Mitigating Bias in Item Retrieval for Enhancing Exam Assembly in Vocational Education Services

Alonso Palomino[1,3] **Andreas Fischer**[2] **David Buschhüter**[1] **Roland Roller**[1]
Niels Pinkwart[1] **Benjamin Paaßen**[1,3]

[1] German Research Center for Artificial Intelligence (DFKI), Germany, `<first>.<last>@dfki.de`
[2] Forschungsinstitut Betriebliche Bildung (f-bb), Germany, `<first>.<last>@f-bb.de`
[3] Bielefeld University, Germany, `<first>.<last>@techfak.uni-bielefeld.de`

Abstract

In education, high-quality exams must cover broad specifications across diverse difficulty levels during the assembly and calibration of test items to effectively measure examinees' competence. However, balancing the trade-off of selecting relevant test items while fulfilling exam specifications without bias is challenging, particularly when manual item selection and exam assembly rely on a pre-validated item base. To address this limitation, we propose a new mixed-integer programming re-ranking approach to improve relevance, while mitigating bias on an industry-grade exam assembly platform. We evaluate our approach by comparing it against nine bias mitigation re-ranking methods in 225 experiments on a real-world benchmark data set from vocational education services. Experimental results demonstrate a 17% relevance improvement with a 9% bias reduction when integrating sequential optimization techniques with improved contextual relevance augmentation and scoring using a large language model. Our approach bridges information retrieval and exam assembly, enhancing the human-in-the-loop exam assembly process while promoting unbiased exam design

1 Introduction

Retrieving and assembling test items into exams from a pre-validated item base that accurately and comprehensively estimates examinees' competence remains a significant challenge in education (Linden et al., 2005; Lane et al., 2016; Kurdi et al., 2020). Despite the practical importance of exam assembly, few methods exist to support educators during manual item retrieval for exam assembly tasks (Palomino et al., 2024; Bißantz et al., 2024).

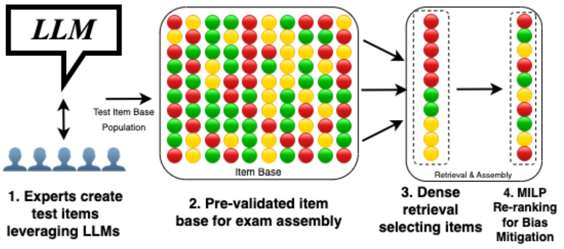
LLM
Test Item Base
Population
Item Base
Retrieval & Assembly
1. Experts create test items leveraging LLMs
2. Pre-validated item base for exam assembly
3. Dense retrieval selecting items
4. MILP Re-ranking for Bias Mitigation

Figure 1: Test item retrieval workflow for exam assembly. VET experts use LLMs to generate and refine test items, adjusting difficulty based on expertise. Items are color-coded by difficulty: green (easy), yellow (medium), and red (hard). Experts populate a test item base, then retrieve and assemble items into formative exams. The initial search skews toward harder items, but our MILP-driven bias mitigation re-ranks difficulty distribution for a balanced ranking.

A key limitation in high-quality exam assembly, especially when relying on a pre-validated test item base, is attribute bias, which typically arises when the retrieved items' ranking order reflects imbalances in specific attributes, such as difficulty or source, while prioritizing the relevance to a topic. For instance, during the retrieval phase, while items of a given difficulty level may be overrepresented (or underrepresented) in a ranking, manually or synthetically generated items via large language models (LLMs) could be omitted (or overly included). As a result, assembled exams may differ significantly in psychometric selection, raising concerns about the exams' quality and comprehensiveness.

Information retrieval (IR) research has extensively documented how information access systems may retrieve specific content while systematically and inadvertently omitting relevant but underrepresented content (Baeza-Yates, 2018; Gao and Shah, 2021). This phenomenon, also known as an instance of algorithmic bias, typically leads to "skewed or unfair" system behavior, potentially compromising system accuracy and integrity while perpetuating and reinforcing incomplete or distorted results (Singh and Joachims, 2018; Biega et al., 2019; Mehrabi et al., 2021; Shahbazi et al., 2023). While several bias mitigation methods exist in IR, and although linear optimization methods could be employed to assemble exams (Linden

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 183–193
April 30, 2025 ©2025 Association for Computational Linguistics

et al., 2005; Bißantz et al., 2024), their application in supporting unbiased item retrieval for manual exam assembly still needs to be explored. This paper bridges this gap by introducing a new IR bias mitigation technique based on relevance and bias metric-based balancing.

We addressed difficulty and source bias in item retrieval to enhance the human-in-the-loop exam assembly process, a critical requirement for testing and educational organizations (Lane et al., 2016; Bißantz et al., 2024). Specifically, we examined bfz's[1] internal item retrieval platform for exam assembly, EdTec-QBuilder. As Germany's largest vocational education and training (VET) provider, bfz employs this system for test item selection, which we evaluated using an industry-standard TREC-style benchmark comprising 5,624 validated items (Palomino et al., 2024). On this benchmark, we employed an ad-hoc retrieval methodology to evaluate nine debiasing techniques to mitigate item difficulty and source bias, conducting 225 experiments overall[2]. We propose a new bias mitigation method incorporating a novel mixed-integer linear programming (MILP) approach with enhanced relevance generation via LLM-based contextual augmentation, finding that our approach best optimizes the trade-off between bias mitigation and the relevance of the retrieved test items (see Table 2). Figure 1 illustrates our approach for test item retrieval and difficulty calibration in exam assembly. After VET experts query a test item base, our method reorders the retrieved items to mitigate difficulty bias while enhancing topical relevance. This approach ensures that manual exam designers receive balanced test item rankings that reflect a broad range of difficulty levels and topics, ultimately facilitating the creation of well-balanced exams by surfacing relevant items that might otherwise be omitted. We elaborate on the industry application of our new bias mitigation re-ranking method. Finally, we present conclusions and future lines of research.

2 Related Work

Bias and fairness in information retrieval (IR) pertain to how systems rank objects, potentially favoring or disadvantaging specific groups or categories unintentionally. Numerous approaches have emerged to measure and address bias and unfairness in IR. For instance, Kırnap et al. (2021) proposed a probabilistic weighted sampling and Horvitz-Thompson inference approach to measuring bias based on proportional item exposure. Raj and Ekstrand (2020, 2022) evaluated and compared existing bias and fairness metrics, finding conceptual similarities but differences in the effect of ranking attributes, such as group/category distribution. Recently, Bernard and Balog (2023) and Dai et al. (2024) surveyed 75 and 100 papers on bias and fairness in IR, respectively, finding that current notions of bias in IR are complex and multidimensional; most current approaches to tackle bias intervene at the in- or post-processing level. Regarding in-processing interventions to address bias in IR, Celis et al. (2018) introduced a theoretical framework based on bipartite matching constraints, packing integer programming, and greedy-based diversification methods to incorporate fairness constraints during ranking generation. Thonet and Renders (2020) developed an efficient sequential greedy brute-force ranker that combines greedy selection to produce fair rankings when target groups are unknown. Morik et al. (2020) proposed a dynamic learning-to-rank approach that mitigated exposure bias by amortizing group allocation fairness while estimating relevance scores. Li et al. (2022) mitigated bias in neural retrieval systems with an in-batch balancing regularization method enforcing fairness constraints during neural retrieval model training. Wang et al. (2023) proposed a hyperbolic mitigation model for news recommendations, which employs a re-weighting aggregation module to reduce conformity bias while improving user intrinsic interests. Hager et al. (2024) proposed a regression expectation maximization model for learning-to-rank to address position bias with click data. As for post-processing interventions to mitigate bias in IR Zhu et al. (2020) debiased a Bayesian personalized ranking method with an adversarial learning model that enhances predicted preferences among groups while ensuring statistical parity. Burke et al. (2021) introduced a candidate ranking multi-model aggregation method to enhance the protected group representation, enforcing fairness over hiring decisions. Feng and Shah (2022) introduced an ϵ-greedy post-reranking method to tackle gender bias by reducing imbalanced representations over gender groups while maintaining the original ranking's relevance. In contrast to this prior work, we consider a new

[1] https://www.bfz.de

[2] Code and search runs available at: https://dfki-kiperweb.de

mixed-integer linear programming re-ranking formulation, which maximizes the retained relevance while minimizing the difference between actual and desired group distribution. Furthermore, we are the first to consider bias in the context of item retrieval for manual exam assembly tasks, a sought-after capability by educational and assessment organizations. Additionally, we explore LLM prompting and optimization strategies for contextual query generation and improved relevance generation (Sun et al., 2023) to boost our method's performance.

3 Bias Framework

Below, we outline the bias measurement framework, testbed, and analysis of ANN+CE, the core search and retrieval method used by EdTec-QBuilder for manual item retrieval and exam assembly.

Bias Measuring Bias and unfairness in IR can be modeled from the user's perspective. As users' visual attention distribution is higher for top-ranked items, bias and unfairness increase if higher-ranked items from specific classes are over-represented among the top-ranked entries. For our use case, we operationalized bias measurement and mitigation using the framework proposed by Sapiezynski et al. (2019). We measured bias as how balanced an item's difficulty and source classes are represented across the top search results (i.e., difficulty and source bias). Due to its stability and robustness, we employed attention-weighted rank fairness (AWRF) (Ekstrand et al., 2022; Raj and Ekstrand, 2022; Cachel and Rundensteiner, 2024) as a metric to evaluate the difficulty and source bias. Additionally, to measure the tradeoff between relevance and bias equally, we calculated the following joint metric (JM):

$$\text{JM} = \text{nDCG}(L_{r,c}) \cdot (1 - \text{AWRF}(L_{r,c})) \qquad (1)$$

where $L_{r,c}$ represents a ranked list of relevant items with their corresponding group information (difficulty and source), and where nDCG represents the normalized discounted gain (higher is better). We inverted the AWRF scale to make higher values better and multiplied both scales to create a joint metric.

Testbed To measure and operationalize bias mitigation methods that improve our industry partner's item retrieval and assembly platform's performance, we employed our previous TREC-style testbed for the manual item retrieval and exam assembly task (Palomino et al., 2024). The testbed includes 25 different top-performing frozen search runs, each comprising top-100 rankings for 15 queries across 5,624 items focused on VET for the German job market. Each test item is accompanied by its corresponding 3-level graded query relevance judgments, attribute labels for difficulty (e.g., easy, medium, or hard), and source (i.e., manually created by a VET expert or generated via ChatGPT3.5).

Bias Analysis We analyzed bias in our testbed's top 50 search results, focusing on the most interacted ranking positions. Table 1 summarizes the best-performing nearest neighbors with cross-encoder (ANN+CE) searches at a cutoff of 50, a legacy item retrieval method previously transferred to our industry partner; this method was selected as the core item retrieval method due to its strong performance in our previous benchmark, as described in (Palomino et al., 2024). Each listed ANN+CE method combines its corresponding core embedding model. We included standard IR metrics, with AWRF and JM scores, to measure difficulty and source bias. From a relevance standpoint, while ANN+CE methods #1 and #2 reported the highest nDCG values of 0.28 and 0.25, respectively, method #3 reported the lowest nDCG of 0.24. From a bias handling standpoint, while method #3 reported the lowest bias with an average AWRF of 0.47, method #1 reported the highest average AWRF with a score of 0.52. While method #3 decreased the difficulty bias effect with an AWRF score of 0.33, method #1 underperformed when handling the item's difficulty classes, showing an AWRF score of 0.47. However, regarding the source bias, method #1 performed best with a score of 0.57, while method #3 performed worst with 0.62, indicating the highest source bias score. Ultimately, when considering relevance and bias equally, method #1 performed the best with a JM of 0.15, while methods #2 and #3 reported 0.12, suggesting more loss of relevance performance. This performance highlights the importance of addressing the multidimensional aspects in balancing the relevance/bias tradeoff in retrieval methods.

4 MILP-Driven Bias Mitigation

Our task is to mitigate the difficulty and source bias in EdTec-QBuilder (bfz's item retrieval and exam assembly platform). Given a pre-ranked list retrieved items for a given query, we wish to re-

Base Models Performance Metrics @50														
#	Method	Core Embedding Model	nDCG	MRR	Prec.	Rec.	F1	MAP	AWRF (Dif.)	AWRF (Src.)	AWRF (Avg.)	JM (Src.)	JM (Dif.)	JM (Avg.)
1	ANN+CE	gbert-large-paraphrase-euclidean	0.28	0.56	0.08	0.23	0.19	0.20	0.47	0.57	0.52	0.12	0.15	0.13
2		gbert-large-paraphrase-cosine	0.25	0.44	0.07	0.22	0.18	0.19	0.45	0.54	0.50	0.11	0.13	0.12
3		e5-multi-sml-torch	0.24	0.46	0.06	0.21	0.17	0.18	0.33	0.62	0.47	0.09	0.16	0.12

Table 1: Retrieval and bias metrics, for the top-3 baseline ANN+CE search models from (Palomino et al., 2024), ranked in descending order of their average joint metric.

rank the items such that bias regarding difficulty and source among the top search results is reduced. We formalize this task as follows. Let $r_1, \ldots, r_N \mathbb{R}$ be real-valued relevance scores for the top N retrieved items, as provided by some ranking scheme; let $y_{i,k} \in \{0, 1\}$ indicate whether item i belongs to class k; and let $p_k \in [0, 1]$ indicate the desired fraction of class k among the top m ranked items. Then, we wish to re-rank a subset of $m \leq N$ items to the top, such that high relevance is maintained, but bias is reduced. For our specific scenario, $N = 100$, and $m = 50$. The classes are the Cartesian product of the difficulty level (easy, medium, hard) and the source (human-written, GPT-3.5 written) of the items, and the observed class counts divided by the total number of items gives the target distribution p.

To simplify optimization, we do not target Eq. (1) directly but a linear surrogate objective, namely a linear combination of the sum of relevance scores in the subset and the total variation distance between the target distribution p and the actual class distribution among the included items. As such, our fairness term can also be regarded as a measure of demographic parity in the top-m results. Based on our linear surrogate objective, our fair re-ranking scheme can be formulated as a mixed-integer linear program (MILP):

$$\begin{aligned} \min_{\vec{x} \in \{0,1\}^n, \vec{d} \in \mathbb{R}^K} \quad & -\sum_{i=1}^{n} x_i \cdot r_i + \lambda \cdot \sum_{k=1}^{K} d_k \qquad (2) \\ \text{such that} \quad & \vec{1}^T \cdot \vec{x} \leq m \\ & \frac{1}{m} \boldsymbol{Y}^T \cdot \vec{x} - \vec{p} \leq \vec{d} \\ & \vec{p} - \frac{1}{m} \boldsymbol{Y}^T \cdot \vec{x} \leq \vec{d}, \end{aligned}$$

Where x_i is 1 if and only if item i is selected for the top m search results, d_k is a slack variable representing the total variation distance for class k, and λ controls the trade-off between relevance and bias. Equation 2 ensures that the final ranking maintains a class distribution close to the target distribution p, balancing difficulty levels and sources. The fairness constraint minimizes the total variation distance between the observed and desired distributions across all classes.

5 Experiments

Section 3 bias analysis shows that EdTec-QBuilder ANN+CE search model only partially addresses difficulty and source bias. We applied a re-ranking approach on our testbed to evaluate bias mitigation, optimizing the relevance/bias tradeoff within the top 50 ranked results. For a given query, our re-ranking framework ensures fair representation of all relevant difficulty levels and sources at the top of the ranking. We assessed the proposed methods using IR metrics, AWRF via the `ranx` and `FairRankTune` libraries (Bassani, 2022; Cachel and Rundensteiner, 2024), and the JM metric to evaluate the relevance/bias tradeoff.

Mitigation methods Below, we summarize the nine re-ranking methods benchmarked to mitigate difficulty and source bias on EdTec-QBuilder for our task.

1. **Random:** A randomized re-ranking method that sets a proportionate target class distribution constraint inferred from the initial ranking's class distribution.
2. **DetConstSort:** A deterministic constrained sorting method that re-balances the initial ranking input by enforcing a balanced class distribution constraints, ensuring equal group representation (Geyik et al., 2019; Cachel and Rundensteiner, 2024).
3. **MMR:** A maximal marginal relevance ranking diversification method that ensures that highly relevant and distinct items vary from the original ranking (Carbonell and Goldstein, 1998). By selecting items that maximize the weighted combination of relevancy to the query and dissimilarity with the chosen initial items, MRR

minimizes redundancy across items by penalizing items that are highly similar to the original selected. We employed GPT-4o embeddings (Hurst et al., 2024) to calculate the relevancy and similarity terms.

4. **ϵ-greedy :** A re-ranking method to re-balance a given ranking by associating an ϵ probability of swapping positions with a random element below it (Berry and Fristedt, 1985; Feng and Shah, 2022; Cachel and Rundensteiner, 2024). The method greedily explores new random swaps while discovering potentially better rankings and maintaining the original ranking as much as possible.
5. **CMAB:** A LinUCB contextual multi-armed bandit (Strong et al., 2021) for re-ranking. For each item in the ranking (i.e., arm), we attached information such as item class distribution, item's length, query length, and group statistics counts such as standard deviation, entropy, skewness, and gini coefficients (i.e., context). The CMAB method iteratively ranks and selects items, balancing relevance and fairness scores via nDCG and AWRF rewards.
6. **FA*IR:** A greedy statistical method that uses priority queues to re-rank by processing candidate items sequentially selecting them based on fairness constraints inferred using random Bernoulli trials selection, the algorithm operates by internally creating a tabular structure representing a minimum of protected classes candidates needed at each position to pass a statistical fairness test (Zehlike et al., 2017).
7. **MILP:** Our new bias mitigation re-ranking method (see Section 4) is implemented via `SciPy` library. To handle the relevance/bias tradeoff equally, we set the λ parameter to 0.5.
8. **MILP-LLM:** An extension of our MILP method that incorporates the approach of Sun et al. (2023) to improve query expansion and relevance scoring. Using LLM prompting, each query is expanded with related skill topics, enhancing its coverage of relevant test items. Candidate items are then updated with improved relevance scores, computed based on the expanded query and candidate item similarities using GPT-4o embeddings. Finally, MILP optimally re-ranks the items.
9. **MILP-BOpt:** A refinement of MILP-LLM that leverages Head et al. (2021) bayesian optimization to further optimize the bias/relevancy tradeoff of selecting the λ parameter based on optimizing JM scores.

We leveraged our testbed to benchmark the above re-ranking methods for bias mitigation. This evaluation enabled us to effectively address biases present in the current ANN+CE-based search and retrieval method of EdTec-QBuilder (see Table 1).

5.1 Results

Overall, we conducted 225 experiments over our previous item retrieval and assembly benchmark (Palomino et al., 2024). Table 2 summarizes the top three best-performing re-rankers per method with their corresponding core embedding model at a cutoff of 50. From a relevance standpoint, MILP-based methods, such as MILP-LLM (#13) and MILP-BOpt (#16), showed the best performance in comparison with other evaluated methods, with nDCG scores of 0.45 and MRR scores of 0.67. As for the lowest relevance performance, DetConstSort (#5) and MMR (#24) models demonstrated the lowest scores, displaying nDCG values between 0.21 and 0.22 and MRR values ranging from 0.40 to 0.52. From the difficulty bias mitigation standpoint, FA*IR (#12) and MILP-BOpt (#18) performed best, showing the lowest AWRF scores with 0.26 and 0.27 respectively. Methods like CMAB (#1), DetConstSort (#4), and MRR (#22) showed the highest AWRF values, ranging from 0.47 to 0.49, indicating low performance when mitigating difficulty bias. Among the methods showing lower source bias, MILP-BOpt (#18) and MILP-LLM (#21) performed best, displaying both 0.35 AWRF scores. ϵ-greedy (#8) and CMAB (#2) struggled when mitigating the source bias; these methods reported the highest AWRF values, 0.63 and 0.62, respectively. When considering equally the relevance and source bias via the proposed joint metric, MILP-based models performed best; more specifically, MILP-BOpt (#16) and MILP-LLM (#19), both with 0.25. The lowest-performing methods handling equally relevance and bias were based on MMR (#23 and #24) with an average JM score of 0.12. When considering all performance aspects, MILP-BOpt (#16) and MILP-LLM (#19) methods best controlled the relevance/bias tradeoff, both high in nDCG scores of 0.45 and 0.43 while maintaining low average AWRF of 0.43 and 0.41.

Overall, MILP-based methods significantly improve relevance while decreasing difficulty and source bias when compared to our previous search and retrieval approach (see Table 1). In general, when comparing with top previous results, we observed that MILP-BOpt and MILP-LLM outperformed ANN+CE methods in terms of relevance (e.g., method #1 from Table1) by 17%. Regarding mitigating both source and difficulty bias, MILP-BOpt (#16) and MILP-LLM (#13) effectively mitigated bias, showing a decrease of 9% in average AWRF, with respect to method #1 and #2 from our previous results. Finally, judging solely from an nDCG Vs. average AWRF tradeoff perspective, MILP and FA*IR models achieved the best balance by effectively minimizing bias while improving relevance, as demonstrated in their positions on the Pareto frontier (see Appendix A.2), when considering all tested method's nDCG and average AWRF scores, in general MILP methods display improved performance without sacrificing either nDCG or average AWRF (nDCG=0.45 and Avg. AWRF=0.42).

6 Industry Application

We collaborated with bfz, Germany's largest VET provider, to enhance EdTec-QBuilder[3], their internal exam assembly platform. Performance and bias auditing (see Section 4) showed that while Palomino et al. (2024) method effectively retrieved relevant test items for assembling exams, it showed attribute biases related to the difficulty and source of the items, resulting in imbalanced exams, potentially compromising exams' comprehensiveness during the manual assembly process. To address this issue, we intend to deploy the MILP-BOpt re-ranking method (#16), which achieved a 9% reduction in AWRF and a 17% improvement in nDCG compared by solely relying on the previous approach (see Table 1). To prepare for the future integration of our MILP-BOpt re-ranking method into the EdTec-QBuilder platform, we completed a pre-deployment testing phase (see Appendix A.3), which aims to maintain system scalability and reliability by leveraging the legacy retrieval capabilities but optimizing it via MILP-BOpt. Our new method is compatible with current architecture dependencies, so it can be integrated seamlessly with the existing environment without causing dependency conflicts.

Our bias mitigation approach leads to a more balanced exam assembly, mitigating bias on EdTec-QBuilder, our partner's exam assembly platform, by optimizing test item selection while maintaining a well-distributed mix of difficulty levels and preserving high topical relevance. The proposed MILP-driven re-ranking strategy functions as the backend search mechanism of the enhanced system version, ensuring items align with fairer difficulty level constraints. To enhance EdTec-QBuilder's transparency and user control, the updated version introduces a graphical user interface that visualizes difficulty imbalance, allowing exam designers to monitor and refine the overall difficulty distribution for an exam more effectively.

Beyond improving fairness in ranking, our approach holds practical significance for VET services, particularly in manual exam assembly and assessment settings where exam validity depends on diverse and unbiased test item selection. EdTec-QBuilder's former item retrieval and exam assembly system failed to account for difficulty and source-based imbalances, leading to biased test compositions that affected learners' evaluation outcomes. By integrating MILP-driven bias mitigation, our method ensures that exams are more representative, supporting psychometric integrity in vocational assessment. This advancement aligns with broader trends in fair information retrieval and algorithmic transparency, where unbiased ranking is increasingly valued in education, commercial search applications, hiring platforms, and recommendation systems.

The demand for unbiased exam assembly methods is growing among educational and high-stakes assessment organizations (Linden et al., 2005; Lane et al., 2016; Palomino et al., 2024; Bißantz et al., 2024). More broadly, ensuring fairness in information retrieval is essential not only in education but also in commercial domains where ranking biases impact access to opportunities and decision-making, such as e-commerce and hiring platforms (Yin and Jeffries, 2021; Bhadani, 2021; Özer et al., 2024). By enhancing the fairness of test item retrieval and assembly, our approach contributes to both assessment quality in VET services and broader advancements in unbiased ranking methodologies.

[3] Demo fork available at: https://www.dfki.de/kiperweb/about.html

Performance Metrics for Re-Ranking Methods @50

#	Method	Core Embedding Model	nDCG	MRR	Prec.	Rec.	F1	MAP	AWRF (Dif.)	AWRF (Src.)	AWRF (Avg.)	JM (Src.)	JM (Dif.)	JM (Avg.)
1	CMAB	gbert-large-paraphrase-euclidean	0.28	0.63	0.23	0.19	0.20	0.08	0.47	0.58	0.52	0.12	0.15	0.13
2		e5-multi-sml-torch	0.25	0.54	0.21	0.18	0.18	0.06	0.32	0.62	0.47	0.09	0.16	0.13
3		gbert-large-paraphrase-cosine	0.26	0.52	0.22	0.19	0.19	0.07	0.47	0.55	0.51	0.11	0.13	0.12
4	DetConstSort	gbert-large-paraphrase-euclidean	0.28	0.50	0.24	0.20	0.21	0.09	0.48	0.55	0.52	0.12	0.14	0.13
5		e5-base-multilingual-4096	0.21	0.40	0.20	0.17	0.17	0.05	0.32	0.47	0.40	0.11	0.14	0.13
6		gbert-large-paraphrase-cosine	0.27	0.48	0.23	0.20	0.20	0.09	0.50	0.55	0.53	0.12	0.13	0.12
7	ϵ-greedy	gbert-large-paraphrase-euclidean	0.31	0.53	0.29	0.24	0.25	0.09	0.41	0.55	0.48	0.14	0.18	0.16
8		multilingual-mpnet-base-v2	0.32	0.45	0.30	0.26	0.27	0.10	0.39	0.63	0.51	0.11	0.19	0.15
9		e5-multi-sml-torch	0.27	0.48	0.25	0.21	0.22	0.07	0.33	0.57	0.45	0.11	0.18	0.14
10	FA*IR	gbert-large-paraphrase-cosine	0.37	0.41	0.38	0.34	0.34	0.13	0.41	0.51	0.46	0.18	0.22	0.20
11		multilingual-mpnet-base-v2	0.36	0.44	0.36	0.31	0.32	0.12	0.32	0.53	0.42	0.16	0.24	0.20
12		gbert-large-paraphrase-euclidean	0.29	0.38	0.30	0.27	0.27	0.09	0.26	0.36	0.31	0.18	0.21	0.20
13	MILP-LLM	gbert-large-paraphrase-cosine	0.45	0.67	0.39	0.34	0.35	0.20	0.32	0.52	0.42	0.21	0.30	0.25
14		gbert-large-paraphrase-euclidean	0.44	0.71	0.38	0.33	0.34	0.20	0.34	0.52	0.43	0.21	0.29	0.25
15		efederici_e5-base-multilingual-4096	0.34	0.59	0.29	0.26	0.27	0.14	0.27	0.35	0.31	0.22	0.25	0.23
16	MILP-BOpt	gbert-large-paraphrase-cosine	0.45	0.67	0.39	0.34	0.35	0.20	0.32	0.54	0.43	0.20	0.30	0.25
17		gbert-large-paraphrase-euclidean	0.43	0.71	0.38	0.33	0.34	0.20	0.33	0.52	0.43	0.20	0.29	0.24
18		e5-base-multilingual-4096	0.34	0.59	0.29	0.26	0.27	0.14	0.27	0.35	0.31	0.22	0.25	0.23
19	MILP	gbert-large-paraphrase-cosine	0.43	0.64	0.39	0.34	0.35	0.18	0.30	0.52	0.41	0.20	0.29	0.25
20		gbert-large-paraphrase-euclidean	0.41	0.65	0.38	0.33	0.34	0.17	0.33	0.52	0.42	0.19	0.28	0.24
21		multilingual-mpnet-base-v2	0.40	0.60	0.36	0.31	0.32	0.16	0.36	0.53	0.44	0.18	0.25	0.22
22	MMR	gbert-large-paraphrase-euclidean	0.28	0.58	0.23	0.19	0.19	0.08	0.49	0.58	0.54	0.11	0.14	0.12
23		e5-base-multilingual-4096	0.22	0.52	0.19	0.16	0.17	0.05	0.36	0.50	0.43	0.11	0.14	0.12
24		multilingual-e5-base	0.21	0.50	0.19	0.15	0.16	0.04	0.35	0.50	0.42	0.10	0.13	0.12
25	Random	gbert-large-paraphrase-euclidean	0.33	0.54	0.30	0.26	0.26	0.11	0.37	0.47	0.42	0.17	0.20	0.19
26		multilingual-mpnet-base-v2	0.35	0.55	0.32	0.27	0.28	0.13	0.37	0.60	0.49	0.13	0.22	0.17
27		gbert-large-paraphrase-cosine	0.31	0.40	0.30	0.26	0.26	0.10	0.38	0.52	0.45	0.14	0.19	0.17

Table 2: Performance metrics for various re-ranking methods, evaluated at a cutoff of 50. These methods optimize performance over an initial pool of 100 items retrieved using ANN+CE model #1 described in (Palomino et al., 2024).

7 Conclusions

We conducted 225 experiments using the industry benchmark from Palomino et al. (2024) as a baseline to evaluate nine distinct bias mitigation re-rankers, each designed to address the difficulty and source bias in EdTec-QBuilder, bfz's item retrieval and exam assembly platform. Enhanced by advanced contextualization through refined query and relevance generation and optimized via Bayesian hyperparameter tuning, our new MILP-driven re-ranking method achieved a 17% increase in nDCG while reducing AWRF by 9% compared to previous results. Our approach outperformed popular bias mitigation re-ranking methods in our task, underscoring the suitability of mathematical optimization techniques for mitigating bias in commercial search systems. Future work should explore leveraging alternative optimization paradigms, such as multi-objective and nonlinear programming, and in-training techniques, including bias-aware loss functions and regularization for bias mitigation in neural ranking models.

Limitations and Ethics Statement

We anonymized all sensitive information in the data used for this work and maintained strict confidentiality to protect our partner's product and intellectual property, in full compliance with required privacy standards. Unbiased exam assembly is paramount to ensuring assessment equality and fairness; when exams are not optimally assembled, attribute biases may skew evaluations and undermine the validity of the assessment process—particularly in high-stakes scenarios where test takers must demonstrate competence at a specific knowledge level. Leveraging algorithmic and transparent methods, as presented in our approach, fosters transparency in exam construction. While our MILP-driven re-ranking approach improved performance by reducing bias and enhancing ranking relevance on EdTec-QBuilder (bfz's exam assembly platform), it may struggle to mitigate other attribute-based biases as exam specifications, constraints, and candidate rankings become more complex.

A potential limitation arises when incorporating additional test item attributes into the MILP formulation, especially with a larger item base. Expanding the model to account for attributes such as topic relevance to specific skills, cognitive complexity (e.g., recall vs. application), item format (e.g., multiple-choice vs. open-ended), language level, or domain-specific prerequisites could significantly increase computational complexity. As more attributes are introduced, the problem may become harder to solve efficiently, potentially impacting runtime performance. Nevertheless, in our setup—given the specific use case and restrictions—our approach demonstrated computational efficiency, consistently finding solutions within milliseconds, thereby making it suitable for real-time or near real-time applications, as evidenced by the demo fork of our tool. Approximation heuristics, such as warm starts, cutting strategies, and parallel solving, could help maintain efficiency even in more complex scenarios.

Although our method does not determine contextual study group cohorts for making recommendations, it is not yet capable of identifying the most relevant items for a given learning group's progress. Consequently, we delegate this decision to vocational trainers using our tool.

Acknowledgements

This research was funded by the Federal Ministry of Education and Research (BMBF) in Germany [Grant 21IVP056C] as part of AZUBOT. The project was implemented under the leadership of f-bb in collaboration with the German Research Center for Artificial Intelligence (DFKI), the Institute for Vocational Education and Training (IFBB), the Bildungswerk der Niedersächsischen Wirtschaft (BNW), oncampus GmbH, and Provadis Partner für Bildung und Beratung GmbH. Additionally, this research aligns with KIPerWeb, which explores AI-supported personalization in vocational training. The views expressed are those of the authors and do not necessarily reflect those of the BMBF.

References

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Elias Bassani. 2022. ranx: A blazing-fast python library for ranking evaluation and comparison. In *ECIR (2)*, volume 13186 of *Lecture Notes in Computer Science*, pages 259–264. Springer.

Nolwenn Bernard and Krisztian Balog. 2023. A systematic review of fairness, accountability, transparency and ethics in information retrieval. *ACM Comput. Surv.* Just Accepted.

Donald A Berry and Bert Fristedt. 1985. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5(71-87):7–7.

Saumya Bhadani. 2021. Biases in recommendation system. In *Proceedings of the 15th ACM conference on recommender systems*, pages 855–859.

Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, and Sebastian Kohlmeier. 2019. Overview of the trec 2019 fair ranking track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*.

Steven Bißantz, Susanne Frick, Filip Melinscak, Dragos Iliescu, and Eunike Wetzel. 2024. The potential of machine learning methods in psychological assessment and test construction. *European Journal of Psychological Assessment*, 40(1):1–4. [Editorial].

Ian Burke, Robin Burke, and Goran Kuljanin. 2021. Fair candidate ranking with spatial partitioning: Lessons from the siop ml competition. In *HR@ RecSys*.

Kathleen Cachel and Elke Rundensteiner. 2024. Fairranktune: A python toolkit for fair ranking tasks. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5195–5199, New York, NY, USA. Association for Computing Machinery.

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*, pages 335–336.

L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2018. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447.

Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. 2022. Overview of the trec 2021 fair ranking track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*.

Yunhe Feng and Chirag Shah. 2022. Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11882–11890.

Ruoyuan Gao and Chirag Shah. 2021. Addressing bias and fairness in search systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2643–2646, New York, NY, USA. Association for Computing Machinery.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 2221–2231.

Philipp Hager, Romain Deffayet, Jean-Michel Renders, Onno Zoeter, and Maarten de Rijke. 2024. Unbiased learning to rank meets reality: Lessons from baidu's large-scale search dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1546–1556, New York, NY, USA. Association for Computing Machinery.

Tim Head, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi. 2021. scikit-optimize/scikit-optimize.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ömer Kırnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of fair ranking metrics with incomplete judgments. In *Proceedings of the Web Conference 2021*, pages 1065–1075.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

Suzanne Lane, Mark R Raymond, Thomas M Haladyna, et al. 2016. *Handbook of test development*, volume 2. Routledge New York, NY.

Yuantong Li, Xiaokai Wei, Zijian Wang, Shen Wang, Parminder Bhatia, Xiaofei Ma, and Andrew Arnold. 2022. Debiasing neural retrieval via in-batch balancing regularization. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 58–66, Seattle, Washington. Association for Computational Linguistics.

Wim J Linden et al. 2005. *Linear models for optimal test design*. Springer.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 429–438, New York, NY, USA. Association for Computing Machinery.

Özalp Özer, A Serdar Şimşek, Xiaoxi Zhao, Ethan Dee, and Vivian Yu. 2024. Measuring the efficacy of amazon recommendation systems. In *Tutorials in Operations Research: Smarter Decisions for a Better World*, pages 224–243. INFORMS.

Alonso Palomino, Andreas Fischer, Jakub Kuzilek, Jarek Nitsch, Niels Pinkwart, and Benjamin Paassen. 2024. EdTec-QBuilder: A semantic retrieval tool for assembling vocational training exams in German language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 26–35, Mexico City, Mexico. Association for Computational Linguistics.

Amifa Raj and Michael D Ekstrand. 2020. Comparing fair ranking metrics. *arXiv preprint arXiv:2009.01311*.

Amifa Raj and Michael D Ekstrand. 2022. Measuring fairness in ranked results: an analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–736.

Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 553–562, New York, NY, USA. Association for Computing Machinery.

Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. 2023. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 55(13s):1–39.

Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2219–2228.

Emily Strong, Bernard Kleynhans, and Serdar Kadıoğlu. 2021. Mabwiser: Parallelizable contextual multi-armed bandits. *International Journal on Artificial Intelligence Tools*, 30(04):2150021.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Thibaut Thonet and Jean-Michel Renders. 2020. Multi-grouping robust fair ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2077–2080.

Shicheng Wang, Shu Guo, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2023. Hdnr: A hyperbolic-based debiased approach for personalized news recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 259–268.

L Yin and A Jeffries. 2021. How we analyzed amazon's treatment of its "brands" in search results. *The Markup*, 1.

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*, pages 1569–1578.

Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 449–458, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Pre-trained Sentence Similarity Models

Table 3 provides a comprehensive summary of the pre-trained semantic sentence similarity models utilized in our experiments. These models formed the foundation of the embedding-based ANN+CE search framework described in prior work (Palomino et al., 2024). The outputs of these core embedding models served as the candidate pools for applying the proposed re-ranking methods (Section 5), enabling the benchmarking of bias mitigation strategies and relevance optimization techniques for our item retrieval for exam assembly task.

A.2 Pareto Methods

From our exhaustive analysis, we observed that mitigating bias in our task depends on optimally balancing the relevance/bias tradeoff as much as possible. Figure 2 shows the Pareto frontier trade-off between relevance and fairness, highlighting the optimal methods that best balance nDCG and AWRF, where improved performance on one metric could worsen the other. We observed that our proposed MILP-driven bias re-rankers successfully balanced the relevance/bias tradeoff represented by nDCG and average AWRF as optimally as possible.

#	Models for ANN Search
1	paraphrase-multilingual-mpnet-base-v2
2	German_Semantic_STS_V2
3	LaBSE
4	bi-encoder_msmarco_bert-base_german
5	e5-base-multilingual-4096
6	multilingual-e5-base
7	mfaq
8	sts_paraphrase_xlm-roberta-base_de-en
9	gbert-large-paraphrase-euclidean
10	all-MiniLM-L12-v2-embedding-all
11	paraphrase-multilingual-mpnet-base-v2-embedding-all
12	distiluse-base-multilingual-cased-v1
13	distiluse-base-multilingual-cased-v2
14	gbert-large-paraphrase-cosine
15	text2vec-base-multilingual
16	German-semantic
17	LaBSE
18	sn-xlm-roberta-base-snli-mnli-anli-xnli
19	musterdatenkatalog_clf
20	debatenet-2-cat
21	LEALLA-large
22	lt-wikidata-comp-de
23	e5-multi-sml-torch
24	text2vec-base-multilingual
25	Llama-2-7b-chat-hf

Table 3: Complete list of tested language models for ANN-based nearest neighbor search

A.2.1 Path to an Enhanced System Architecture for Improved Retrieval Performance

All experiments were conducted on a macOS with an ARM64 processor (32 GB RAM, 12 cores). We plan to deploy MILP-BOpt—our best-performing bias mitigation re-ranking method—by integrating it into the EdTec-QBuilder architecture (see Figure 3). The system starts with a standard ANN+CE search over a 100-item candidate pool using pre-calculated, offline-stored item embeddings for efficiency. This is followed by query expansion via asynchronous API calls to GPT-4o, which generates real-time embeddings to boost relevance scores. MILP-BOpt then dynamically computes

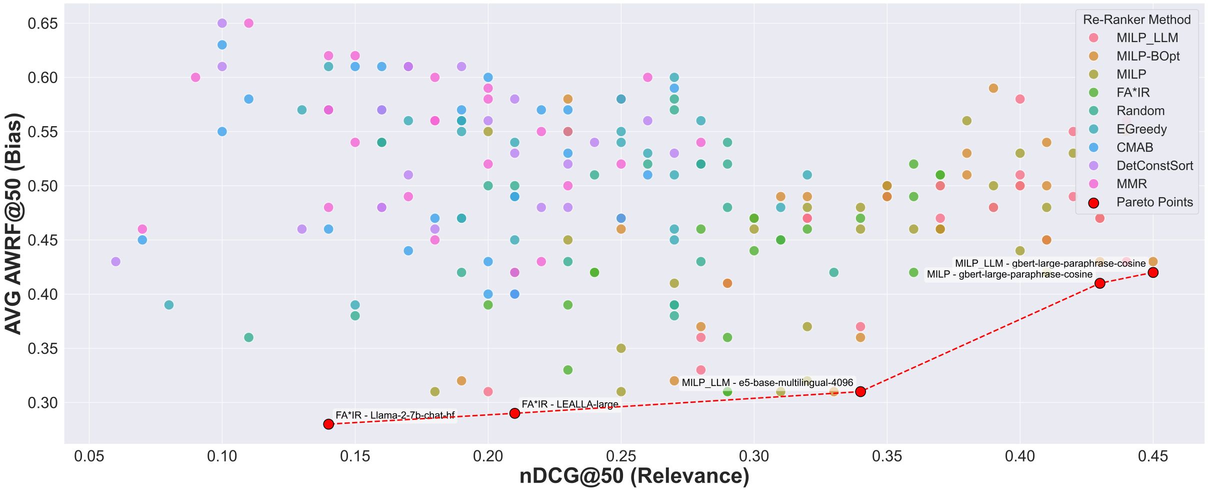

Figure 2: Comparing re-ranking methods: achieving optimal balance between relevance (nDCG) and bias (Avg. AWRF) trade-offs.

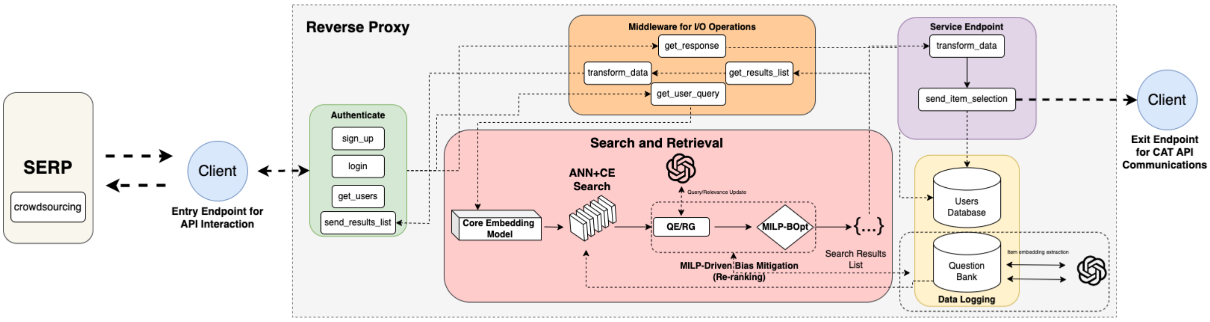

Figure 3: Pre-deployment testing architecture of EdTec-Builder, illustrating client interactions, API endpoints, the core ANN+CE search with the MILP-BOpt bias mitigation re-ranker method, and integration with authentication, logging, and external data sources.

the optimal lambda trade-off between relevance and bias mitigation using Bayesian optimization with multithreading via Head et al. (2021). Finally, the platform's UI displays an improved ranking that enables manual exam designers to select items more comprehensively. A live pre-deployment demo was developed to evaluate MILP-BOpt in a real-world integration test. Pre-deployment tests using the `SciPy` library indicate that minimal architectural changes are needed for this enhancement; however, as the system scales, computational efficiency may be further improved with advanced parallelization and warm-start techniques.

A.3 Prompting Strategy for LLM-Based Query Expansion

Building on (Sun et al., 2023), we used a zero-shot prompting strategy with strict output validation to improve skill-based query expansion and contextual relevance by incorporating related terms. The prompting process was structured as follows:

1. We generated a prompt for each query, strictly requesting the top essential skill terms related to the original query.
2. We configured a deterministic output by setting GPT-4 with: (a) Temperature: 0.0 and (b) Top_p: 1.0 (no nucleus sampling).
3. We used a Pydantic model to validate a list-based schema, ensuring consistent skill extraction. The expanded query is updated by concatenating it with the newly extracted skills.

Our approach yields deterministic skill expansion, consistent output handling, and prevents malformed responses. We employed OpenAI's `text-embedding-three` large model to compute semantic similarity scores. After query expansion, we calculated cosine similarity between the expanded queries and embedded items. This process complements MILP-BOpt and improves relevance scores.