# Refairmulate: A Large-Scale Dataset for Gender-Fair Query Reformulations

Hai Son Le
Toronto Metropolitan University
Toronto, Canada

Shirin Seyedsalehi
Toronto Metropolitan University
Toronto, Canada

Morteza Zihayat
Toronto Metropolitan University
Toronto, Canada

Ebrahim Bagheri
University of Toronto
Toronto, Canada

## Abstract

Information Retrieval systems can amplify societal inequalities when training data and ranking algorithms encode biases toward certain gender identities. Query reformulation methods are known to improve retrieval effectiveness, yet they seldom treat retrieval fairness as a first-class criterion. We introduce `Refairmulate`, an open resource that supports *training and benchmarking* gender-fair query reformulation models through a multi-objective procedure that jointly balances retrieval effectiveness and gender bias. The resource contains three dataset subsets. The *Optimal* subset consists of 112,261 query pairs, where each pair includes an original query and a reformulated variant that attains perfect RR@10 (=1) alongside a complete reduction of measured gender bias. The *Effective* subset includes 209,343 query pairs, where the reformulated variant is guaranteed to achieve both higher retrieval effectiveness and lower gender bias than the original query, enabling direct supervision for joint optimization. The *Fair* subset contains 321,604 query pairs and guarantees reduced gender bias regardless of changes in retrieval effectiveness, supporting fairness-focused training and analysis. We benchmark `Refairmulate` with BM25 and several dense retrievers, including SPLADE, SBERT, TCT-ColBERT, and ANCE, all of which show consistent gains, with up to 76.0% relative improvement in MRR@10 and up to 48.5% reduction in gender bias. To our knowledge, `Refairmulate` is the first dataset explicitly designed for fairness-aware query reformulation, providing large-scale training and benchmarking data for the community.

## 1 Introduction

Information Retrieval (IR) systems are central to how individuals retrieve online content, supporting critical applications in search engines, recommendation systems, and digital assistants. Yet, despite their widespread adoption, these systems risk reinforcing and amplifying societal inequalities, such as *gender bias* as the focus of this resource paper, due to entrenched patterns in training data and algorithmic design [4, 18, 21]. Gender bias in IR exhibits itself when ostensibly neutral queries return results disproportionately associated with a particular gender, thereby perpetuating stereotypes and limiting representational fairness [18].

Addressing such biases requires not only interventions at the data and algorithmic levels but also at the query formulation stage, where user intent interacts most directly with retrieval models. In this context, query reformulation methods through which user queries are rewritten or expanded have emerged as a powerful mechanism for enhancing search performance [12, 19, 26]. However, despite demonstrated improvements in traditional retrieval effectiveness metrics such as Mean Reciprocal Rank (MRR) and Average Precision (AP), these methods largely disregard fairness. As a result, reformulation methods are typically trained and evaluated as if "better queries" are defined solely by relevance, even though reformulations can systematically shift the gender composition and representational balance of retrieved results. This limitation is not merely incidental but stems from methodological oversights in current query reformulation approaches [1, 22, 27]. Specifically, (1) training corpora used for reformulation often encode gendered biases that models inadvertently learn and reproduce [21]; (2) retrieval effectiveness is optimized in isolation, sidelining fairness as a competing or complementary objective [1]; and (3) no standard evaluation frameworks or benchmarks currently exist to assess or enforce fairness constraints in query reformulation.

This situation presents both a practical and theoretical gap, namely the absence of large-scale, systematically constructed datasets that support fairness-aware query reformulation, which not only limits the development of socially responsible IR models but also restricts reproducibility and comparative evaluation. Without such a resource, the community lacks the empirical foundations to study and possibly mitigate gender bias in query reformulation approaches. Importantly, it also prevents training reformulation models with a data that explicitly reflects the intended objective of producing reformulations that are simultaneously more effective and less biased.

To address this, we introduce `Refairmulate`, an open-source toolkit and resource suite that directly responds to this critical need. `Refairmulate` includes three large-scale datasets, *Optimal*, *Effective*, and *Fair*, which contain 112,000 to 321,604 query pairs specifically designed to enable the joint optimization of fairness and retrieval effectiveness. In addition to serving as a benchmark for evaluation, `Refairmulate` is constructed to function as **training data** for supervised and contrastive learning setups (e.g., cross-encoders and rerankers over candidate reformulations), where the training data consists of samples that carry effectiveness improvements along with measurable bias reduction. Each of the proposed `Refairmulate` datasets are derived from MS MARCO [2] and filtered through a novel pipeline that classifies gender bias, generates diverse query variants using LLMs, and selects optimal reformulations based on multi-objective optimization strategies. Notably, the *Optimal* subset achieves perfect bias and effectiveness for the included queries, with an average precision of 1 and average bias scores of 0, offering a theoretical upper bound for fairness-performance trade-offs. Our contributions in this paper can be enumerated as follows: **(1)** Practically, `Refairmulate` enables the

training, and benchmarking of fairness-aware query reformulation models at scale, addressing a previously unmet need in responsible IR. In particular, it provides paired supervised training data that can be used to learn reformulation strategies that optimize for both retrieval effectiveness and fairness, rather than treating fairness as a post-hoc constraint. **(2)** Theoretically, it formalizes multi-objective optimization for bias and retrieval effectiveness as core query reformulation criteria, challenging the often over-reliance on the sole retrieval effectiveness objective. **(3)** Methodologically, it introduces a pipeline that integrates classification, generation, and optimization, applicable beyond gender bias to possibly address other fairness dimensions in the future. **(4)** From a community standpoint, it provides the first public, large-scale benchmark set of datasets purpose-built for studying gender fairness in query reformulation enabling reproducibility, and evaluation consistency.

In our experiments across five retrieval models, `Refairmulate` demonstrates up to 76.0% improvements in effectiveness and 48.5% reduction in gender bias, establishing its utility and robustness. The key point of this open resource is that `Refairmulate` makes it possible to *train* and *evaluate* query reformulation methods under a joint objective, i.e., learning reformulations that improve ranking quality while measurably reducing representational gender bias. To our knowledge, `Refairmulate` is the first resource of its kind, and all of its datasets and code are released publicly to support community adoption and further research in this space: https://github.com/haisonle001/Refairmulate.

## 2 Related Work

**Query Reformulation Datasets.** Constructing reliable ground-truth query pairs is a longstanding challenge in query reformulation research. Session-based approaches extract pairs from user search logs, treating the final query in a session as the best formulation [5]. While scalable, these methods are prone to query drift when users pursue multiple topics within a single session. Tamannaee et al. [23] address this by applying unsupervised revision strategies and selecting the variant with the highest MAP, but the approach is computationally prohibitive for large collections such as MS MARCO and yields only silver-standard data. Another interesting approach to generate a query pair dataset has been to pair those queries that have an overlapping set of relevant judgement documents [28]. Most closely related to our work, Arabzadeh et al. [1] introduced the *Matches Made in Heaven* toolkit and three large-scale query reformulation datasets (Diamond, Platinum, and Gold) built on MS MARCO, where query pairs are constructed to guarantee improvements in retrieval effectiveness. While their datasets have significantly advanced supervised query reformulation, they are solely optimized for retrieval effectiveness and do not account for fairness considerations.

**Gender Bias in Retrieval.** Rekabsaz and Schedl [18] showed that neural ranking models can amplify gender bias present in training corpora and introduced the ARaB (Average Rank Bias) family to measure bias in ranked lists. ARaB provides a position-sensitive score by estimating the gender magnitude of top-ranked documents and aggregating it across the ranking. Positive values indicate male-skewed results, while negative values indicate female-skewed results. Different ARaB variants (e.g., TF-, Boolean-, and term-count–based) differ in how document-level gender magnitude is computed (e.g, ARaB-TF usethe logarithm of the term frequencies of gender-definitional words, capturing how strongly gendered terms appear in a document). Complementary lexical-level analysis can be conducted using LIWC [15].

Rekabsaz et al. [17] further released human-annotated query sets labeled as male, female, or neutral, and Bigdeli et al. [3] proposed a BERT-based classifier to scale this annotation process. While these works provide tools for measuring and detecting gender bias, they do not address mitigation through query reformulation. `Refairmulate` builds on this foundation by constructing large-scale reformulation pairs explicitly optimized for both retrieval effectiveness and gender bias reduction. While these works provide valuable metrics and classifiers for detecting gender bias, they do not offer mechanisms or resources for mitigating bias through query reformulation.

## 3 The Proposed Refairmulate Dataset

### 3.1 Purpose and Utility

The `Refairmulate` dataset supports gender-fair query reformulation through three subsets:

- *Optimal* includes only those query pairs with perfect scores (maximum effectiveness, zero bias),
- *Effective* includes query pairs that are guaranteed to have reformulated variants with better retrieval effectiveness and lower gender bias compared to the original query,
- *Fair* includes query pairs whose reformulated variant has a lower gender bias compared to the original query regardless of its retrieval effectiveness.

Each example contains an original query $q$, a reformulated query $q'$, the retrieved rankings for both, and the corresponding effectiveness (e.g., RR@10/MRR@10) and bias scores (e.g., ARaB variants and LIWC). This structure makes the dataset suitable for training supervised query selection models (e.g., cross-encoders) and for benchmarking reformulation methods under explicit fairness and effectiveness objectives. Unlike existing resources, `Refairmulate` does not assume fairness-neutral generation or post-hoc correction but integrates fairness directly into the query reformulation process. This makes it a uniquely valuable asset for model selection, ablation studies, and fairness-performance trade-off analysis across multiple retrieval models.

### 3.2 Approach Overview

Our goal has been to build a large-scale dataset of query reformulation pairs that offers controlled degrees of fairness and effectiveness in semantically-comparable (equivalent) queries. Each instance consists of an original gender-neutral query and its reformulated variant that, when retrieved, leads to reduced gender bias and improved retrieval effectiveness. To achieve this, we adopt a pipeline-based strategy grounded in multi-objective optimization. The pipeline begins with a curated base of queries and applies successive stages of filtering, generation, and scoring to produce high-quality query reformulation pairs. The core idea is to treat fairness and effectiveness as competing but reconcilable objectives and to explicitly encode both into the query variant generation and selection process.

**Algorithm 1** Refairmulate: Fair and Effective Query Reformulation

**Require:** Query set $Q$, relevant documents $D_q$ for each $q \in Q$
**Ensure:** Reformulated query pairs $QP = \{(q, q') \mid q \in Q, q' \in Q'\}$

1: **Initialize** $QP \leftarrow \emptyset$
2: **for all** $q \in Q$ **do**
3: **if** $C(q) \neq 0$ **then continue**
4: **end if** ▷ Skip biased queries
5: Compute $\text{bias}(q, D_q)$, $\text{eff}(q, D_q)$ and categorize $q$
6: $V_q \leftarrow G(q, D_q)$ ▷ Generate variants
7: **for all** $v_q^{(i)} \in V_q$ **do**
8: Compute $\text{bias}(v_q^{(i)}, D_{v_q^{(i)}})$, $\text{eff}(v_q^{(i)}, D_{v_q^{(i)}})$
9: $S(q, v_q^{(i)}) \leftarrow w_e \Delta\text{eff} + w_b \Delta\text{bias}$
10: **end for**
11: $q' \leftarrow \arg\max_{v_q^{(i)} \in V_q} S(q, v_q^{(i)})$
12: $QP \leftarrow QP \cup \{(q, q')\}$
13: **end for**
14: **return** $QP$

Concretely, we start with the MS MARCO passage ranking dataset[1] and apply a BERT-based query gender classifier to retain only gender-neutral queries. For each such query, we retrieve its top documents and use them as contextual input to a fine-tuned large language model, which generates a diverse set of candidate query reformulations. Each candidate is then evaluated for retrieval effectiveness (e.g., MRR@10) and gender bias (e.g., ARaB and LIWC metrics). A multi-objective scoring function compares each candidate to the original query and selects the best-performing variant under group-specific criteria. This process yields labeled query pairs along with ranking data and metadata, ultimately partitioned into three subsets: *Optimal*, *Effective*, and *Fair*.

## 3.3 Dataset Construction

Given a source query $q$, the objective of `Refairmulate` is to generate a reformulated query $q'$ that preserves the information need of $q$ while reducing bias (e.g., measured by a gender bias score), and improving retrieval effectiveness (e.g., measured by Mean Average Precision (MAP) or Mean Reciprocal Rank (MRR)). Formally, let $Q = \{q_1, q_2, \ldots, q_n\}$ denote the set of source queries, and $D_{q_i} = \{d_{i_1}, d_{i_2}, \ldots, d_{i_m}\}$ represent the set of relevant documents for query $q$. The pipeline produces a set of query pairs $QP = \{(q, q') \mid q \in Q, q' \in Q'\}$, where $Q'$ is the set of reformulated queries. Each $q'$ is designed to maximize retrieval effectiveness, $\text{eff}(q', D_{q'})$, and minimize a bias score, $bias(q', D_{q'})$, ensuring $q'$ aligns with the information need of $q$ while reducing societal biases, particularly gender bias.

**Query Classification.** The first step in the `Refairmulate` construction pipeline is to identify and retain only gender-neutral queries. This filtering step is critical for isolating bias introduced by the retrieval system or reformulation model, rather than the query itself. We define stereotypical gender bias as the tendency of a ranking system to favor one gender in the retrieved documents for a gender-neutral query, even though neutral representation is expected. Consequently, the `Refairmulate` dataset focuses exclusively on gender-neutral queries, where such bias may emerge. To identify gender-neutral queries, we employ a BERT-based classifier following the methodology and trained model of Bigdeli et al. [3]. The classifier is trained on a human-annotated dataset of 3,709 queries [18], where each query is labeled as one of three categories: *Neutral*, *Female*, or *Male*. This dataset has been widely adopted as a benchmark for query gender classification in information retrieval [20, 21]. Our classifier achieves up to 86% accuracy, with per-class F1-scores of 0.81, 0.87, and 0.86 for the Female, Male, and Neutral classes, respectively.

Formally, we define a classification function $C : Q \rightarrow 0, 1, 2$, where each query $q \in Q$ is assigned a label indicating its gender association: 0 for gender-neutral, 1 for male-associated, and 2 for female-associated queries. For example, queries such as *"best hairstyles for women"* or *"men's workout routines"* are labeled as gendered, whereas queries like *"healthy breakfast ideas"* or *"top-paying engineering jobs"* are considered gender-neutral. Only queries classified as neutral ($C(q) = 0$) are retained for further processing and the rest are discarded (Algorithm 1, Lines 3–4). This filtering produces the final source query set $Q_0 \subseteq Q$, which serves as the input to the reformulation and optimization stages. By constraining the dataset to gender-neutral queries at input time, we ensure that any downstream bias captured in document rankings or reformulated variants reflects system behavior rather than query formulation. After filtering, we assess the retrieval effectiveness and bias score of remaining gender-neutral queries against the relevant documents $D_q$. Based on bias and effectiveness, we categorize queries into four groups (Algorithm 1, Line 5):

- **Group 1: High Effectiveness, Low Bias** ($bias(q, D_q) \leq \theta_{\text{bias}}$, $\text{eff}(q, D_q) \geq \theta_{\text{eff}}$): Queries with strong performance and minimal bias, requiring minimal reformulation.
- **Group 2: High Effectiveness, High Bias** ($bias(q, D_q) > \theta_{\text{bias}}$, $\text{eff}(q, D_q) \geq \theta_{\text{eff}}$): Effective but biased queries, needing reformulation to reduce bias while preserving effectiveness.
- **Group 3: Low Effectiveness, Low Bias** ($bias(q, D_q) \leq \theta_{\text{bias}}$, $\text{eff}(q, D_q) < \theta_{\text{eff}}$): Unbiased but underperforming queries, requiring reformulation to improve effectiveness.
- **Group 4: Low Effectiveness, High Bias** ($bias(q, D_q) > \theta_{\text{bias}}$, $\text{eff}(q, D_q) < \theta_{\text{eff}}$): Queries that are both biased and ineffective, requiring comprehensive reformulation.

Here, $\theta_{\text{eff}}$ and $\theta_{\text{bias}}$ are the effectiveness and gender bias thresholds. This categorization enables tailored reformulation strategies for each group, unlike prior approaches that apply uniform transformations.

**Query Generation.** For each gender-neutral query retained from the classification stage, the goal of the Query Generation step is to produce a diverse set of semantically equivalent reformulations that vary in lexical or syntactic form. These reformulations serve as candidates for subsequent bias and effectiveness evaluation. Formally, for a given query $q \in Q_0$ and its associated top-retrieved documents $D_q$, we define a generation function $G : Q_0 \cup D_q \rightarrow Q'_q$ that maps the original query and its context to a set of reformulated queries

[1] We utilized MS MARCO as the seed corpus because it remains the standard large-scale training ground for neural retrievers. While the BEIR [9] benchmark is valuable for zero-shot evaluation, it lacks the massive scale of training query-passage pairs necessary for training the generative components of our pipeline effectively.

**Table 1: Performance and bias metrics of the original and reformulated queries for the Optimal, Effective, and Fair datasets, evaluated at a cut-off rank of 10. Absolute values are reported alongside relative improvement percentages.**

| | MRR | ARaB-TC | ARaB-TF | ARaB-BOOL | LIWC |
|---|---|---|---|---|---|
| **Optimal (112,261)** | | | | | |
| Original Query | 0.161 | 0.063 | 0.033 | 0.035 | 0.132 |
| Reformulated Query | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Relative Impr. | 521.1% | 100.0% | 100.0% | 100.0% | 100.0% |
| **Effective (209,343)** | | | | | |
| Original Query | 0.081 | 0.139 | 0.071 | 0.070 | 0.273 |
| Reformulated Query | 0.425 | 0.029 | 0.015 | 0.016 | 0.081 |
| Relative Impr. | 424.9% | 79.1% | 78.9% | 77.7% | 70.3% |
| **Fair (321,604)** | | | | | |
| Original Query | 0.109 | 0.113 | 0.058 | 0.058 | 0.224 |
| Reformulated Query | 0.626 | 0.019 | 0.009 | 0.012 | 0.053 |
| Relative Impr. | 474.3% | 83.2% | 84.5% | 78.8% | 76.3% |

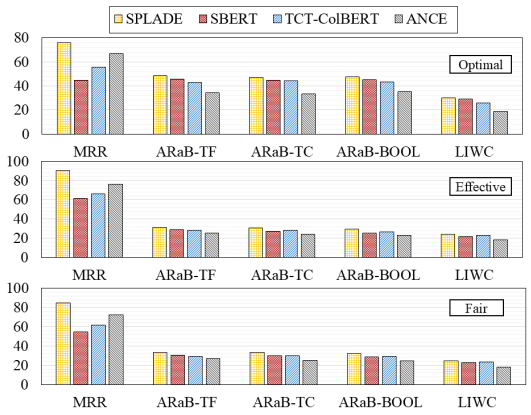


**Figure 1: Mean percentage improvements in effectiveness and bias metrics for the Optimal, Effective, and Fair datasets across four dense retrievers: SPLADE, SBERT, TCT-ColBERT, and ANCE.**

$Q'_q = \{q'_1, q'_2, \ldots, q'_k\}$ (Algorithm 1, Line 6). We implement $G$ using docTTTTTquery[2] [14]. To promote diversity, we use stochastic decoding with beam search, varying generation parameters such as maximum sequence length, top-$k$ sampling, and temperature. We follow the same setup, and training parameters as specified in [14]. For each source query, the model generates reformulations conditioned on individual documents in $D_q$, cycling through the top passages. Duplicate variants are removed, and additional generations are invoked as needed to ensure a complete candidate set of size $K$. This stage produces the pool of reformulation candidates $V_q = \{v_q^{(1)}, \ldots, v_q^{(K)}\}$, which are subsequently evaluated in the selection module.

**Query Selection.** In this stage, we identify the optimal reformulated query for each original query $q$ by jointly evaluating candidate variants in terms of retrieval effectiveness and gender bias. This stage ensures that selected queries offer measurable improvements along one or both dimensions, in alignment with the multi-objective optimization goals of the dataset. To ensure that the generated reformulations do not drift semantically from the original user intent, we first apply a semantic filtering step using a pre-trained Sentence-BERT model [6]. For each candidate $v_q^{(i)} \in V_q$, we compute two metrics: retrieval effectiveness: $\text{eff}(v_q^{(i)}, D_{v_q^{(i)}})$ and gender bias: $\text{bias}(v_q^{(i)}, D_{v_q^{(i)}})$ (Algorithm 1, Line 8). Here, $D_{v_q^{(i)}}$ denotes the top documents retrieved using the variant $v_q^{(i)}$. These values are compared against the original query's scores, $\text{eff}(q, D_q)$ and $\text{bias}(q, D_q)$, to assess improvement or degradation introduced by each candidate. For retrieval effectiveness, we use Reciprocal Rank at cut-off 10 (RR@10), and for gender bias we use the Average Rank Bias with term-frequency gender magnitude (ARaB-tf) introduced by Rekabsaz et. al. in [18]. To support structured dataset construction, we assign each variant a discrete label using a rule-based function:

$$L(v_q^{(i)}) = \begin{cases} 0 & \text{if } \Delta\text{eff} > 0 \text{ and } \Delta\text{bias} < 0, \\ 1 & \text{if } \Delta\text{eff} > 0 \text{ and } \Delta\text{bias} = 0, \\ 2 & \text{if } \Delta\text{eff} = 0 \text{ and } \Delta\text{bias} < 0, \\ 3 & \text{if } \Delta\text{eff} = 0 \text{ and } \Delta\text{bias} = 0, \\ 4 & \text{otherwise,} \end{cases} \tag{1}$$

In this equation, the gain in effectiveness and reduction in bias are defined as:

$$\Delta\text{eff}(q, v_q^{(i)}) = \text{eff}(v_q^{(i)}, D_{v_q^{(i)}}) - \text{eff}(q, D_q), \tag{2}$$

$$\Delta\text{bias}(q, v_q^{(i)}) = \text{bias}(v_q^{(i)}, D_{v_q^{(i)}}) - \text{bias}(q, D_q). \tag{3}$$

We note that $\Delta\text{bias}(q, v_q^{(i)}) < 0$ indicates that the reformulated query has lower bias than the original query, which is the desired outcome. Similarly, $\Delta\text{eff}(q, v_q^{(i)}) > 0$ indicates an improvement in the effectiveness of the reformulated query. On this basis, Label 0 indicates improvement in both dimensions, Labels 1 and 2 indicate improvement in one dimension only, Label 3 indicates no change, and Label 4 captures degradation in both effectiveness and bias simultaneously, i.e., the reformulated query performs worse than the original on both objectives. Label 4 cases are excluded from the three positive subsets but are retained as negative samples for contrastive learning.

We use these labels to decide which reformulations are acceptable for each query group. Group 4 queries need improvements in both effectiveness and bias, so we keep only Label 0 reformulations (better on both). Group 3 queries mainly need better effectiveness, so we keep Label 0 (best case) and Label 1 (effectiveness improves and bias does not get worse). Group 2 queries mainly need lower bias, so we keep Label 0 (best case) and Label 2 (bias improves and effectiveness does not get worse). Group 1 queries require little or no change, so we allow Labels 0–3 (anything that improves at least one objective or leaves both unchanged). After this filtering, we pick the single best reformulation for each query using a scoring function that combines effectiveness gain and bias reduction (Algorithm 1, Line 9):

[2]We selected docTTTTTquery due to its computational efficiency (220M parameters) and its specific fine-tuning on MS MARCO, which aligns with our base dataset. However, our proposed pipeline is model-agnostic and can be instantiated with larger generative models if computational resources allow.

**Table 2: Example query pairs from each subset showing original and reformulated queries with their effectiveness and bias scores.**

| Subset | Original Query | Reformulated Query | ΔMRR (%) | ΔARaB (%) |
|---|---|---|---|---|
| Optimal | Mig 29 fighter vs f16 | what were the f-16 and mig-29 planes intended to do | 100 | 100.0 |
| Effective | what size nappies does a 6 month old wear | what size nappies should kids be in at 6 months | 0 | 77.9 |
| Fair | how long does it take for marigolds to grow from seed | how fast will marigold seeds germinate indoors | 50 | 100.0 |

$$S(q, v_q^{(i)}) = w_e \cdot \Delta\text{eff}(q, v_q^{(i)}) + w_b \cdot \Delta\text{bias}(q, v_q^{(i)}), \quad (4)$$

The weights $w_e$ and $w_b$ control the relative importance of effectiveness and fairness, and can be tuned depending on the optimization priority. The final selection for each query $q$ is the variant $v_q^*$ that maximizes $S(q, v_q^{(i)})$, subject to group-specific constraints.

## 4 Dataset Characteristics

Building on the multi-stage construction pipeline described above, we curate `Refairmulate`, a collection of large-scale resources specifically designed for gender-fair aware query reformulation. The `Refairmulate` dataset is generated using our end-to-end pipeline over the MS MARCO Passage Ranking dataset [13], which contains 8.8 million passages and over 500,000 query-document relevance annotations. Starting from these MS MARCO training queries identified as gender-neutral by our gender classifier, we construct high-quality reformulation pairs that encode varying trade-offs between retrieval effectiveness and fairness. In the proposed dataset, we assigned equal importance (weight) to effectiveness, and fairness. Therefore, we selected $w_e = 1$, and $w_b = 1$ in Equation 4. However, one can vary $w_e$, and $w_b$ for a desired balance based on the application. We evaluate the effectiveness of each generated alternative query based on BM25 using the Anserini implementation [25] and organize the dataset into three distinct subsets:

***Optimal*:** Contains 112, 261 query pairs that achieve dual ideal objectives, i.e., maximum retrieval performance (RR@10 = 1) and complete removal of measured bias (all bias metrics = 0).

***Effective*:** Includes 209, 343 query pairs whose reformulated variants are guaranteed to have at least comparable retrieval effectiveness to the original query but always better gender bias, i.e. the query pairs whose reformulated queries are selected from Label 0, and 2 in Equation 1).

***Fair*:** A broader collection that combines *Optimal* and *Effective* subsets with additional pairs exhibiting partial or mixed retrieval effectiveness, offering coverage for more flexible optimization scenarios. This set includes 321, 604 query pairs. These query pairs include reformulated queries matching Labels 0, 1, 2, or 3 in Equation 1.

In addition to positive reformulations, we include *negative samples*, which are query pairs where the reformulated variant performs worse than the original in both effectiveness and bias (query pairs whose reformulated queries are selected from Label 4 in Equation 1). These cases are valuable for contrastive learning settings, enabling the training of models to distinguish between desirable and undesirable reformulations.

**Table 3: Retrieval effectiveness and gender bias on 76,289 overlapping queries between `Refairmulate` and *Matches Made in Heaven* [1].**

| | RR@10 | ARaB-TC | ARaB-TF | ARaB-BOOL |
|---|---|---|---|---|
| Original Query | 0.075 | 0.355 | 0.180 | 0.180 |
| Matches Made in Heaven | 0.669 | 0.192 | 0.097 | 0.097 |
| `Refairmulate` (Ours) | 0.643 | **0.059** | **0.031** | **0.032** |

**Note.** We ensured that the `Refairmulate` dataset was constructed and released following reproducibility guidelines. All source data originate from the publicly available MS MARCO Passage Ranking corpus, which is distributed under the Microsoft Research License. `Refairmulate` includes only automatically generated and anonymized query–reformulation pairs without any personally identifiable information (PII) or user-specific attributes. No human subject data were collected or inferred during the dataset curation process.

Table 1 summarizes average effectiveness and bias for original vs. reformulated queries in each subset under BM25. As expected, the *Optimal* subset demonstrates the most dramatic improvements, including an 521.118% increase in MRR@10 and a complete reduction in bias. The *Effective* subset yields substantial but targeted gains (e.g., 424.941% in MRR@10 and up to 80% in bias improvements), while the *Fair* subset balances both objectives (474.312% MRR@10 gain, 84% bias improvement), making it suitable for end-to-end training or evaluation of fairness-aware models. An MRR@10 of one indicates that the reformulated queries were selected so that the dataset achieves a perfect MRR, rather than representing the results of an evaluation. To assess generalizability, we evaluate each subset across four dense retrieval architectures, namely SPLADE [7], SBERT [16], TCT-ColBERT [10], and ANCE [24], on a held-out sample of 10,000 queries per dataset. As shown in Figure 1, the reformulated queries yield consistent gains in both retrieval performance and fairness metrics, indicating that selecting reformulations using BM25 still yields consistent gains across architectures. In Table 2, we present sample queries from the three sets, along with the corresponding improvements in performance and bias achieved by the reformulated queries compared to the original ones. For example for the query "*Mig29 fighter vs f16*, the reformulated query is "*What were the f-16, and mig-29 planes intended to do*. We can observe that while both queries are gender neutral, for the reformulated query, there is a 100% improvement in ranking performance, and 100% reduction in gender bias. Table 3 further contextualizes `Refairmulate` against *Matches Made in Heaven* [1] on overlapping queries, showing substantially lower measured bias with comparable RR@10.

## 5 Benchmarking the Refairmulate Dataset

To demonstrate the practical utility and effectiveness of our dataset for bias-aware query reformulation models, we benchmarked our models using an experimental setup adapted from [8]. Our evaluation specifically targets the joint optimization of retrieval effectiveness and fairness metrics, showcasing how the `Refairmulate`

**Table 4: Mean percentage improvements in performance and bias metrics evaluated on the 1,765 and 215 query sets at a cut-off rank of 10.**

| Queries | MRR | ARaB-TC | ARaB-TF | ARaB-BOOL | LIWC |
|---|---|---|---|---|---|
| 215 Queries | 154.64% | 81.53% | 78.61% | 75.60% | 72.08% |
| 1,765 Queries | 88.71% | 63.32% | 59.95% | 55.57% | 56.34% |

datasets can be leveraged to simultaneously improve search performance while reducing gender bias. We train a BERT-based cross-encoder that learns to predict whether a reformulated query $q'$ represents an improvement over the original query $q$ with retrieval effectiveness (measured by MRR@10) knowing the bias score (quantified through ARaB metrics).

The model is trained using contrastive learning with positive examples from the `Refairmulate` query pairs where the reformulated query improves effectiveness and/or fairness, and negative examples where performance worsens. Each input pair $(q, q')$ is encoded jointly using the BERT-based cross-encoder architecture, where the `[CLS]` representation captures the interaction between the original and reformulated queries. The model outputs a scalar value $\hat{y} \in [0, 1]$ indicating the likelihood that $q'$ represents a beneficial reformulation of $q$. The network is trained using the binary cross-entropy loss with a learning rate of $2 \times 10^{-5}$, a warm-up ratio of 10%, a batch size of 16, and one training epoch, consistent with prior query selection frameworks [8]. We evaluate the performance of the model on two human-annotated neutral query sets (215 and 1,765 queries) offered for this purpose by Rekabsaz et al. [17]. For each of the queries, we select the best reformulated version based on the BERT scores for the alternative queries. The best alternative is selected, and then we retrieve for that query. The experimental results in Table 4 show that the proposed approach significantly improves both retrieval effectiveness and bias reduction. On the 215-query dataset, it yields a 154.64% mean improvement in MRR10, while with larger 1,765-query set, MRR10 improves by 88.71%. Substantial reductions in gender bias are also observed where in the 215-query set, ARaB-TC, ARaB-TF, ARaB-BOOL, and LIWC improve by 81.53%, 78.61%, 75.60%, and 72.08%, respectively; for the 1,765-query set, improvements are 63.32%, 59.95%, 55.57%, and 56.34%, respectively. These findings confirm the approach's effectiveness in enhancing fairness and mitigating gender bias across both small and large datasets.

Figure 2 presents the Help–Hurt diagram, illustrating the relative changes in performance (measured by MRR) and bias (measured by ARaB) across the two query sets. The diagram provides an intuitive view of how individual query pairs respond to the applied optimization. Specifically, for the 215-query set, approximately 32.22% of the queries demonstrate performance improvements, while only 6.67% show degradation. Similarly, within the 1765-query set, 33.27% of the queries improve and only 14.61% experience a decline. Furthermore, substantial reductions in gender bias are also observed across all ARaB dimensions and LIWC-based metrics. This consistent upward trend across both sets points to the effectiveness of the proposed dataset in enhancing the balance between performance and fairness, leading to more reliable and equitable query reformulations.

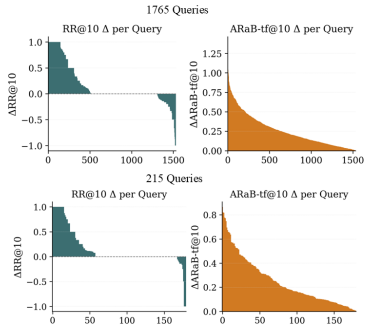


**Figure 2: Help–Hurt diagram illustrating the comparative changes in performance (measured by MRR) and bias (measured by ARaB-TF) across the 1765 and 215 query sets.**

## 6 Limitations

We acknowledge that `Refairmulate` comes with at least the following inherent limitations from its own design, including: (1) The evaluation framework used in our paper depends on a limited set of quantitative fairness indicators such as ARaB and LIWC. While these metrics offer measurement consistency, they provide only a partial view of equity and may fail to capture subtle or intersectional disparities present in natural queries [11]. (2) The effectiveness of the `Refairmulate` dataset can vary across linguistic and cultural contexts, which we have not captured or qualified in our work. Gender bias might differ by region or language, and applying models trained on English web data, predominantly from the Global North, may lead to incomplete representation of content from other regions and languages, and hence bias mitigation on this dataset might not be generally transferrable. (3) Efforts to mitigate one bias dimension, for instance gender, can modify the distribution of others such as occupation or geography. This interdependence may negatively impact other bias dimensions in practice and hence points to the importance of evaluating fairness holistically rather than along isolated attributes. (4) Since `Refairmulate` is constructed from MS MARCO, it inevitably inherits demographic and topical skews present in the original corpus. Such inherited patterns constrain the diversity of linguistic and contextual representations, may reduce the dataset's ability to generalize beyond those observable in the MS MARCO collection.

## 7 Concluding Remarks

We introduce `Refairmulate`, the first large-scale training and benchmarking dataset explicitly constructed to support gender-fair query reformulation. By operationalizing a multi-objective optimization framework over retrieval effectiveness and gender bias, and demonstrating consistent gains across multiple dense retrieval architectures, our proposed dataset offers a reliable foundation for the training and evaluation of fairness-sensitive IR systems. Our results validate the feasibility of improving retrieval effectiveness while simultaneously reducing representational harm, challenging the assumption that fairness must come at the cost of performance. Beyond benchmarking utility, `Refairmulate` advances a methodological shift where it treats fairness as a first-class optimization objective rather than a post-hoc correction.

# References


[1] Negar Arabzadeh, Amin Bigdeli, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2021. Matches made in heaven: Toolkit and large-scale datasets for supervised query reformulation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4417–4425.

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[3] Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2021. Exploring gender biases in information retrieval relevance judgement datasets. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 216–224.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[5] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to Attend, Copy, and Generate for Session-Based Query Suggestion. arXiv:1708.03418 [cs.IR] https://arxiv.org/abs/1708.03418

[6] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2288–2292.

[8] Seyed Mohammad Hosseini, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2024. Enhanced Retrieval Effectiveness through Selective Query Generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3792–3796.

[9] Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. 2024. Resources for Brewing BEIR: Reproducible Reference Models and Statistical Analyses. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1431–1440. doi:10.1145/3626772.3657862

[10] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 39–48.

[11] Anja Klasnja, Negar Arabzadeh, Mahbod Mehrvarz, and Ebrahim Bagheri. 2022. On the characteristics of ranking-based gender bias measures. In *Proceedings of the 14th ACM Web Science Conference 2022*. 245–249.

[12] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 1929–1932.

[13] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268 http://arxiv.org/abs/1611.09268

[14] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375* (2019).

[15] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

[16] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[17] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 306–316.

[18] Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias?. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2065–2068.

[19] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. 2016. Using word embeddings for automatic query expansion. *arXiv preprint arXiv:1606.07608* (2016).

[20] Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. 2022. Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases.. In *EDBT*. 2–435.

[21] Shirin Seyedsalehi, Sara Salamat, Negar Arabzadeh, Sajad Ebrahimi, Morteza Zihayat, and Ebrahim Bagheri. 2025. Gender disentangled representation learning in neural rankers. *Machine Learning* 114, 5 (2025), 1–33.

[22] Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. 2020. Reque: a configurable workflow and dataset collection for query refinement. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3165–3172.

[23] Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. 2020. ReQue: A Configurable Workflow and Dataset Collection for Query Refinement. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3165–3172. doi:10.1145/3340531.3412775

[24] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[25] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 1253–1256.

[26] Hamed Zamani and W Bruce Croft. 2017. Relevance-based word embedding. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. 505–514.

[27] George Zerveas, Ruochen Zhang, Leila Kim, and Carsten Eickhoff. 2020. Brown university at trec deep learning 2019. *arXiv preprint arXiv:2009.04016* (2020).

[28] George Zerveas, Ruochen Zhang, Leila Kim, and Carsten Eickhoff. 2020. Brown University at TREC Deep Learning 2019. arXiv:2009.04016 [cs.IR] https://arxiv.org/abs/2009.04016