% of the intents have the loss always smaller than 1.0.*

4.1 Property of VRisk

Our VRisk metric is controllable and generalizable. Specifically, $\beta \in (0, 1]$ is a tunable parameter that controls which worst fractions we would like to care about. An interesting property is described following proposition.

Proposition 1. *When $\beta = 1$, VRisk reduces to the expected loss*

$$V_{Risk}(R_k \mid q; \beta = 1) = \sum_{c \in C(q)} Pr(c \mid q) \cdot \ell(R_k, q, c). \tag{11}$$

Thus, our framework generalizes the average intent-weighted optimization, while enabling control of tail risk via β. Specifically, a smaller β means a *strictly* safer system.

In short, VRisk converts diversification into an intuitive **risk-minimization** problem driven by a single, intuitive parameter β. Using VRisk, we are able to explicitly measure the robustness of a ranking, unlike existing metrics that measure the average.

Why not minimax? The minimax criterion, which maximizes the value of the single worst–served case, is typically more common in the IR literature [14, 15, 28, 47]. In our case, minimax solves

$$\min_{R_k} \max_{c \in C(q)} \ell(R_k, q, c). \tag{12}$$

In Figure 1, the minimax task would be to minimize the loss of only the left-most intent (i.e., intent with the most loss).

However, minimax is not suited for our problem, as it is not tunable. Minimax must optimize for the worst intent, even if that intent only occurs at a 0.0001% chance. In reality, we should not lower the average utility just to fulfill such intent. In contrast, VRisk looks at the worst *fraction* of the intents, where β controls the size of the fraction, so it addresses a 0.0001% intent *only if* it lowers fraction loss the most. Moreover, *VRisk subsumes minimax* as the special case of $\beta \to 0$.

[1]Even better, 95% of the intents have loss $\le \zeta \le 1.0$

Algorithm 1 VRisker

Require: Query q, candidates $\mathcal{D}$, length k, risk level β

```
Ranking R ← ∅
for each rank i = 1, ..., k do
    // select document that minimizes VRisk
    d* ← arg min_{d ∈ D\R} V_Risk(R ∪ {d} | q; β)
    Tie-break by maximizing V_IW(R ∪ {d} | q)
    // append document to rank i
    R ← R ∪ {d*}
end for
return Ranking R
```

5 Optimization of VRisk: VRisker

The VRisk objective gives us an explicit target of *minimizing tail risk*, but finding the global optimum is computationally intractable, as shown in the following proposition.

Proposition 2 (NP-hardness). *For variable k and any $\beta \in (0, 1]$, minimizing VRisk*

$$\min_{R_k \subseteq \mathcal{D},\, |R_k| = k} V_{Risk}(R_k \mid q; \beta)$$

is NP-hard. Proof in Appendix A.1.

Due to the NP-hardness of the optimization problem, we propose an efficient ranking algorithm called **VRisk E**fficient **R**anker (**VRisker**), shown in Algorithm 1. VRisker is a greedy algorithm that picks documents iteratively from the top position. At each step, it adds the document that most reduces VRisk. When there is a tie, it adds the document that maximizes the IW metric $V_{\text{IW}}(R \mid q)$.

VRisker is practical in terms of complexity. Let $m = |C(q)|$ and $n = |\mathcal{D}|$. Each risk evaluation involves $O(m)$ arithmetic operations. The run time is $O(k^2 nm)$, but drops to $O(knm)$ with incremental updates, which is identical to the speed of prior greedy diversification algorithms [1, 6, 37, 40]. Empirical comparison is in Figure 8.

5.1 Approximation Guarantee of VRisker

The VRisker algorithm has strong theoretical guarantees. When the base metric $V(\cdot \mid q, c)$ is *modular* (e.g., average relevance, Precision@k), the per-intent loss $\ell(R, q, c)$ is monotone non-increasing, and the *risk-reduction* function

$$\Phi(R) = V_{\text{Risk}}(\varnothing \mid q; \beta) \;-\; V_{\text{Risk}}(R \mid q; \beta) \tag{13}$$

is *monotone submodular*.[2] Given the submodularity, we are able to provide an approximation guarantee on VRisker, in Theorem 1.

Theorem 1 ($(1 - 1/e)$ Optimality Guarantee). *For modular base metric V, let $R_k^\star$ be the optimal length-k ranking. Then, VRisker returns R_k such that*

$$\Phi(R_k) \;\ge\; \left(1 - \tfrac{1}{e}\right) \Phi(R_k^\star), \tag{14}$$

i.e., it captures at least 63% of the optimal risk drop. Proof in Appendix A.2.

[2]Submodularity follows because (i) adding a document can only lower each hinge-loss term, and (ii) the convex combination inside CVaR is linear in these losses.

5.2 Guarantee for Non-Modular Metrics

When the base metric V is non-modular, such as nDCG, the risk-reduction function $\Phi(R)$ is no longer submodular, so Theorem 1 does not hold. However, we can establish a formal approximation guarantee for VRisker using the concept of the *submodularity ratio* [13]. A function has a submodularity ratio $\gamma \in (0, 1]$ if the marginal gain of adding an element to a larger set is at least a γ-fraction of the marginal gain of adding it to a smaller subset.

For example, for VRisk optimization with nDCG as the base metric, we can show that the risk-reduction function $\Phi(R)$ has a data-independent submodularity ratio $\gamma > 0$. Building on Bian et al. [2], we prove the following theorem.

Theorem 2 (nDCG-Risk Approximation). *For any query q, risk level β, and cut-off k, VRisker returns a ranking R_k that satisfies*

$$\Phi(R_k) \;\geq\; \left(1 - e^{-\gamma}\right) \Phi(R_k^{\star}), \tag{15}$$

where $\gamma \geq w_k / \sum_{i=1}^{k} w_i$ is the submodularity ratio, with $w_i = 1/\log_2(1+i)$ being the nDCG discount at rank i. Proof in Appendix A.3.

Theorem 2 provides a worst-case guarantee for VRisker's performance. The ratio γ depends only on the ranking depth k, where for typical depths such as $k = 5, 10, 20$, the lower bound on γ is $0.15, 0.09, 0.05$, respectively. While this theoretical bound is looser than when V is modular, we show that VRisker is not an arbitrary heuristic. We also examine the optimality of VRisker on non-modular base metrics in Section 6.

To sum up, VRisker is a theoretically grounded and efficient algorithm that explicitly minimizes the risks of a user receiving a poor ranking.

6 Experiments

In this section, we extensively evaluate our metric and method by comparing with other re-ranking approaches on NTCIR INTENT-2 [34], TREC Web 2012 [11], and on MovieLens 32M datasets [20].

6.1 Experiment Setup

Datasets. NTCIR INTENT-2 and TREC Web 2012 provide a list of intents for each query, with their intent-specific graded relevance $\mathrm{rel}(d \mid q, c)$. NTCIR INTENT-2 provides 5 relevance grades, and TREC Web 2012 provides 6. While NTCIR INTENT-2 provides their intent probabilities $\Pr(c \mid q)$, TREC Web does not provide intent probabilities. However, following Clarke et al. [11], we let all intents in the TREC Web datasets occur with equal probabilities.

To test generalization beyond search, we also evaluate on MovieLens 32M [20], where intent-based diversity is of increasing interest [5, 22, 26, 41, 45, 51, 52]. We treat each user as a query and each genre as an intent. We select users with more than 200 ratings to ensure sufficient per-user data. Otherwise, different re-ranking algorithms can collapse to similar outputs under extreme sparsity. Following Steck [40], we estimate $\Pr(c \mid q)$ from the user's historical genre proportions and use explicit ratings as raw relevance $\mathrm{rel}(d \mid q)$. For multi-genre items, we define per-intent relevance via a Bayes-consistent allocation

$$\mathrm{rel}(d \mid q, c) = \mathbb{1}[c \in C(d)] \frac{\mathrm{rel}(d \mid q)}{\sum_{c' \in C(d)} \Pr(c' \mid q)}, \tag{16}$$

where $C(d)$ is the set of genre labels for d and $\mathbb{1}$ is an indicator function. This construction ensures Eq. (1). We adopt an evaluation-only protocol, where no recommender is trained, so as to isolate the effect of the re-ranking objective. Unobserved ratings are treated as non-relevant for evaluation only.

Additional statistics of the datasets are summarized in Table 2.

Table 2: Statistics of the datasets. (Note that the #Queries denote #Users in MovieLens.)

Dataset	#Queries	#Docs	Avg. #Intents per Query
INTENT-2 (JP) [34]	95	5,085	6.1
TREC Web '12 [11]	50	15,200	6.0
MovieLens 32M [20]	42,902	71,933	8.0

Experimental Parameters. We set $\beta = 0.10$ as the default, meaning that we focus on the average loss of the worst 10% of intents when computing V_{Risk}. We also provide experiments where we sweep β to different values. The ranking length k is also an experimental parameter, with the default length $k = 10$. The baseline performance $V_{\mathrm{tgt}}(q, c)$ is also an experimental parameter, but set to the oracle (Eq. (8)).

Importantly, we set the base metric V to average relevance (which we denote AvgRel) as in the main text. Therefore, for the majority of experiments, we only show results on VRisk and V_{std}, as $V_{\mathrm{std}} = V_{\mathrm{IW}}$ holds (refer to Section 3.4). However, we also test on nDCG, DCG, ERR [7], RBP [29], and Precision@k to show the generalizability of VRisk and VRisker.

Compared Methods. We compare VRisker with the following:

Naive. The naive ranking approach naively optimizes the standard metric V_{std} as in Eq. (3). It does not care about diversity.

IW-Greedy. IW-Greedy is a greedy maximization method that aims to maximize V_{IW} in Eq. (5). For linear base metrics (e.g., average relevance, DCG), IW-Greedy is *strictly optimal*, exactly matching the Naive method. For non-linear base metrics, IW-Greedy has a $(1 - 1/e)$ optimality guarantee [6].

xQuAD [37]. An intent-based diversification method. λ_{xQuAD}, the weight controls the balance between relevance and diversity, is set to 0.5[3].

MMR [4]. A diversification method that intends to maximize the coverage without relying on intents. Similarity of documents is computed by measuring the cosine similarity of TF-IDF vectors of document texts for search tasks, and tags and titles for the recommendation tasks. λ_{MMR}, the weight controls the balance between relevance and diversity, is set to 0.5[3].

For all main experiments, we also give comparisons with IA-SELECT [1], FA*IR [56], and Calibrated Recommendations (CR) [40] in Appendix B.4.

Unless specified otherwise, instead of reporting the raw metrics, results are shown as a percentage compared to the Naive ranking, so that all curves share the same 100% center line, making the trade-off between V_{Risk} and average performance (i.e., V_{std} / V_{IW}) more visible. For each obj $\in$ {Risk, IW, std}, we set

$$\Delta V_{\mathrm{obj}}(R_k) = \frac{V_{\mathrm{obj}}(R_k)}{V_{\mathrm{obj}}(\mathrm{Naive}(q))} \times 100. \tag{17}$$

[3]We sweep this in Appendix B.4 and confirm that it does not change the conclusion.

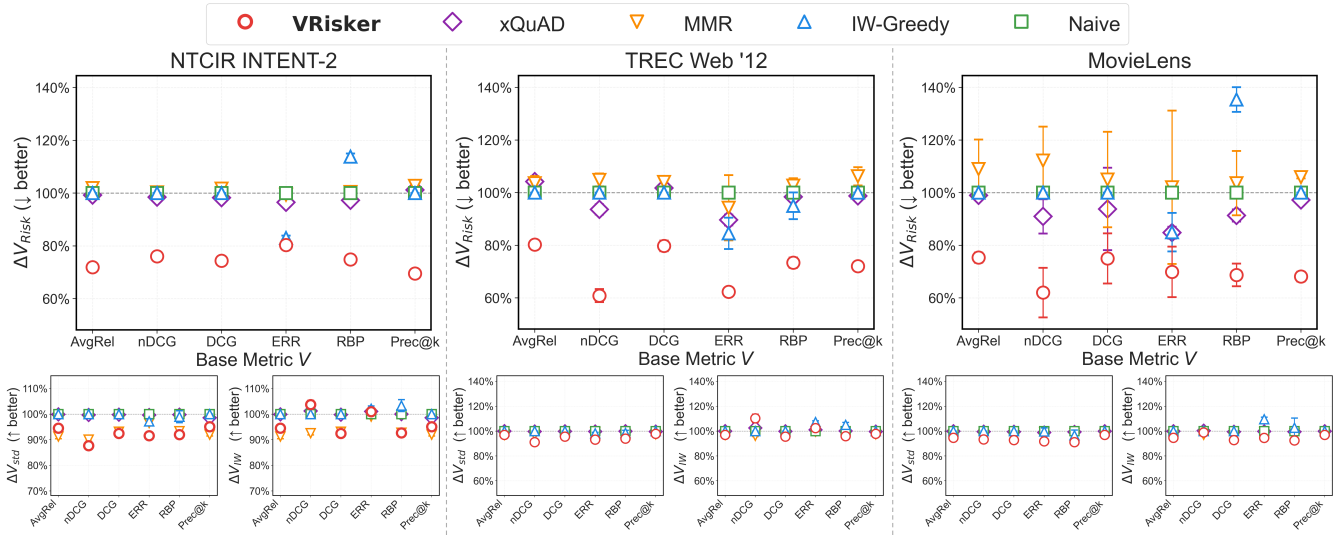

Figure 2: *VRisker is robust across base metrics* V. The plots show how different methods perform on varying base metrics V, tested on NTCIR INTENT-2 (left), TREC Web 2012 (middle), and MovieLens 32M (right). The top figures evaluate risk, ΔV_{Risk} (smaller is more robust), which measures the expected loss of the worst β-fraction of intents. The bottom figures evaluate the average performance, ΔV_{IW} and ΔV_{std} (larger is better). All values are relative to Naive = 100%. ($k = 10$, $\beta = 0.10$)

For transparency, for each experiment, we also report the results on the raw metrics in Appendix B.4. Note that $\Delta V_{\text{IW}} = \Delta V_{\text{std}}$ for linear bases, so we do not show both in most results.

The values shown are the $\Delta V_{\text{obj}}(R_k)$ averaged on all queries[4]. The error bars indicate the 95% confidence interval, calculated by treating each query/user as an independent observation drawn at random from the population. We additionally provide statistical significance testing in Appendix B.1.

6.2 Results and Q&As

We present our results via the following Q&As.

Q: Does VRisker work on different base metrics?

A: VRisker performs robustly across various base metrics, while existing methods are unstable and vulnerable.

Figure 2 compares the use of different base metrics V: AvgRel (average relevance, the default), nDCG, DCG, Expected Reciprocal Rank (ERR) [6, 7], Rank Biased Precision (RBP) on $p = 0.8$ [29], and Precision@k. For Precision@k, we binarize the relevance labels via threshold $(\text{relmax} + \text{relmin})/2$. We compare methods on naive-relative VRisk (ΔV_{Risk}) with $\beta = 0.10$, standard metric (ΔV_{std}), and IW-metric (ΔV_{IW}), where Naive is at 100%.

While VRisker has a $(1 - 1/e)$ guarantee for modular base metrics (e.g., AvgRel and Precision@k), we observe that VRisker is about 20-40% more robust than naive on other non-modular base metrics as well. Furthermore, VRisker only decreases the standard performance (V_{std}) by about 0-10%.

Other diversification methods are barely more robust, or even less robust, than the naive baseline. As we have argued in Section 3.4, IW-based diversification methods (i.e., IW-Greedy and xQuAD) are often very similar to the naive baseline, even when the base metrics are not linear. Interestingly, on base metric nDCG, IW-Greedy and the Naive algorithm behave exactly the same on all experimented settings and queries, despite nDCG not being linear (see Eq. (6)). This is because the greedy marginal at each rank reduces to sorting documents by expected gain under intent weights.

Q: How does the risk level, β, affect the performance?

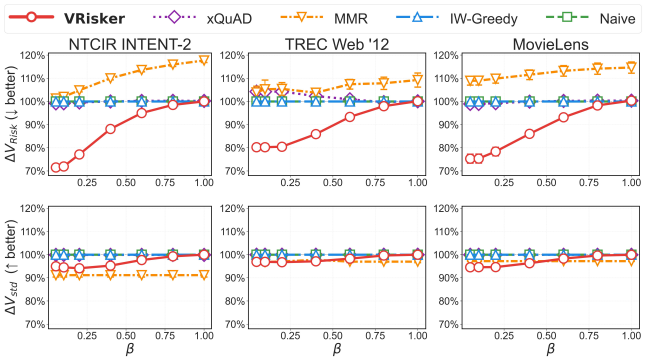

Figure 3: *VRisk/VRisker's pessimism is controllable via* β. The plots show results on VRisk and V_{std} (= V_{IW}).

A: As we make evaluation more pessimistic (smaller β), average performance decreases, demonstrating the tunability of VRisk & VRisker.

Figure 3 shows the performance of different methods when we vary the pessimism, β. Recall that a smaller β focuses on a smaller worst-case tail of intents. The results demonstrate a clear trade-off between standard performance (i.e., V_{std} / V_{IW}) and VRisk, which

[4]Note that we cannot compare with query-level robustness methods [17, 49], as they do not take the average on all queries, so cannot be compared on the same scale.

is a practical property discussed in Section 4.1. For example, when $\beta = 0.05$, VRisker has about 20-30% less worst-case loss, while sacrificing about 3-5% in standard performance. In contrast, when $\beta = 0.8$, VRisker reduces risk by about 2-3%, while having lost almost nothing in terms of the standard performance. When the platform should prioritize robustness over standard performance, β should be tuned lower.

Q: How does ranking length affect performance?

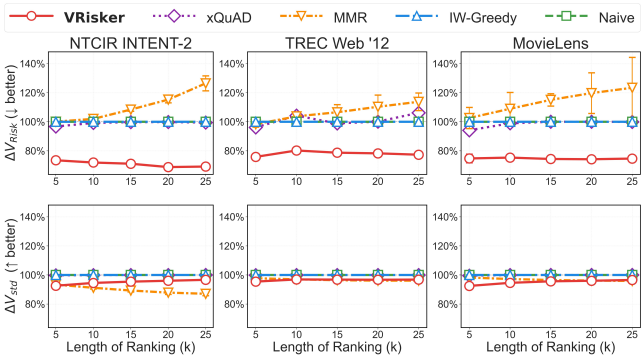

Figure 4: *VRisker is robust to ranking length.*

A: The average performance of VRisker improves as the ranking length increases, while keeping consistent robustness.

Figure 4 compares the performance on different ranking lengths on the three datasets. VRisker minimizes the worst β-fraction loss consistently compared to the alternatives, with a 20-30% decrease. We also observe that in terms of the standard performance ($V_{\text{std}}/V_{\text{IW}}$), VRisker performs better as the ranking is longer. In the best case, in INTENT-2 $k = 25$, we observe 33% reduction in VRisk while sacrificing only 2% in standard performance.

Q: Do we need perfectly accurate intent probabilities?

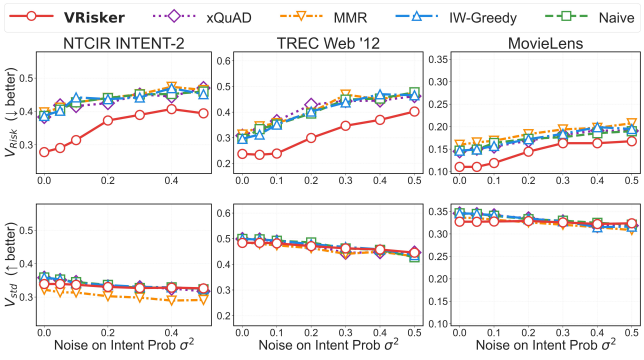

Figure 5: *The results are consistent when intent probabilities with noise added.* VRisk and V_{std} shown with raw values.

A: Preferable, but VRisker degrades more gracefully under noise.

Figure 5 plots the performance of diversification methods on various noise levels. For each intent probability, we perturb each probability as $\Pr(c \mid q) \leftarrow \Pr(c \mid q) + \epsilon_c$, with $\epsilon_c \sim \mathcal{N}(0, \sigma^2)$. We then clip to $[0, 1]$ and renormalize across intents. σ^2 controls the noise level, where the larger the value, the more noise when running the algorithms. We then evaluate using the true intent probabilities.

We observe that the VRisk increases for all methods, meaning less robustness. This is predictable since noisy intent probabilities trigger noise in both raw relevance calculation and IW metric calculation. Yet, VRisker maintains the lowest risk no matter the noise. Additionally, we observe that in terms of the standard performance, VRisker performs comparatively better as more noise is injected. At $\sigma^2 = 0.5$, VRisker performs the best on all datasets. This is counter-intuitive, but we hypothesize that this is because VRisker satisfies the main intents even with low predicted intent probabilities.

Q: How optimal is VRisker?

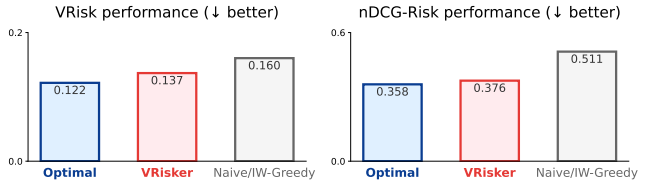

Figure 6: *VRisker is near optimal.* Values are averaged over all three datasets. Results on other bases are in Figure 10.

A: VRisker is nearly optimal for various base metrics.

Figure 6 shows the comparison of VRisker against the optimal performance, averaged over all three datasets. Here, Optimal is computed exactly per query by solving the VRisk objective as a MILP (PuLP + CBC solver). We show results tested on VRisk (i.e., base metric is average relevance) and nDCG, but we show results on other base metrics in Appendix B.2. As discussed in Section 5.1, VRisker is $(1 - 1/e)$ optimal when the base metric is modular (left-hand chart). As discussed in Section 5.2, for non-modular bases VRisker is not merely a heuristic. As a result, we observe that VRisker is near optimal in all experimented settings, reassuring the robustness of the approach.

Q: How does target level V_{tgt} affect the evaluation?

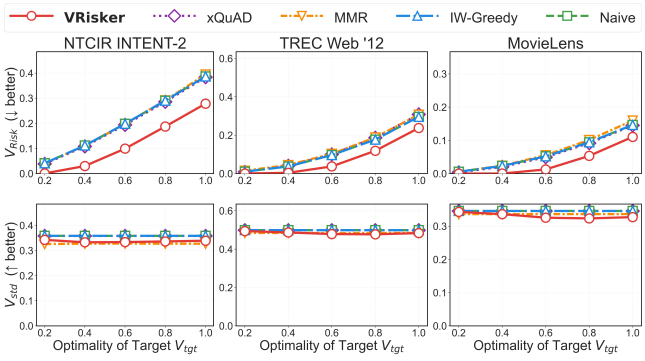

Figure 7: *VRisker works on various baseline target levels, and the optimality of it is a tunable parameter for adjusting safety.* Results shown in raw values.

A: VRisk/VRisker works on various baseline target levels, and target optimality provides a knob to trade off robustness and utility.

In our framework, $V_{\text{tgt}}(q, c)$ can be set to the oracle value or any baseline system. The latter is common in prior IR robustness literature [17, 49], but is now applied within a query.

We test how changing the target level (i.e., baseline system) changes the evaluation and performance. We do this by testing on different optimality of V_{tgt}, where a smaller value means less optimal. For example, if the optimality is 0.2, this means 20% of the oracle

$$V_{\text{tgt}}(q, c) = 0.2 \times \max_{R'_k} V(R'_k \mid q, c). \tag{18}$$

If VRisk=0.0 at target optimality 0.2, this implies that every intent attains at least 20% of its per-intent oracle value.

From Figure 7, we observe that VRisk is smaller as the target level is less optimal. This is simply because the target level is more achievable, resulting in lower loss. At optimality 0.2, we observe that VRisker achieves VRisk=0.0, where all intents satisfy at least 20% of the oracle value. Once perfect VRisk is achieved, VRisker could focus solely on raising the standard performance, because the IW tie-break takes over (see Algorithm 1). This can be observed in the bottom figures, where VRisker achieves near-optimal or optimal standard performance at 0.2. Moreover, in the Web'12 and MovieLens datasets, VRisker achieves perfect VRisk and near-optimal standard performance at 0.4 target optimality as well.

This is a clear and interesting trade-off. Lowering the target optimality can assure minimal satisfaction for all intents and also raise the standard performance. On the other hand, higher target optimality makes VRisker strictly safer and makes VRisk more pessimistic. Lastly, another interesting property is that when the target level is the oracle, minimization of loss is exactly the maximization of value. This property is especially relevant in production scenarios where service-level objectives are framed in terms of minimum per-intent quality guarantees.

By tuning V_{tgt} (and β), practitioners can directly express and enforce these guarantees while preserving flexibility for the remainder of the ranking.

Q: How fast is VRisker?

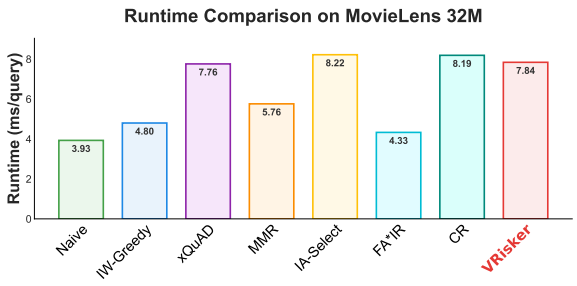

Figure 8: Average runtime on MovieLens 32M (ms/query). Per-query candidate document size is 71,933.

A: Runtime is comparable to standard diversification methods.

Figure 8 reports per-query runtime[5] to produce a top-10 ranking on MovieLens 32M (per-query candidate set size is $n = 71,933$). VRisker achieves 7.84 ms/query, which is within 1% of xQuAD, and faster than IA-SELECT and Calibrated Recommendation (CR). As expected, greedy diversifiers incur overhead relative to Naive, a simple expected-relevance sort, but the absolute latencies remain very small. These results indicate that VRisker matches the runtime profile of standard greedy diversifiers while delivering its robustness benefits.

[5]MacBook Pro (M2, 2022, 16 GB), single-threaded NumPy/BLAS.

Q: Do tie-breakers matter?

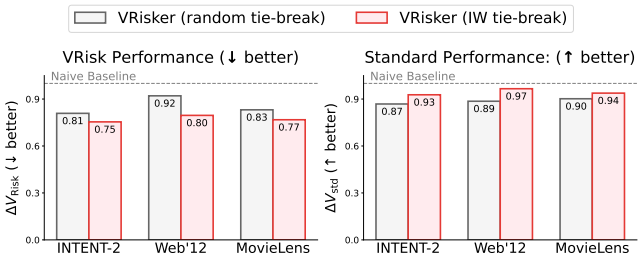

Figure 9: Tie-breaking ablation for VRisker.

A: IW tie-breaker improves both robustness and utility.

A tie occurs when multiple candidates yield identical VRisk decreases at a position. When there is a tie, VRisker picks the document that maximizes the IW metric (see Algorithm 1). Figure 9 compares the performance of VRisker with the IW tie-break versus a random tie-break. We observe that the IW tie-breaker reduced tail risk by 7–14% while improving average utility by 4–9% across INTENT-2, TREC Web'12, and MovieLens.

Q: Are existing diversification algorithms robust?

A: Generally, no.

In most experimental settings (Figures 3, 4, 5, 6, 7), existing diversification algorithms are no more robust than the naive ranking. This is because the optimization of IW-metrics is identical to the optimization of the standard metrics, as discussed in Section 3.4. In Figure 2, we observe that while some (non-linear) base metrics result in less risk, it is not consistent between datasets, and yet the risk is consistently larger than VRisker. We observe, as Chapelle et al. [6] remark, that IW-Greedy and xQuAD on ERR are fairly robust compared to the other metrics. Thus, if you really need to use an IW-based diversification, we suggest using it with ERR. However, in general, we recommend using VRisker, a tunable, intuitive, baseline-relative, and most importantly robust method.

An anonymized repository with code to reproduce all results and the appendix (w/ additional experimental results) is provided in https://github.com/anonymousIR26/wsdm-sup. The appendix provides results on raw values (not normalized values) with three additional baselines methods. Figure 6 is reported on all base metrics as well. Appendix B.3 also sweeps the weights of the prior methods to show that the observed results and discussions are not weight-dependent.

7 Limitations

A key limitation, that is shared with prior intent-aware diversification work [1, 9, 36, 37, 51], is the need to estimate intent probabilities $\Pr(c \mid q)$. While recent progress with LLMs makes intent discovery and labeling increasingly tractable [38, 42], such estimates can be

biased and typically require calibration against logs or human judgments. Encouragingly, our noise study (Figure 5) shows that VRisk remains comparatively robust under perturbed intent distributions.

Our formulation and prior work [1, 9, 36, 37, 51] also assume mutually exclusive intents per query, which simplifies analysis but may not capture overlapping or hierarchical intents. In addition, VRisk's behavior depends on two policy parameters: the target level $V_{\text{tgt}}(q, c)$ and the risk level β. We study both empirically (Fig. 7, Fig. 3), but different applications may prefer different settings.

8 Conclusion

This paper reframes diversification as *within-query risk minimization*. First, we showed mathematically and empirically that the most common diversification metrics favor majority intents just like standard metrics, prioritizing vulnerable rankings. To address this problem, we introduced **VRisk**, a CVaR-style, β-tunable metric that quantifies tail risk. VRisk is intuitive and has various properties that meet practitioner needs. Minimization of VRisk explicitly minimizes the chances of a user failing a search session. To minimize VRisk efficiently, we propose **VRisker**, a greedy optimizer with a $(1 - 1/e)$ guarantee for modular base metrics and a data-dependent bound for non-modular bases. Empirically, across INTENT-2, TREC Web'12, and MovieLens 32M, VRisker reduced tail risk by up to 33% with only $\sim$ 2% loss in average utility, while classic diversification often matches Naive ranker in robustness.

For future work, we plan to learn intents jointly, extend to session-level objectives, and integrate the idea of robustness in generative texts like question answering.

Ethical Considerations

This work adheres to established ethical standards for research. All evaluations were performed on publicly available datasets containing no personally identifiable information. The proposed methods are intended to enhance user experience by improving diversity and coverage in search and recommendation results.

References

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (Barcelona, Spain) *(WSDM '09)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/1498759.1498766

[2] Andrew An Bian, Joachim M. Buhmann, Andreas Krause, and Sebastian Tschiatschek. 2019. Guarantees for Greedy Maximization of Non-submodular Functions with Applications. arXiv:1703.02100 [cs.DM] https://arxiv.org/abs/1703.02100

[3] Christina Brandt, Thorsten Joachims, Yisong Yue, and Jacob Bank. 2011. Dynamic ranked retrieval. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11)*. Association for Computing Machinery, New York, NY, USA, 247–256. https://doi.org/10.1145/1935826.1935872

[4] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 335–336. https://doi.org/10.1145/290941.291025

[5] Bo Chang, Alexandros Karatzoglou, Yuyan Wang, Can Xu, Ed H. Chi, and Minmin Chen. 2023. Latent User Intent Modeling for Sequential Recommenders. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*. 427–431. https://doi.org/10.1145/3543873.3584641

[6] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. 2011. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval* 14, 6 (Dec. 2011), 572–592.

[7] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[8] Charles L.A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11)*. Association for Computing Machinery, New York, NY, USA, 75–84. https://doi.org/10.1145/1935826.1935847

[9] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[10] Charles L. Clarke, Maheedhar Kolla, and Olga Vechtomova. 2009. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory* (Cambridge, UK) *(ICTIR '09)*. Springer-Verlag, Berlin, Heidelberg, 188–199. https://doi.org/10.1007/978-3-642-04417-5_17

[11] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. 2012. Overview of the TREC 2012 Web Track. In *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012) (NIST Special Publication, Vol. 500-298)*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. http://trec.nist.gov/pubs/trec21/papers/WEB12.overview.pdf

[12] Kevyn Collins-Thompson. 2009. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 837–846. https://doi.org/10.1145/1645953.1646059

[13] Abhimanyu Das and David Kempe. 2011. Submodular meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection. arXiv:1102.3975 [stat.ML] https://arxiv.org/abs/1102.3975

[14] Fernando Diaz. 2024. Pessimistic Evaluation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2024)*. ACM, 115–124. https://doi.org/10.1145/3673791.3698428

[15] Fernando Diaz, Michael D. Ekstrand, and Bhaskar Mitra. 2024. Recall, Robustness, and Lexicographic Evaluation. arXiv:2302.11370 [cs.IR] https://arxiv.org/abs/2302.11370

[16] 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. https://doi.org/10.1145/2600428.2609625

[17] B. Taner Dinçer, Craig Macdonald, and Iadh Ounis. 2016. Risk-Sensitive Evaluation and Learning to Rank using Multiple Baselines. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 483–492. https://doi.org/10.1145/2911451.2911511

[18] Yingqiang Ge, Shuyuan Xu, Shuchang Liu, Zuohui Fu, Fei Sun, and Yongfeng Zhang. 2020. Learning Personalized Risk Preferences for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, 409–418.

[19] Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. 2023. Safe Deployment for Counterfactual Learning to Rank with Exposure-Based Risk Minimization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. ACM, 249–258.

[20] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[21] Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. 2021. Off-Policy Risk Assessment in Contextual Bandits. arXiv:2104.08977 [cs.LG] https://arxiv.org/abs/2104.08977

[22] Dietmar Jannach and Markus Zanker. 2024. A Survey on Intent-aware Recommender Systems. *Comput. Surveys* (2024). https://doi.org/10.1145/3700890 arXiv:2406.16350.

[23] Madian Khabsa, Aidan Crook, Ahmed Hassan Awadallah, Imed Zitouni, Tasos Anastasakos, and Kyle Williams. 2016. Learning to Account for Good Abandonment in Search Success Metrics. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 1893–1896. https://doi.org/10.1145/2983323.2983867

[24] Haruka Kiyohara, Ren Kishimoto, Kosuke Kawakami, Ken Kobayashi, Kazuhide Nakata, and Yuta Saito. 2024. Towards Assessing and Benchmarking Risk-Return

Tradeoff of Off-Policy Evaluation. arXiv:2311.18207 [cs.LG] https://arxiv.org/abs/2311.18207

[25] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good abandonment in mobile and PC internet search. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 43–50. https://doi.org/10.1145/1571941.1571951

[26] Jiayu Li, Peijie Sun, Zhefan Wang, Weizhi Ma, Yangkun Li, Min Zhang, Zhoutian Feng, and Daiyue Xue. 2023. Intent-aware Ranking Ensemble for Personalized Recommendation. In *Proceedings of the 46th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. 1004–1013. https://doi.org/10.1145/3539618.3591702

[27] Maria Maistro, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. 2021. Principled Multi-Aspect Evaluation Measures of Rankings. 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, 1232–1242. https://doi.org/10.1145/3459637.3482287

[28] Omid Memarrast, Ashkan Rezaei, Rizal Fathony, and Brian Ziebart. 2021. Fairness for Robust Learning to Rank. arXiv:2112.06288 [cs.LG] https://arxiv.org/abs/2112.06288

[29] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (Dec. 2008), 27 pages. https://doi.org/10.1145/1416950.1416952

[30] Debmalya Panigrahi, Atish Das Sarma, Gagan Aggarwal, and Andrew Tomkins. 2012. Online selection of diverse results. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. Association for Computing Machinery, New York, NY, USA, 263–272. https://doi.org/10.1145/2124295.2124329

[31] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. 2008. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning* (Helsinki, Finland) *(ICML '08)*. Association for Computing Machinery, New York, NY, USA, 784–791. https://doi.org/10.1145/1390156.1390255

[32] Stephen Robertson. 2006. On GMAP: and other transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (Arlington, Virginia, USA) *(CIKM '06)*. Association for Computing Machinery, New York, NY, USA, 78–83. https://doi.org/10.1145/1183614.1183630

[33] R. Tyrrell Rockafellar and Stanislav Uryasev. 2000. Optimization of conditional value-at risk. *Journal of Risk* 3 (2000), 21–41. https://api.semanticscholar.org/CorpusID:854622

[34] Tetsuya Sakai, Zhicheng Dou, Takehiro Yamamoto, Yiqun Liu, Min Zhang, Makoto P. Kato, Ruihua Song, and Mayu Iwata. 2013. Summary of the NTCIR-10 INTENT-2 task: subtopic mining and search result diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 761–764. https://doi.org/10.1145/2484028.2484104

[35] Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2022. A Versatile Framework for Evaluating Ranked Lists in terms of Group Fairness and Relevance. arXiv:2204.00280 [cs.IR] https://arxiv.org/abs/2204.00280

[36] Tetsuya Sakai and Ruihua Song. 2011. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 1043–1052. https://doi.org/10.1145/2009916.2010055

[37] Rodrygo L. T. Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. 2010. Explicit search result diversification through sub-queries. In *Proceedings of the 32nd European Conference on Advances in Information Retrieval* (Milton Keynes, UK) *(ECIR'2010)*. Springer-Verlag, Berlin, Heidelberg, 87–99. https://doi.org/10.1007/978-3-642-12275-0_11

[38] Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Snigdha Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, Nagu Rangan, Tara Safavi, Siddharth Suri, Mengting Wan, Leijie Wang, and Longqi Yang. 2024. Using Large Language Models to Generate, Validate, and Apply User Intent Taxonomies. arXiv:2309.13063 [cs.IR] https://arxiv.org/abs/2309.13063

[39] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining (KDD '18)*. ACM, 2219–2228. https://doi.org/10.1145/3219819.3220088

[40] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 154–162. https://doi.org/10.1145/3240323.3240372

[41] Zhu Sun, Hongyang Liu, Xinghua Qu, Kaidong Feng, Yan Wang, and Yew-Soon Ong. 2024. Large Language Models for Intent-Driven Session Recommendations. In *Proceedings of the 47th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 324–334. https://doi.org/10.1145/3626772.3657688

[42] Rikiya Takehi, Ellen M. Voorhees, Tetsuya Sakai, and Ian Soboroff. 2025. LLM-Assisted Relevance Assessments: When Should We Ask LLMs for Help?. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 95–105. https://doi.org/10.1145/3726302.3729916

[43] Riku Togashi, Tatsushi Oka, Naoto Ohsaka, and Tetsuro Morimura. 2024. Safe Collaborative Filtering. arXiv:2306.05292 [cs.IR] https://arxiv.org/abs/2306.05292

[44] Ming-Feng Tsai and Chuan-Ju Wang. 2012. Post-Modern Portfolio Theory for Information Retrieval. In *Proceedings of the 35th European Conference on Information Retrieval*. 80–85.

[45] Saul Vargas, Pablo Castells, and David Vallet. 2011. Intent-oriented Diversity in Recommender Systems. In *Proceedings of the 34th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. 1211–1212. https://doi.org/10.1145/2009916.2010124

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

[47] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, 515–524. https://doi.org/10.1145/3077136.3080786

[48] Jun Wang and Jianhan Zhu. 2009. Portfolio Theory of Information Retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 115–122. https://doi.org/10.1145/1571941.1571963

[49] 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. https://doi.org/10.1145/2348283.2348385

[50] Xiaojie Wang, Ji-Rong Wen, Zhicheng Dou, Tetsuya Sakai, and Rui Zhang. 2018. Search Result Diversity Evaluation Based on Intent Hierarchies. *IEEE Transactions on Knowledge and Data Engineering* 30, 1 (2018), 156–169. https://doi.org/10.1109/TKDE.2017.2729559

[51] Yuyan Wang, Cheenar Banerjee, Samer Chucri, Fabio Soldo, Sriraj Badam, Ed H. Chi, and Minmin Chen. 2025. Beyond Item Dissimilarities: Diversifying by Intent in Recommender Systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25)*. ACM, 2672–2681. https://doi.org/10.1145/3690624.3709429

[52] Yuling Wang, Xiao Wang, Xiangzhou Huang, Yanhua Yu, Haoyang Li, Mengdi Zhang, Zirui Guo, and Wei Wu. 2023. Intent-aware Recommendation via Disentangled Graph Contrastive Learning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI '23)*. 2343–2351. https://doi.org/10.24963/ijcai.2023/260

[53] Michael J. Welch, Junghoo Cho, and Christopher Olston. 2011. Search result diversity for informational queries. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/1963405.1963441

[54] Hongyi Wen, Xinyang Yi, Tiansheng Yao, Jiaxi Tang, Lichan Hong, and Ed H. Chi. 2022. Distributionally-robust Recommendations for Improving Worst-case User Experience. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3606–3610. https://doi.org/10.1145/3485447.3512255

[55] Ryen W. White and Susan T. Dumais. 2009. Characterizing and predicting search engine switching behavior. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 87–96. https://doi.org/10.1145/1645953.1645967

[56] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, 1569–1578. https://doi.org/10.1145/3132847.3132938

[57] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. 2024. Fairness and Diversity in Recommender Systems: A Survey. arXiv:2307.04644 [cs.IR] https://arxiv.org/abs/2307.04644

A Proofs

A.1 Proof of Proposition 2: NP-hardness

Proof. We reduce *Weighted Max-k-Cover* to VRisk minimization (with k part of the input).[6]

Given ground set U, weights $w : U \to \mathbb{R}_{\geq 0}$, family $\{S_1, \ldots, S_n\} \subseteq 2^U$, and budget k, build a single-query instance: intents $C(q) = U$ with

$$\Pr(u \mid q) = \frac{w(u)}{W}, \quad W = \sum_{u \in U} w(u).$$

Create one document d_j per set S_j and set binary relevance

$$\mathrm{rel}(d_j \mid q, u) = \mathbf{1}[u \in S_j].$$

Use the modular base metric and the target $V_{\mathrm{tgt}}(q, u) = 1/k$; fix $\beta = 1$.

For any length-k ranking R,

$$V(R \mid q, u) = \frac{1}{k} \sum_{d \in R} \mathbf{1}[u \in S(d)],$$

$$\ell(R, q, u) = \max\left\{0, \frac{1}{k} - V(R \mid q, u)\right\} = \frac{1}{k}\,\mathbf{1}\left[u \notin \bigcup_{d \in R} S(d)\right].$$

Thus

$$\begin{aligned} V_{\mathrm{Risk}}(R \mid q; 1) &= \sum_u \Pr(u \mid q)\, \ell(R, q, u) \\ &= \frac{1}{k}\left(1 - \frac{1}{W} \sum_{u \in \cup_{d \in R} S(d)} w(u)\right). \end{aligned}$$

Minimizing $V_{\mathrm{Risk}}(R \mid q; 1)$ is equivalent to maximizing the covered weight in Weighted Max-k-Cover. Hence VRisk minimization is NP-hard. □

A.2 Proof of Theorem 1: $(1 - 1/e)$-Optimality Guarantee

Proof of Theorem 1. Let

$$\mathrm{VRisk}(R) := V_{\mathrm{Risk}}(R \mid q; \beta) = \min_{\zeta \in R}\left[\zeta + \frac{1}{\beta} \sum_i \Pr(c_i \mid q)\, (\ell_i(R) - \zeta)_+\right].$$

For the discrete nonnegative losses $\ell_i(R)$, a minimizer $\zeta^\star$ always lies in $[0, \max_i \ell_i(R)]$, hence $\zeta^\star \geq 0$. For any fixed $\zeta \geq 0$ define

$$\begin{aligned} H_\zeta(R) &:= \zeta + \frac{1}{\beta} \sum_i p_i\, (\ell_i(R) - \zeta)_+, \\ p_i &:= \Pr(c_i \mid q). \end{aligned}$$

Then $\mathrm{VRisk}(R) = \min_\zeta H_\zeta(R)$.

Step 1. Fix $\zeta \geq 0$ and intent i. Write $C_i := V_{\mathrm{tgt}}(q, c_i) - \zeta \geq 0$ and $s_i(R) = \sum_{d \in R} v_i(d)$ with $v_i(d) := \mathrm{rel}(d \mid q, c_i)/k \geq 0$. For $e \notin R$,

$$\begin{aligned} \Delta_i^{(\zeta)}(R; e) &:= \big(C_i - s_i(R)\big)_+ - \big(C_i - s_i(R) - v_i(e)\big)_+ \\ &= \max\{0, \min\{v_i(e),\, C_i - s_i(R)\}\}. \end{aligned}$$

Hence if $R \subseteq S$ then $s_i(R) \leq s_i(S)$ and thus $\Delta_i^{(\zeta)}(R; e) \geq \Delta_i^{(\zeta)}(S; e)$. Summing with nonnegative weights p_i/β gives, for all $R \subseteq S$ and $e \notin S$,

$$H_\zeta(R) - H_\zeta(R \cup \{e\}) \;\geq\; H_\zeta(S) - H_\zeta(S \cup \{e\}). \qquad (*)$$

That is, the *risk drop at fixed* ζ has diminishing returns.

Step 2 (one-step greedy progress). Let R_t be the greedy set after t steps and let $\zeta_t \in \arg\min_\zeta H_\zeta(R_t)$. For any $e \notin R_t$,

$$\mathrm{VRisk}(R_t) - \mathrm{VRisk}(R_t \cup \{e\}) \;\geq\; H_{\zeta_t}(R_t) - H_{\zeta_t}(R_t \cup \{e\})$$

since $\mathrm{VRisk}(R_t) = H_{\zeta_t}(R_t)$ and $\mathrm{VRisk}(R_t \cup \{e\}) \leq H_{\zeta_t}(R_t \cup \{e\})$. Let O be an optimal k-set. Using $(*)$ and averaging over $e \in O$,

$$\begin{aligned} \max_{e \notin R_t}\big[H_{\zeta_t}(R_t) - H_{\zeta_t}(R_t \cup \{e\})\big] &\;\geq\; \frac{1}{k} \sum_{e \in O}\big[H_{\zeta_t}(R_t) - H_{\zeta_t}(R_t \cup \{e\})\big] \\ &\;\geq\; \frac{1}{k}\Big(H_{\zeta_t}(R_t) - H_{\zeta_t}(R_t \cup O)\Big). \end{aligned}$$

Monotonicity of H_{ζ_t} in R gives $H_{\zeta_t}(R_t \cup O) \leq H_{\zeta_t}(O)$, and by definition $\mathrm{VRisk}(O) = \min_\zeta H_\zeta(O) \leq H_{\zeta_t}(O)$. Therefore

$$H_{\zeta_t}(R_t) - H_{\zeta_t}(R_t \cup O) \;\geq\; H_{\zeta_t}(R_t) - H_{\zeta_t}(O) \;\geq\; \mathrm{VRisk}(R_t) - \mathrm{VRisk}(O).$$

Combining the displays and taking greedy e_{t+1},

$$\mathrm{VRisk}(R_t) - \mathrm{VRisk}(R_{t+1}) \;\geq\; \frac{1}{k}\big(\mathrm{VRisk}(R_t) - \mathrm{VRisk}(O)\big).$$

Step 3. Let $\Delta_t := \mathrm{VRisk}(R_t) - \mathrm{VRisk}(O)$. Then

$$\Delta_{t+1} \leq \left(1 - \frac{1}{k}\right)\Delta_t,$$

so

$$\Delta_k \leq \left(1 - \frac{1}{k}\right)^k \Delta_0 \leq e^{-1}\Delta_0.$$

Equivalently,

$$\mathrm{VRisk}(\varnothing) - \mathrm{VRisk}(R_k) \;\geq\; \left(1 - \frac{1}{e}\right)\big(\mathrm{VRisk}(\varnothing) - \mathrm{VRisk}(O)\big),$$

i.e., the greedy R_k captures at least $(1 - 1/e)$ of the optimal risk reduction. □

A.3 Proof of Theorem 2: NDCG-Risk Approximation

Proof of Theorem 2. Let

$$F(R) = V_{\mathrm{Risk}}(\varnothing) - V_{\mathrm{Risk}}(R) = \Delta(R)$$

be the *risk–reduction* set function. F is *monotone* because adding a document can only lower V_{Risk}.

Step 1: submodularity ratio. For any two prefixes $A \subseteq B \subseteq \mathcal{D}$ and any document $d \notin B$ define the marginal gains

$$\Delta(d \mid X) = F(X \cup \{d\}) - F(X), \quad X \in \{A, B\}.$$

Write $|A| = r$, $|B| = t$ $(r < t < k)$. With the NDCG discount vector $w_1 \geq \cdots \geq w_k$, d is placed at position $r + 1$ in $A \cup \{d\}$ and $t + 1$ in $B \cup \{d\}$. For each intent c the per–intent gain satisfies

$$\mathrm{NDCG}_c(A \cup \{d\}) - \mathrm{NDCG}_c(A) = \frac{w_{r+1}\, g_{d,c}}{\mathrm{IDCG}_c},$$

$$\mathrm{NDCG}_c(B \cup \{d\}) - \mathrm{NDCG}_c(B) = \frac{w_{t+1}\, g_{d,c}}{\mathrm{IDCG}_c},$$

[6] For fixed constant k, brute force $O(n^k)$ is polynomial.

where $g_{d,c} = 2^{\mathrm{rel}(d|q,c)} - 1 \geq 0$. Because V_{Risk} is a positive convex combination of these per–intent gains clipped by a hinge at ζ, the clipping can *only reduce* both numerators *by the same amount*. Hence

$$\begin{aligned}\Delta(d \mid B) &\geq \frac{w_{t+1}}{w_{r+1}}\,\Delta(d \mid A)\\ &\geq \frac{w_k}{w_1}\,\Delta(d \mid A)\\ &\geq \underbrace{\frac{w_k}{\sum_{i=1}^{k} w_i}}_{=\gamma}\,\Delta(d \mid A).\end{aligned}$$

Thus the *submodularity ratio* [13] of F is lower-bounded by γ.

Step 2: greedy approximation. For any monotone set function with submodularity ratio γ, the standard greedy algorithm attains the guarantee

$$F(R_k) \geq (1 - e^{-\gamma})F(R_k^{\star})$$

under a cardinality constraint k. Applying this to $F = \Delta$ proves Theorem 2. □

B Additional Experiments and Results

B.1 Statistical Significance Testing

Following IR best practice, we treat the set of queries in each benchmark as a sample from a larger population and test the null hypothesis that two systems have equal *expected* performance across that population. For every query q we compute the paired difference $d_q = M_A(q) - M_B(q)$ where M is either our risk metric $V_{\mathrm{Risk}}(R_k|q;\beta = 0.10)$ or the underlying expected-utility metric V_{std}. We then apply (i) a two-sided Wilcoxon signed–rank test and (ii) a paired randomization test with $B = 100{,}000$ permutations.

Table 3: Paired significance tests at $\beta = 0.10$, $k = 10$. Asterisks mark values that remain below $\alpha = 0.05$ after Holm–Bonferroni correction over all comparisons.

Dataset	Metric	Wilcoxon p		Perm. p	
		vs Naive	vs xQuAD	vs Naive	vs xQuAD
INTENT-2	V_{Risk}	$1.1 \times 10^{-13^{*}}$	$1.8 \times 10^{-13^{*}}$	$< 10^{-5^{*}}$	$< 10^{-5^{*}}$
	V_{std}	$1.2 \times 10^{-11^{*}}$	$1.2 \times 10^{-11^{*}}$	$< 10^{-5^{*}}$	$< 10^{-5^{*}}$
WEB	V_{Risk}	$7.7 \times 10^{-6^{*}}$	$1.4 \times 10^{-6^{*}}$	$< 10^{-5^{*}}$	$1.0 \times 10^{-5^{*}}$
	V_{std}	$5.8 \times 10^{-9^{*}}$	$5.8 \times 10^{-9^{*}}$	$< 10^{-5^{*}}$	$< 10^{-5^{*}}$
ML 32M	V_{Risk}	$7.5 \times 10^{-65^{*}}$	$9.6 \times 10^{-65^{*}}$	$< 10^{-5^{*}}$	$< 10^{-5^{*}}$
	V_{std}	$2.6 \times 10^{-10^{*}}$	1.4×10^{-1}	$< 10^{-5^{*}}$	$< 10^{-5^{*}}$

Table 3 reports the resulting p-values. Asterisks mark values that remain below $\alpha = 0.05$ after Holm–Bonferroni correction over all comparisons.

B.2 Additional Results on Optimality

Figure 10 provides the full list of experiments on the optimality of VRisker (extended version of Figure 6 in the main text). We observe that VRisker is nearly optimal on all metrics. We also observe that the optimal performance, in return, lowers the standard performance on all metrics. This demonstrates that VRisker is robust while being strong in standard performance.

B.3 Sensitivity Analysis on Prior Methods

xQuAD, MMR, FA*IR, and CR all have a weight that balances between relevance and diversity/fairness/calibration. Although they have different notations for weights, we generalize as $\lambda_{\mathrm{method}}$ in our paper. Essentially, using the weight, they optimize below

$$(1 - \lambda_{\mathrm{method}})V_{\mathrm{std}}(d \mid q, R) \;+\; \lambda_{\mathrm{method}}V_{\mathrm{X}}(d \mid q, R).$$

Thus λ=0 reduces to pure relevance (Naive/IW), and λ=1 to the pure diversity/fairness/calibration.

Figure 11 shows the sensitivity experiment on the same settings as the main text.

Across NTCIR INTENT-2, TREC Web'12, and MovieLens, sweeping λ confirms a consistent pattern: (i) at λ=1 all classical baselines become diversity/calibration-only and suffer catastrophic drops in both V_{std} (=V_{IW}) and V_{Risk}, (ii) for $\lambda \in [0, 0.8]$ they track Naive on both risk and utility, offering little worst-case protection, and (iii) VRisker dominates on $V_{\mathrm{Risk},k}$ while maintaining high V_{std}.

This justifies the fixed λ choice in the main text and shows that the conclusions are not an artifact of particular hyperparameters.

B.4 Additional Results

In this appendix, we provide experiment results on the main text but with more baselines and with the raw metric values if they are normalized on the main text.

The additional baselines are:

IA-SELECT [1]. This method builds a ranking that minimizes the chances that users will not click on any item in a ranking. Since IA-SELECT works best on a cascade assumption with click probabilities instead of relevance, we normalize the relevance scores to $[0, 1]$ when computing.

Calibrated Recommendation (CR) [40]. This method aims to match the intent-probability proportions to the per-intent utility proportions in the ranking. λ_{CR} is the weight which controls the balance between relevance and calibration, and is set to 0.5.

FA*IR [56]. This method aims to fairly expose each group of documents in the ranking. In our case, we target intent-probability distribution as a fairness objective, trading off between relevance and fairness via the weight $\lambda_{\mathrm{FA*IR}}$ (set to 0.5).

Figures 12, 13, 14, 15, and 16 report the additional experimental results Figures 2, 3, 4, 5, and 7 in the main text respectively.

We observe that the additional baselines are not robust or equal as robust as the naive baseline. This is mainly because the settings and the objectives of the methods are conceptually different from our objective. IA-SELECT only functions under the use of click probabilities under the cascade user assumption. FA*IR and CR respectively focus on fairness and calibration, which are different from diversity and robustness.

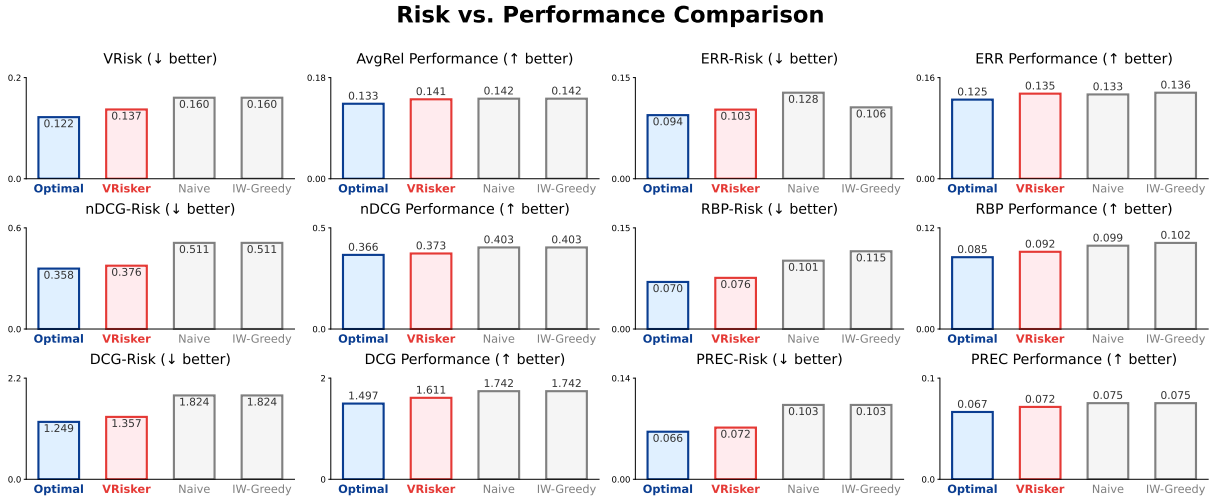

Figure 10: Optimality of VRisker. For each base metric, the left shows VRisk (lower is better) and the right shows V_{IW} (higher is better). Figure 6 of the main text.

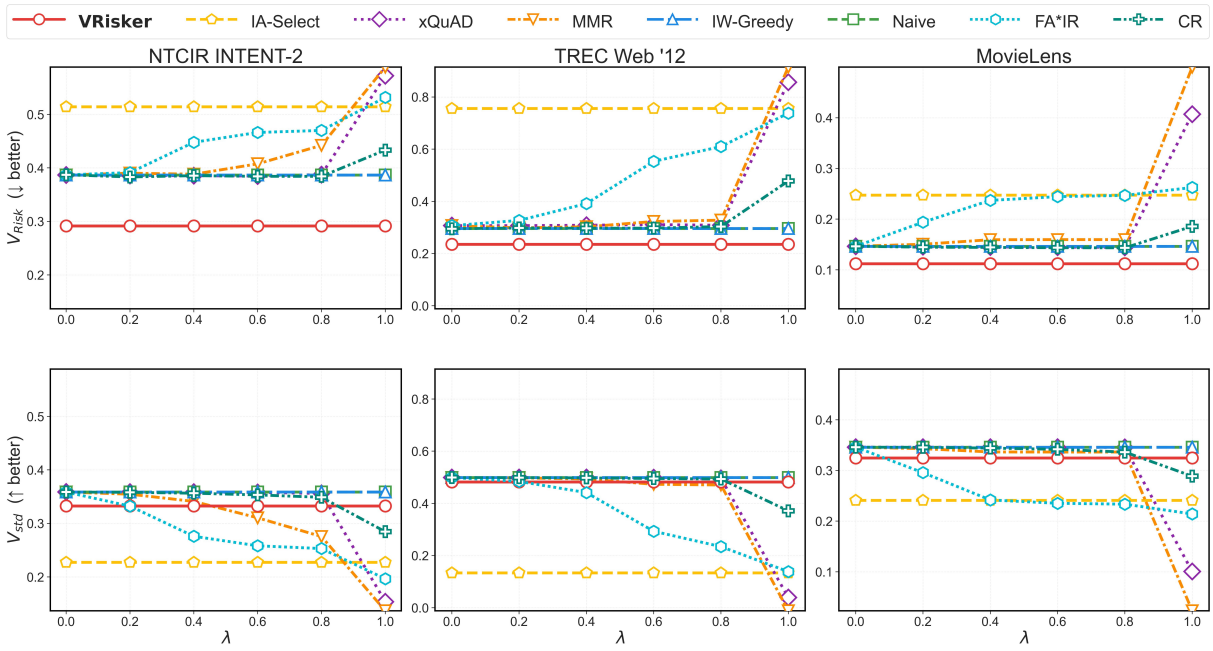

Figure 11: λ-sensitivity (xQuAD, MMR, FA*IR, CR) on three datasets. As $\lambda \to 1$ (pure diversity/calibration), all classical baselines *collapse* in V_{std} and do not provide meaningful tail-risk reduction; for intermediate λ they behave similarly to Naive. In contrast, VRisker (no λ) remains consistently robust with small utility loss.

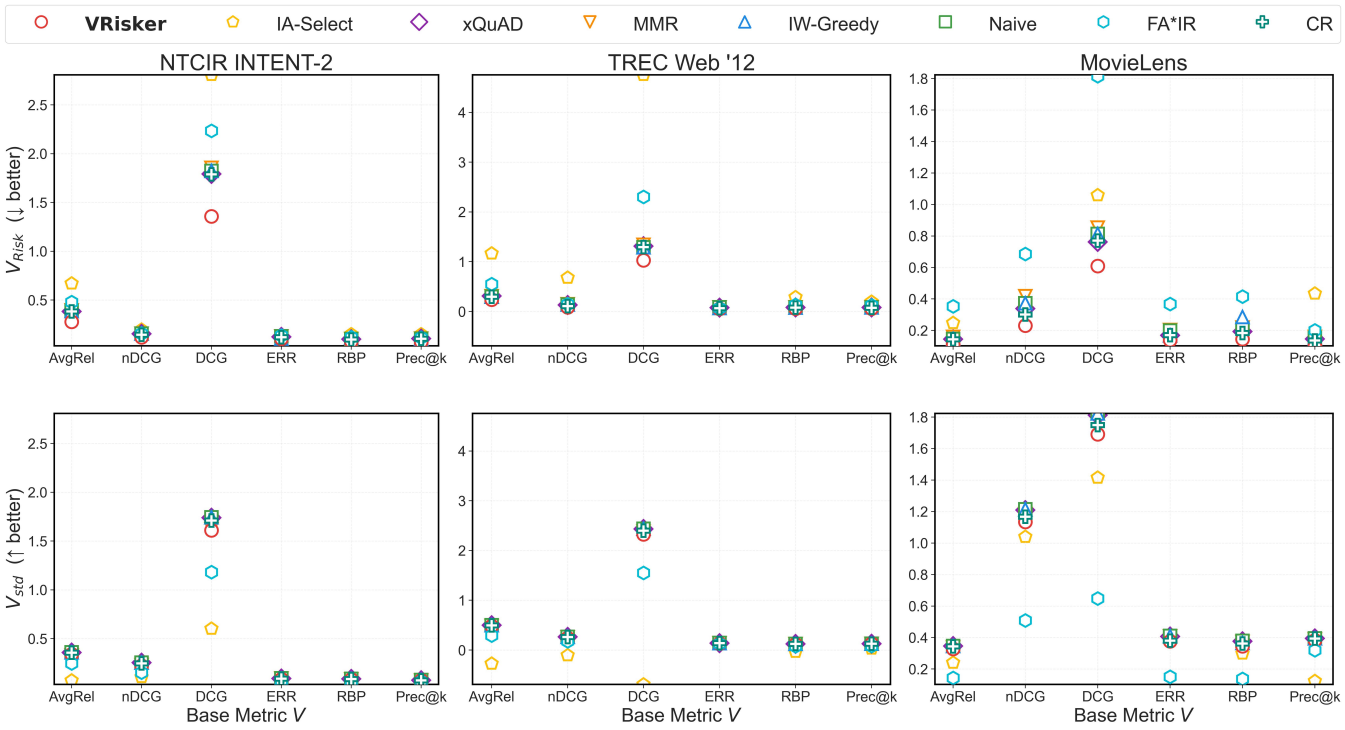

Figure 12: Risk and utility across base metrics. $\Delta V_{\mathbf{Risk}}$ (lower is better) and $\Delta V_{\mathbf{std}}/\Delta V_{\mathbf{IW}}$ (higher is better) for AvgRel, nDCG, DCG, ERR, RBP, and Prec@k on INTENT-2, WEB'12, and MovieLens; Naive = 100%. Figure 2 of the main text.

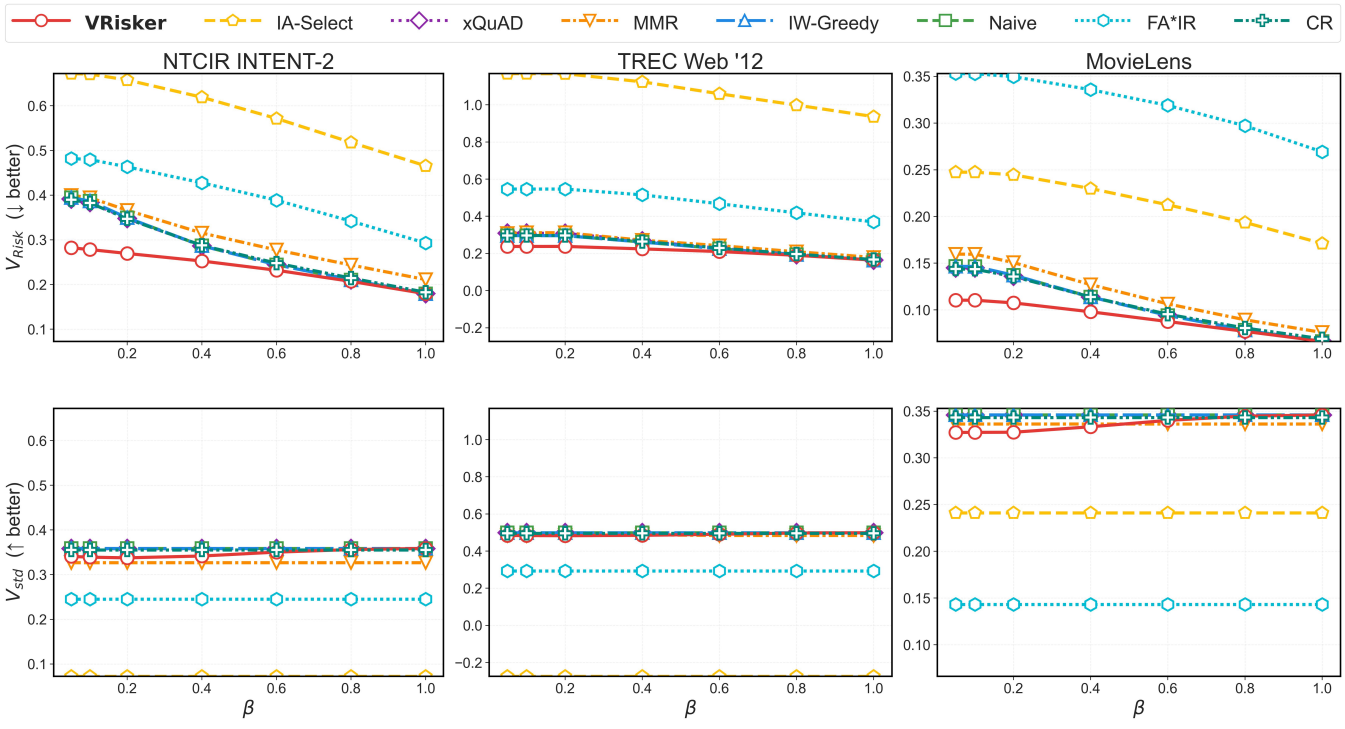

Figure 13: Effect of β (pessimism level). As β decreases, VRisker reduces tail loss more aggressively with a manageable drop in average utility. Figure 3 of the main text.

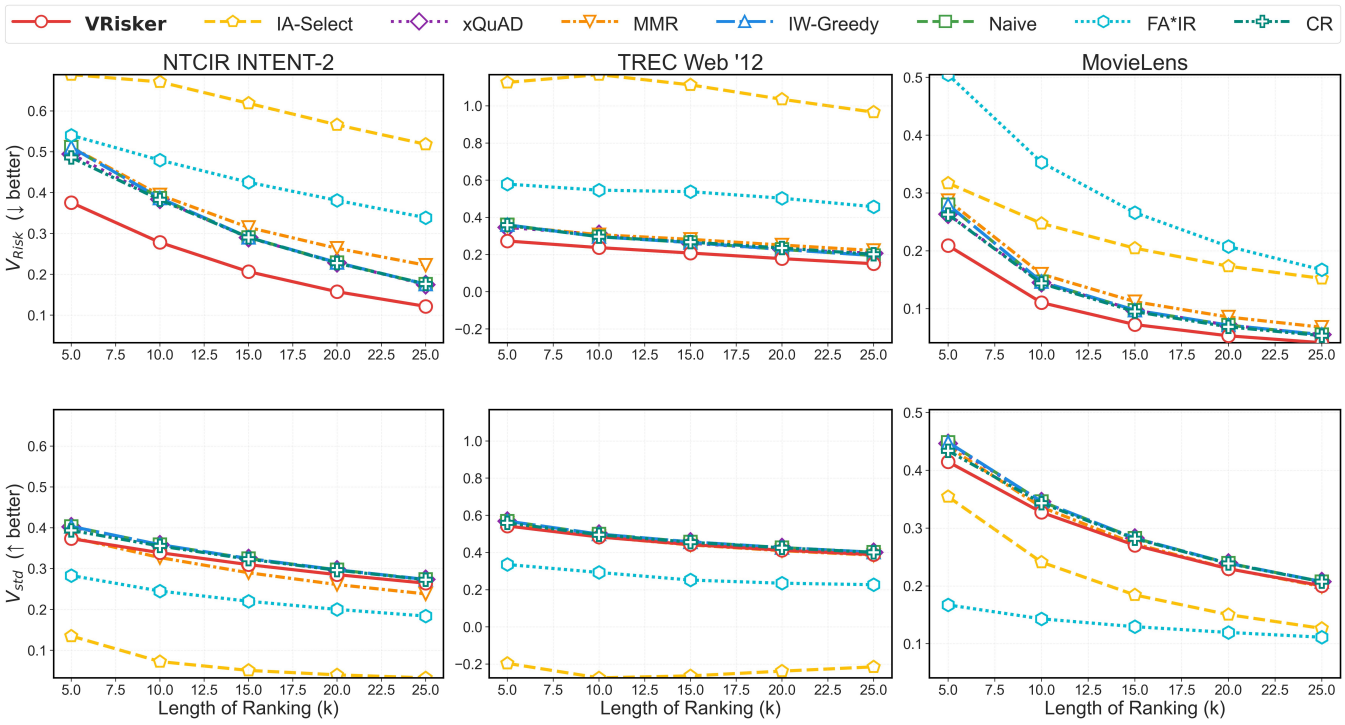

Figure 14: Effect of ranking length k. VRisker's utility improves as k grows while maintaining substantially lower tail risk than baselines. Figure 4 of the main text.

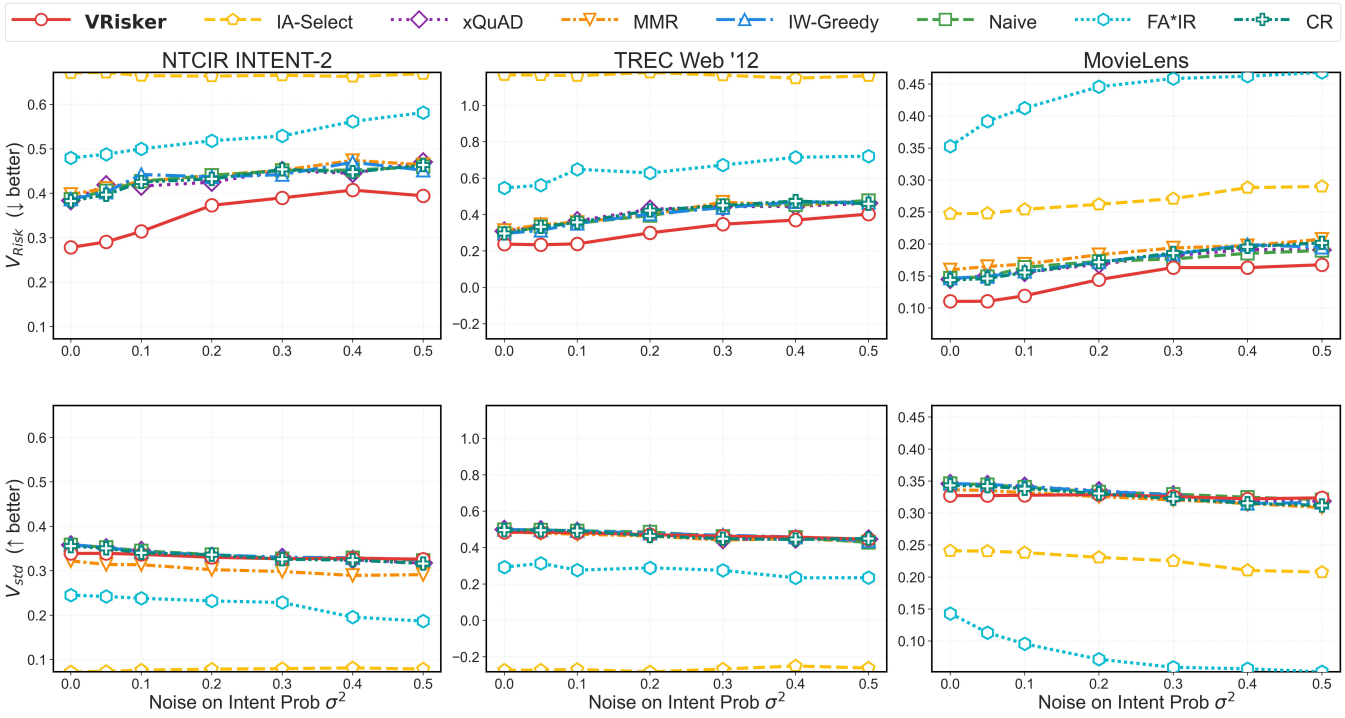

Figure 15: Noise in intent probabilities. VRisker remains most robust under Gaussian noise added to $Pr(c|q)$. Figure 5 of the main text.

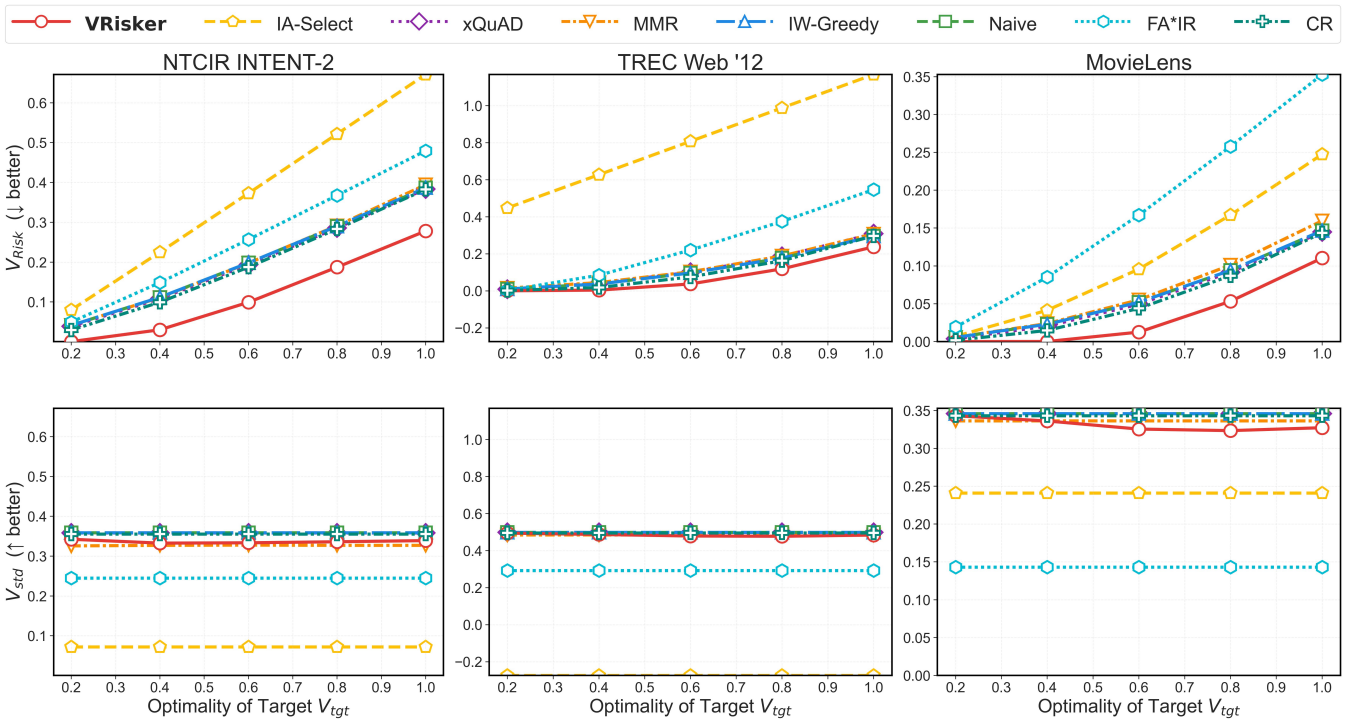

Figure 16: Target optimality sweep. Lower target optimality yields smaller losses and lets VRisker focus on utility once $V_{\mathbf{Risk}} \approx 0$. Figure 7 of the main text.