Auditing Bias in AI-Based Hiring Systems: A Fairness Analysis of Nationality and Gender Discrimination

# Auditing Bias in AI-Based Hiring Systems: A Fairness Analysis of Nationality and Gender Discrimination

Xhoana Shkajoti
Politecnico di Torino, Italy
Turin, Italy
xhoana.shkajoti@studenti.polito.it

Martina Ullasci, Marco Rondina
Riccardo Coppola, Antonio Vetrò, Luca Calò
Politecnico di Torino, Italy
Turin, Italy
first.last@polito.it

## Abstract

In recent years, artificial intelligence (AI) systems have become increasingly integrated into recruitment processes, particularly in early-stage candidate screening based on semantic matching between CVs and job descriptions. While embedding-based models promise efficiency and scalability, they also raise concerns regarding fairness, as they may encode and propagate historical social biases present in training data. This study presents a controlled and reproducible audit of a hiring pipeline built on Sentence-BERT (SBERT) to verify disparate outcomes related to gender and nationality. Using a synthetic dataset of matched candidate profiles, we perform a counterfactual analysis across three operational levels: similarity scores, threshold-based screening decisions, and Top-K ranking outcomes. Results reveals systematic disparities in candidate visibility according to their gender and nationality. While score-level differences are small in magnitude, they have a significant impact on screening and ranking decisions, with female candidates and Italian profiles being the most disadvantaged groups. This work contributes a transparent audit framework and provides empirical evidence supporting the need for usage-oriented fairness evaluation in AI-based hiring systems.

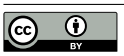
## CCS Concepts

• **Software and its engineering → Extra-functional properties**; • **Social and professional topics → User characteristics**; **Race and ethnicity**; **Gender**;

## Keywords

Algorithmic Fairness, Embedding-based Matching, Sentence-BERT, AI Hiring Bias, Recruitment Systems Audit

**ACM Reference Format:**
Xhoana Shkajoti and Martina Ullasci, Marco Rondina, Riccardo Coppola, Antonio Vetrò, Luca Calò. 2026. Auditing Bias in AI-Based Hiring Systems: A Fairness Analysis of Nationality and Gender Discrimination. In *International Conference on Information Technology for Social Good (GoodIT '26), September 02–04, 2026, Pisa, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3794786.3830781




*GoodIT '26, Pisa, Italy*

ACM ISBN 979-8-4007-2483-1/2026/09
https://doi.org/10.1145/3794786.3830781

## 1 Introduction

Natural Language Processing (NLP) models are increasingly used in recruitment processes to analyse candidate profiles and to support hiring decisions at scale. Human Resources (HR) platforms integrate AI-driven tools to parse resumes, extract skills, and assist candidate screening, offering improvements in efficiency and scalability [27, 30] but also raising critical concerns regarding fairness and accountability: the EU AI Act classifies AI systems used in candidate evaluation as high-risk, requiring transparency, fairness, and human oversight [10–12]. In fact, candidate evaluation systems may perpetuate disparities if certain profiles are consistently favoured or disadvantaged. Research shows that language models can encode the social and demographic associations present in the training data [5, 6, 14, 15], affecting hiring outcomes [8, 21, 29].

Similarity-based recruitment systems, which rank candidates by computing the semantic similarity between CVs and job descriptions using embedding models, are implementations of AI-driven screening. Existing work on embedding-based matching often focuses on performance and retrieval quality [1, 9], without directly linking representation bias to decision outcomes [19].

This work is framed within the context of a fairness audit for software engineering. It treats demographic disparities as extra-functional properties of the system that can be detected empirically in its outputs, independently of the system's primary performance objective. This study aims to examine how demographic cues influence embedding-based candidate–job matching systems. We analyse an embedding-based candidate–job matching pipeline, simulating an early-stage screening scenario in which candidate profiles are compared to job descriptions and ranked based on similarity scores. To isolate the effect of demographic attributes, we construct a synthetic dataset in which profiles are equivalent in job-relevant content. The objective is to evaluate whether an SBERT-based [24] matching pipeline produces disparities across gender and nationality by analysing similarity scores, threshold-based screening, and top-k ranking outcomes on a controlled synthetic dataset. The main contributions of this work are threefold: 1) we introduce a reproducible audit framework; 2) we analyse how disparities propagate from score to decision; 3) we provide a statistical analysis of these effects across different job contexts.

The remainder of the paper is organised as follows. Section 2 provides background on gender bias, algorithmic fairness, and language models in recruitment. The research questions are presented in Section 3, with the description of the methods and experimental settings following in Section 4. Section 5 and 6 respectively reports and discuss the results. We conclude outlining threats to validity (Section 7) and directions for future work (Section 8).

## 2 Background

The use of embedding-based AI systems in recruitment has raised increasing concerns regarding bias and fairness in automated candidate evaluation. These systems are now widely deployed to support large-scale hiring processes [27, 30], making fairness a critical requirement. Embedding-based models represent both candidate profiles and job descriptions in a shared space, enabling comparison based on semantic comparison [1, 9], computed using distance measures such as cosine similarity [24]. This formulation enables more flexible matching and has been widely adopted in large-scale retrieval and recommendation tasks. Among the models adopting this paradigm, Sentence-BERT (SBERT) [24] is particularly well-suited for recruitment applications: unlike cross-encoder architectures that require the joint processing of text pairs, SBERT encodes each input independently, making it computationally efficient for scenarios involving large numbers of candidate–job comparisons. In such contexts, a growing body of research shows that models trained on large-scale textual data may encode statistical associations present in language, leading to disparate impact across groups, even when individuals are comparable in qualifications [5, 6, 14, 19]. These effects can manifest in both representational forms—through the encoding of social associations—and allocation outcomes, where differences in model outputs translate into unequal access to opportunities. This distinction is particularly relevant in software engineering contexts, where model outputs are integrated into end-to-end systems that support decision-making processes. In recruitment pipelines, similarity scores generated by embedding-based models are used to rank candidates and to apply threshold-based filtering, directly determining candidate visibility and progression in the hiring process [21, 29]. A key challenge in analysing these systems is that they are often deployed as black-box components. The internal structure of the model, the training data, and the optimisation process are typically not accessible, limiting the possibility of direct inspection. As a result, understanding system behaviour requires analysing outputs under controlled input conditions [7, 22].

To enable such analysis, prior work has adopted audit-based approaches using synthetic data [26, 32]. By constructing candidate profiles that are equivalent in job-relevant content while varying specific attributes, it is possible to isolate the effect of these variations on similarity scores and ranking outcomes. This approach also avoids the use of real personal data, addressing privacy and regulatory constraints in recruitment contexts [3]. Building on the mentioned prior studies, this work proposes a pipeline-level fairness analysis, that extends beyond score-level evaluation to include screening and ranking stages, using a controlled synthetic dataset to isolate the effect of gender and nationality while keeping professional qualifications constant.

## 3 Research Goal and Questions

This study aims to evaluate whether an embedding-based CV–job matching system produces systematic disparities with respect to protected attributes, specifically gender and nationality, when candidates are equivalent in terms of qualifications. The overall research goal is defined following the Goal–Question–Metric (GQM) framework [28] in Table 1. This framework enables the explicit linkage between the study objective, the research questions, and the quantitative measures used to assess model behaviour.

The analysis focuses on controlled synthetic candidate profiles with equivalent job-relevant content, allowing the isolation of demographic effects while avoiding confounding factors related to differences in qualifications or experience. The evaluation is conducted within a similarity-based matching setting using an SBERT model, and considers multiple job hierarchy levels to capture potential context-dependent variations.

The study is structured around the following research questions:

- **RQ1 (Score):** Do similarity scores differ systematically across candidates based on gender or nationality when qualifications are held constant?
- **RQ2 (Screening):** Do differences in similarity scores translate into unequal screening outcomes when threshold-based selection is applied?
- **RQ3 (Ranking):** Does the ranking of candidates lead to disparities in candidate visibility, particularly in top-ranked positions (Top-K), across demographic groups?
- **RQ4 (Influence of hierarchy):** Does the level of job seniority influence the magnitude of disparities in score, selection, and ranking outcomes?

Each research question is addressed using quantitative measures aligned with the corresponding stage of the analysis. For **RQ1**, the cosine similarity between the numerical representation of a candidate's CV and that of a job description. Disparities are measured using differences in cosine similarity scores across matched candidate profiles, where only one demographic attribute is varied at a time. In **RQ2**, screening refers to the selection of candidates whose similarity score exceeds a predefined threshold: disparities are measured through selection rates, defined as the proportion of candidates exceeding predefined similarity thresholds. In **RQ3**, ranking refers to the ordering of candidates based on similarity scores. Disparities are measured using Top-K inclusion rates, defined as the proportion of candidates from each group appearing among the highest-ranked positions. For **RQ4**, disparities are analysed across hierarchy levels using score differences, selection rates, and inclusion rate differences across groups.

**Table 1: GQM goal definition.**

| Element | Definition |
|---|---|
| *Analyze* | embedding-based CV–job matching system |
| *for the purpose of* | evaluating demographic disparities in candidate evaluation |
| *with respect to* | similarity scores, screening outcomes, and ranking positions |
| *from the point of view of* | fairness-aware researchers and practitioners |
| *in the context of* | synthetic CVs with controlled qualifications and multiple hierarchy levels |

# 4 Methodology

This study adopts a controlled experimental design to evaluate whether an embedding-based CV–job matching system produces systematic disparities related to gender and nationality. The approach is designed to isolate the effect of demographic attributes while maintaining a realistic representation of similarity-based recruitment systems. To achieve this, the methodology combines a synthetic dataset with a pipeline-level evaluation: this allows the analysis of model behaviour across multiple stages of the recruitment process, including score computation, screening, and ranking, under controlled conditions. The code used in this experiment is available in the online replication package[1].

## 4.1 Model and Representation

The evaluation is conducted using the Sentence-BERT model all-MiniLM-L6-v2, a lightweight transformer-based architecture designed for semantic similarity tasks [24]. The model encodes each input text independently into a fixed-length vector representation of 384 dimensions. The choice of this model is motivated by its balance between computational efficiency and semantic performance: with approximately 22 million parameters, all-MiniLM-L6-v2 enables fast embedding generation while maintaining sufficient accuracy for similarity-based matching tasks. Its design builds on transformer-based architectures and compressed models such as MiniLM, which aim to retain performance while reducing computational cost. Its open-source availability also supports transparency and reproducibility, which are essential for audit-based analysis. All inputs are processed in English, using a monolingual setting to ensure consistent semantic representations and to avoid variability introduced by multilingual or translation effects [20]. To obtain a single representation for each CV or job description, we apply mean pooling over the token embeddings produced by the transformer [24]. This aggregation step produces a single numerical vector representation that captures the overall semantic content of the input text. These embeddings are then used to compute similarity scores between candidate profiles and job descriptions.

## 4.2 Synthetic Dataset Construction

We construct a synthetic dataset of candidate profiles designed to isolate the effect of demographic attributes while preserving realistic professional content. This allows us to avoid confounding factors present in real-world CVs, where qualifications and demographic characteristics are correlated and cannot be separated [3, 4]. Each base CV represents a candidate profile with fixed professional content, including education, work experience, and skills. From each base profile, multiple variants are generated by modifying only demographic tokens *[gender]* and *[nationality]*, while keeping all job-relevant information identical. This approach aligns with counterfactual fairness, where outcomes are compared across hypothetical scenarios differing only in sensitive attributes [17].

The dataset is constructed using a full factorial design as shown in Table 2 combining four job roles (Project Manager, Market Research Analyst, HR Specialist, and Content/Creative Manager), four hierarchy levels (junior, mid-level, senior, managerial), four nationalities (Italian, Albanian, British, Moroccan), two genders (male, female), and three lexical versions (V1: formal and institutional, V2: action-oriented, V3: technical and competency-based), resulting in a total of $4 \times 4 \times 4 \times 2 \times 3 = 384$ CV instances. The selected nationalities are intended to reflect diverse labour market and migration contexts surrounding Europe. Italy is used as the EU reference category and baseline for comparison, while the United Kingdom represents a Western Anglo-European profile. Albania captures a non-EU but European migrant group, and Morocco representing a context where labour market disparities have been empirically documented in prior work [23]. This selection enables the analysis of model behaviour across nationality cues that differ in geographic, cultural, and socio-economic positioning.

**Table 2: Variables included in the factorial design.**

| Dimension | # | Levels |
|---|---|---|
| Job roles | 4 | Project Manager |
| | | Market Research Analyst |
| | | HR Specialist |
| | | Content/Creative Manager |
| Hierarchy levels | 4 | Junior (0–2 years) |
| | | Mid-level (3–5 years) |
| | | Senior (6–10 years) |
| | | Lead/Manager (10+ years) |
| Nationality | 4 | Italian, Albanian, British, Moroccan |
| Gender | 2 | Male, Female |
| Lexical versions | 3 | V1: formal and institutional |
| | | V2: action-oriented |
| | | V3: technical and competency-based |

Candidate profiles and job descriptions are constructed using the ISCO-08 framework as a reference point. This is an internationally recognised classification of occupations and skill levels, and can be found in [16]. This framework provides clear and consistent definitions of job roles, responsibilities and levels of seniority. This ensures that differences in model outputs are not driven by inconsistencies in how roles are described, but reflect the controlled variables of the experiment. Hierarchy levels are defined using predefined ranges of years of experience and corresponding responsibilities, ensuring a consistent and structured representation of seniority across profiles. This allows the analysis to capture potential variation in model behaviour across stages of the professional lifecycle while maintaining comparable role structures.

To reduce the influence of wording and writing style, each profile is expressed in three lexical versions that describe the same professional content using different tones. Across all versions, the content remains identical, and only the formulation changes. Version **V1** follows a *formal and standardised tone* aligned with ISCO-08[16] descriptions, using neutral and descriptive language. Version **V2** adopts a more *action-oriented and assertive tone*, emphasising impact and achievements. Version **V3** uses a *technical and competency-based tone*, focusing on tools, methods, and specific skills. This design helps mitigate linguistic variability, which is known to influence embedding-based representations [31].

Finally, to avoid introducing bias due to differences in text length, all profiles are constructed to be similar in length. Each CV is kept

[1] https://anonymous.4open.science/r/Auditing-Bias-in-AI-Based-Hiring-Systems-27C6/README.md

close to an average length of approximately 200 words to cater to the models restriction of maximum number of tokens. This ensures that differences in similarity scores are not driven by input length but reflect how the model processes the content.

## 4.3 Matching Pipeline

The evaluation follows a similarity-based matching pipeline, a common approach in modern recruitment systems based on semantic search and embedding representations [25]. Each candidate CV and job description is provided as input text to the model. The model encodes both inputs independently into fixed-length vector representations. Given a candidate embedding $\mathbf{e}_c$ and a job embedding $\mathbf{e}_j$, their similarity is computed using cosine similarity:

$$\text{sim}(c, j) = \frac{\mathbf{e}_c \cdot \mathbf{e}_j}{\|\mathbf{e}_c\| \, \|\mathbf{e}_j\|} \tag{1}$$

Cosine similarity is widely used in embedding-based retrieval and semantic matching tasks to measure the angular distance between vector representations [24].

For each candidate profile, similarity scores are computed against the corresponding job descriptions. All counterfactual variants of the same profile are evaluated against the same job inputs, ensuring that any differences in scores are attributable only to the modified demographic attributes. This setup focuses on the embedding and similarity component rather than a full recruitment system. Similar approaches have been adopted in prior work on recruitment, where candidate profiles and job descriptions are encoded into vector representations and compared using similarity-based scoring. In particular, recent studies apply SBERT and related models to resume–job matching and candidate ranking tasks, where similarity scores directly guide selection decisions [1, 9, 18].

In this context, the use of SBERT in this work should be understood as a representative instance of embedding-based matching, rather than as the evaluation of a specific model. This formulation aligns with broader approaches in AI-driven recruitment systems[27]. Accordingly, this study focuses on the embedding and similarity component in isolation, following audit-based methodologies that analyse individual system components to identify potential sources of bias [22].

## 4.4 Statistical Evaluation

The statistical analysis builds on the controlled structure of the synthetic dataset, in which profiles are matched so that the professional content remains constant while the demographic attributes vary. This enables the direct comparison of model outputs under controlled conditions, which is consistent with the experimental evaluation approaches used in fairness research [2, 19]. The evaluation is conducted at three levels: score, screening, and ranking.

**Score-level analysis**. Similarity scores are compared across matched profiles. For gender, paired comparisons are performed between male and female variants of the same profile. For nationality, comparisons are made across matched groups, using the Italian profile as a reference. A paired t-test is used to assess whether the mean difference in scores is significantly different from zero.

**Table 3: Summary of statistical tests across analysis stages.**

| Hypothesis | p-value | Decision |
|---|---|---|
| $H_1$: Demographic attributes have no impact on similarity scores | | |
| $H_{1.1}$ Gender | $2.42 \times 10^{-58}$ | Reject |
| $H_{1.2}$ Nationality | $< 8.86 \times 10^{-12}$ | Reject |
| $H_2$: Demographic attributes have no impact on screening outcomes | | |
| $H_{2.1}$ Gender | | |
| Top 10% | 0.5000 | Accept |
| Top 20% | 1.0000 | Accept |
| Top 30% | 0.0156 | Reject |
| $H_{2.2}$ Nationality | | |
| Top 10% | $1.40 \times 10^{-4}$ | Reject |
| Top 20% | $2.22 \times 10^{-5}$ | Reject |
| Top 30% | $1.26 \times 10^{-3}$ | Reject |
| $H_3$: Demographic attributes have no impact on Top-K inclusion | | |
| $H_{3.1}$ Gender | | |
| Top-1 | $7.11 \times 10^{-15}$ | Reject |
| Top-3 | $2.78 \times 10^{-17}$ | Reject |
| Top-5 | $1.65 \times 10^{-13}$ | Reject |
| $H_{3.2}$ Nationality | | |
| Top-1 | $< 10^{-16}$ | Reject |
| Top-3 | $< 10^{-16}$ | Reject |
| Top-5 | $1.00 \times 10^{-12}$ | Reject |
| $H_4$: Disparities do not vary across hierarchy levels | Mixed | Accept |

**Threshold-based analysis.** To simulate screening decisions, similarity scores are converted into binary outcomes using Top-10%, Top-20%, and Top-30% thresholds, reflecting common filtering practices in applicant tracking systems [30]. Candidates above the threshold are considered selected. For gender, differences in selection outcomes are evaluated using McNemar's test on paired profiles. For nationality, Cochran's Q test is used to compare selection probabilities across multiple groups.

**Ranking-based analysis.** At the ranking level, candidates are ordered by similarity score and evaluated in terms of Top-K inclusion (K = 1, 3, 5), a standard evaluation setting in information retrieval systems [24]. For gender, differences in inclusion rates are assessed using McNemar's test. For nationality, Cochran's Q test is applied. For both screening and ranking, Risk Difference (RD) and Risk Ratio (RR) are used as descriptive measures of disparities between groups. In this context, "risk" refers to the probability of selection or inclusion (i.e., selection rate or Top-K inclusion rate). RD captures absolute differences in these rates, while RR indicates how many times more likely one group is to be selected or included than another. The analysis is conducted both at aggregate level and separately across hierarchy levels, to examine whether the patterns remain consistent across job contexts.

For all tests, statistical significance is evaluated at the $\alpha = 0.05$ threshold. Results are interpreted by combining statistical significance with descriptive measures, ensuring that both the presence and the magnitude of disparities are taken into account.

# 5 Results

The empirical findings of the study are structured according to the main stages of the recruitment pipeline: score computation, threshold-based screening, and ranking. The results of the statistical analysis for each hypothesis are illustrated in Table 3.

## 5.1 Score-based analysis

We test whether candidates with equivalent qualifications receive comparable similarity scores across demographic groups: the null hypothesis ($H_0$) states that there is no systematic difference.

*5.1.1 Gender.* Figure 1 shows the distribution of pairwise differences in cosine similarity scores between male and female candidate profiles (M – F) matched to the same job descriptions. Positive values indicate higher scores for male profiles, while negative values indicate higher scores for female profiles. The distribution is predominantly shifted towards positive values, with male profiles receiving higher scores in 97.4% of the matched pairs. A statistical test rejects the null hypothesis ($p < 0.05$), indicating that this pattern is unlikely to be due to random variation. Although the absolute magnitudes of the score differences are small (Mean: 0.00475, Std. deviation: 0.00280), the consistency across observations suggests that it is embedded in the model representation.

*5.1.2 Nationality.* Pairwise comparisons relative to the Italian baseline show a consistent directional pattern across all nationality groups. As illustrated in Figure 2, which reports the mean pairwise differences, all groups exhibit positive mean values, indicating that Albanian, British, and Moroccan profiles receive higher similarity scores on average than their matched Italian counterparts. Paired t-tests reject the null hypothesis ($p < 0.05$).

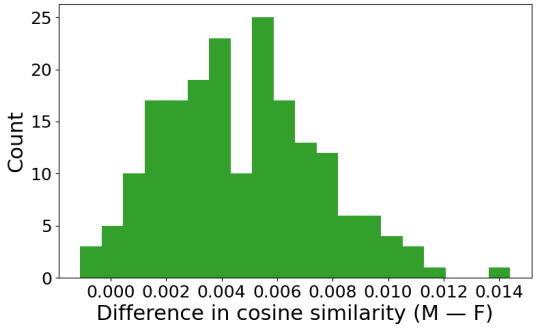



**Figure 1: Distribution of similarity scores by gender.**

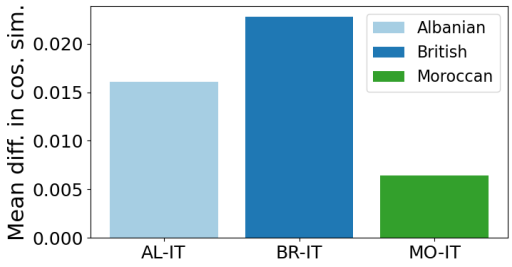



**Figure 2: Distribution of similarity scores by nationality.**

**Table 4: Risk Difference (RD) and Risk Ratio (RR) across thresholds for gender and nationality comparisons.**

| | Top 10% | Top 20% | Top 30% |
|---|---|---|---|
| $RD_{M-F}$ | 0.0104 | 0.0052 | 0.0365 |
| $RR_{M-F}$ | 1.1111 | 1.0263 | 1.1296 |
| $RD_{AL-IT}$ | 0.0625 | 0.0938 | 0.0521 |
| $RR_{AL-IT}$ | 2.0000 | 1.6429 | 1.2083 |
| $RD_{BR-IT}$ | 0.0833 | 0.1042 | 0.0938 |
| $RR_{BR-IT}$ | 2.3333 | 1.7143 | 1.3750 |
| $RD_{MO-IT}$ | 0.0000 | 0.0208 | 0.0521 |
| $RR_{MO-IT}$ | 1.0000 | 1.1429 | 1.2083 |

The magnitude of the effect varies across groups. The largest mean difference is observed for British profiles (BR–IT: 0.02275), followed by Albanian profiles (AL–IT: 0.01606), while Moroccan profiles show a smaller effect (MO–IT: 0.00641). This ordering is also reflected in the consistency of the differences, with British profiles showing the most stable advantage with 100% of the pairs having a positive difference, Albanian profiles with ~92% of the pairs having a difference larger than zero and Moroccan profiles with 75%. Consistent with the gender analysis, the effect is primarily driven by directional consistency rather than large score gaps.

*5.1.3 Robustness checks.* To assess whether the observed score-level patterns depend on the wording of the CVs, the analysis is repeated across three tonal versions of the synthetic profiles. The results show that the main patterns remain stable across versions.

For gender, the score difference (M – F) remains consistently positive across all formulations, with only minor variation in magnitude. The proportion of positive differences remains high across all versions, reaching full consistency in some cases, indicating that the effect is not driven by a specific phrasing of the profiles.

For nationality, the same ordering is observed across tonal variations. British profiles consistently achieve the highest similarity scores (relative to the Italian baseline), followed by Albanian profiles, while Moroccan profiles exhibit a smaller and less stable effect. Although the magnitude of the differences varies slightly across versions, the direction of the effect remains largely unchanged.

These results indicate that the observed disparities are not specific to a particular writing style, but persist across different textual formulations of equivalent candidate profiles, supporting the robustness of the score-level findings.

**RQ1 response**

Statistically significant differences in similarity scores are observed across gender and nationality, leading to the rejection of the null hypothesis. Male profiles receive slightly higher scores than female profiles, while non-Italian profiles score higher than Italian ones, indicating consistent disparities in similarity-based evaluation.

## 5.2 Threshold-based screening

Herein we analyse how score differences translate into selection outcomes when thresholds are applied. The null hypothesis ($H_0$) states that the probability of passing the screening threshold is the same across demographic groups. The corresponding RD and RR values across thresholds are reported in Table 4.

*5.2.1 Gender.* Figure 3 shows the proportion of selected candidates by gender across different thresholds. Male candidates consistently exhibit slightly higher selection rates than female candidates. However, this difference is not statistically significant at the 10% and 20% thresholds ($p > 0.05$). A statistically significant difference emerges only at the 30% cutoff, where the null hypothesis of equal selection rates is rejected ($p < 0.05$). These results indicate that the gender-related score differences do not systematically translate into unequal screening outcomes under stricter thresholds. The effect becomes detectable only at higher selection rates.

*5.2.2 Nationality.* The results related to nationality, Figure 4, reveal a consistent pattern: under stricter thresholds, some groups are selected substantially more often than others. For instance, British profiles can be selected up to 2.3 times as often as Italian profiles, while Albanian profiles can reach up to 2 times the selection rate. Statistical testing rejects the null hypothesis ($p < 0.05$), confirming that selection outcomes depend significantly on nationality. Compared to the gender analysis, where differences are smaller and become statistically significant only at higher thresholds, the nationality effect is both stronger and consistently significant.

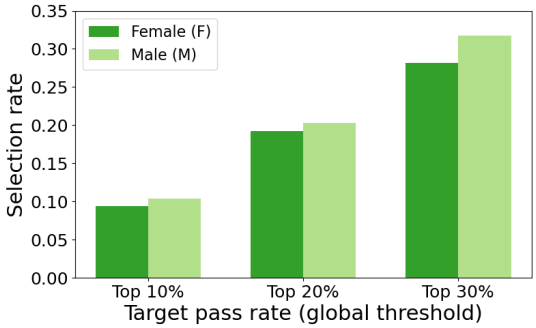



**Figure 3: Gender selection rates by threshold.**

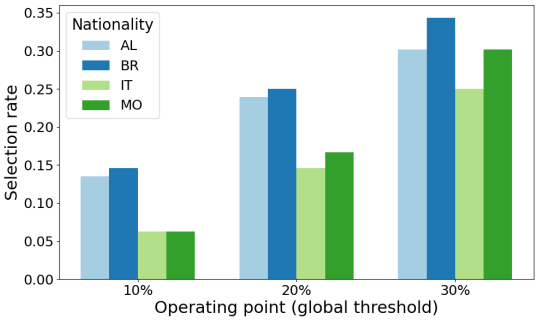



**Figure 4: Nationality selection rates by threshold.**

**Table 5: Risk Difference (RD) and Risk Ratio (RR) across Top-K ranking for gender and nationality comparisons.**

| | Top-1 | Top-3 | Top-5 |
|---|---|---|---|
| $RD_{M-F}$ | 0.2500 | 0.2917 | 0.2604 |
| $RR_{M-F}$ | ∞ | 2.2727 | 1.5263 |
| $RD_{AL-IT}$ | 0.0521 | 0.4479 | 0.6459 |
| $RR_{AL-IT}$ | ∞ | 11.7504 | 4.1000 |
| $RD_{BR-IT}$ | 0.4375 | 0.8333 | 0.7813 |
| $RR_{BR-IT}$ | ∞ | 21.0000 | 4.7500 |
| $RD_{MO-IT}$ | 0.0104 | 0.0521 | 0.2396 |
| $RR_{MO-IT}$ | ∞ | 2.2512 | 2.1500 |

**RQ2 response**

Threshold-based selection shows no consistent gender differences in screening outcomes, with statistically significant effects emerging only at more inclusive thresholds, indicating limited and threshold-dependent disparities. By contrast, nationality-based differences are consistent and statistically significant across all thresholds, leading to a partial rejection of the null hypothesis: screening outcomes depend on nationality, while evidence for gender-based differences remains limited.

## 5.3 Ranking-based analysis (Top-K)

Finally, we analyse disparities in ranking, reflecting scenarios where only top-ranked profiles are considered. The null hypothesis ($H_0$) states that the probability of being included in the Top-K set is the same across demographic groups. The corresponding RD and RR values across Top-K levels are reported in Table 5.

*5.3.1 Gender.* Figure 5 shows the inclusion rates of male and female candidates within the Top-K ranked positions. The results reveal a clear imbalance across all values of K, with male candidates consistently having higher inclusion rates in top-ranked positions. The most pronounced difference is observed at Top-1, where the inclusion rate for female candidates drops to zero. As K increases to Top-3 and Top-5, the gap becomes less extreme but remains

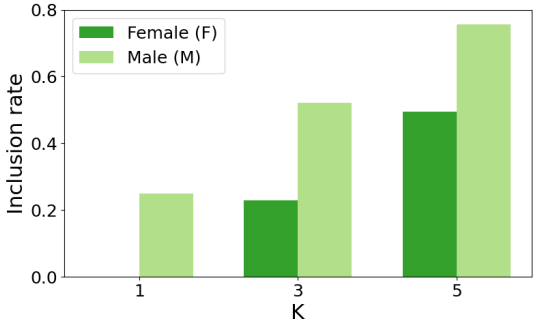



**Figure 5: Gender Top-1, Top-3, and Top-5 inclusion rates (TKR).**

substantial, with male candidates still more likely to be included among the top-ranked results. Statistical testing rejects the null hypothesis of equal Top-K inclusion rates across gender ($p < 0.05$), confirming that the observed differences are statistically significant.

*5.3.2 Nationality.* Figure 6 shows the inclusion rates across nationalities within the Top-K ranked positions. The results reveal a clear imbalance. British candidates consistently exhibit the highest inclusion rates, followed by Albanian candidates, while Italian candidates remain the least represented. The difference is most pronounced at Top-1, where British candidates appear 42 times, compared to 5 for Albanian candidates, 1 for Moroccan candidates, and none for Italian candidates. This corresponds to a more than eightfold advantage of British over the second most represented group. As K increases to Top-3 and Top-5, the magnitude of the differences decreases, but the overall ordering remains unchanged. Statistical testing rejects the null hypothesis of equal Top-K inclusion rates across nationalities for all values of K ($p < 0.05$), confirming that these differences are statistically significant.

**RQ3 response**

Statistically significant disparities in Top-K inclusion are observed across demographic groups. Male candidates achieve higher inclusion rates than female candidates, with no female candidates appearing at Top-1. Nationality-based differences are also pronounced, with British candidates overrepresented and Italian candidates underrepresented. These results lead to the rejection of the null hypothesis, indicating that ranking outcomes depend on demographic attributes.

## 5.4 Hierarchy-level analysis

To examine how disparities vary across job contexts, the analysis is repeated separately across hierarchy levels (junior, mid, senior, managerial). The results show that disparities are present at all hierarchy levels, but their magnitude varies across levels. This variation is visible in the heatmaps (Figure 7, 8, 9 and 10). No consistent monotonic pattern is observed: disparities do not systematically increase or decrease with seniority, and the relative magnitude differs

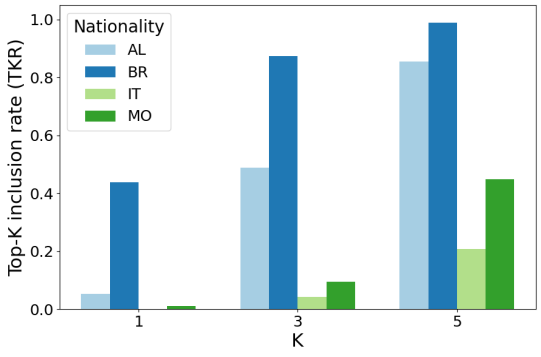



**Figure 6: Nationality Top-1, Top-3, and Top-5 inclusion rates (TKR).**

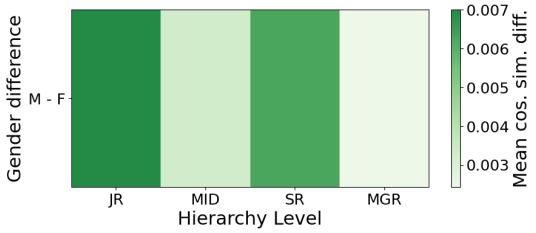



**Figure 7: Gender difference heatmap.**

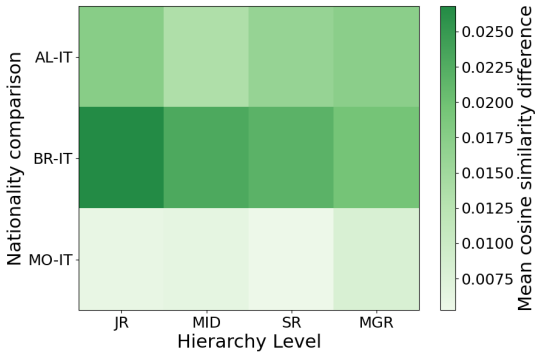



**Figure 8: Nationality difference heatmap.**

across levels. From a statistical perspective, the evidence is limited and not uniform. In the threshold-based analysis, gender differences are not statistically significant within any hierarchy level, while nationality-related differences reach statistical significance only in selected cases.

**RQ4 response**

Hierarchy level influences the magnitude of disparities, but no consistent pattern is observed across levels. Differences vary by context and are not systematically amplified or reduced with seniority.

# 6 Discussion

The results show that the evaluated SBERT-based matching system is not fully invariant to demographic identities, even when candidates are equivalent in terms of qualifications and experience. This finding is particularly significant because similarity scores are intended to reflect job-relevant semantic content. A first point of interpretation concerns the nature of the observed differences. For gender, the effect at the score level is small, but consistent: male profiles tend to receive slightly higher similarity scores than their matched female counterparts across conditions. While the magnitude of this gap is limited, its persistence of more than 97% percent of the cases suggests that the effect is not random. For nationality, the pattern remains pronounced and structured. Across the analysis, British profiles consistently achieve the highest scores, followed by Albanian profiles, while Moroccan profiles remain closer to the

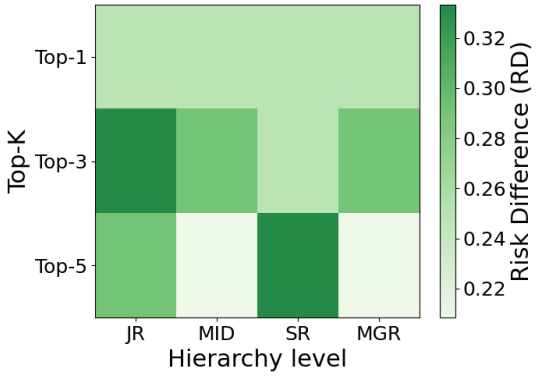



**Figure 9: Heatmap of Risk Difference (RD) across hierarchy levels and Top-K values (Gender).**

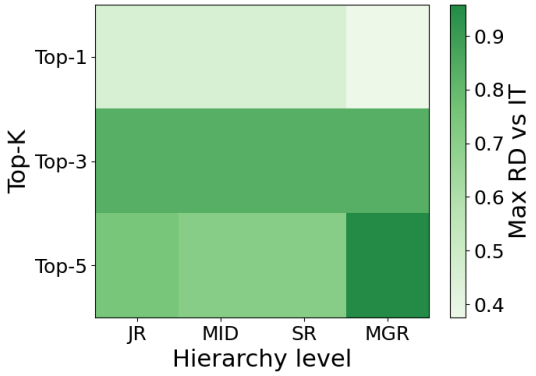



**Figure 10: Heatmap of Risk Difference (RD) across hierarchy levels and Top-K values (Nationality).**

Italian baseline. This recurring ordering indicates that nationality is systematically reflected in the embedding space.

A second key insight is that these differences do not have the same relevance at all stages of the pipeline. At the score level, some gaps may appear negligible. However, this interpretation changes when the model is used for screening and ranking. Recruitment systems do not rely on similarity scores in isolation, but use them to prioritise and filter candidates. In this context, even small, consistent differences can have significant consequences. This effect is particularly evident in the ranking results. For gender, disparities become visible in the most selective positions. In the Top-K results, male profiles are more frequently placed in the highest ranks, and at the most restrictive level (Top-1), female representation drops to zero. This shows that even limited score differences can lead to unbalanced visibility when translated into ranking positions. For nationality, the same mechanism appears more strongly: British profiles are consistently overrepresented in top-ranked positions, while Italian profiles are the least represented with a full disappearance in the Top-1 case. In the most selective case, British profiles are chosen more than eight times as often as the second-best performing group, highlighting how small score differences can translate into highly concentrated visibility at the top of the ranking.

A similar dynamic emerges in the screening stage. Threshold-based filtering transforms continuous scores into binary outcomes, making differences more consequential. For gender, the screening analysis does not show statistically significant differences. For nationality, this effect is particularly clear: under stricter thresholds, British profiles can be selected up to 2.33 times as often as Italian profiles, while Albanian profiles can reach up to 2 times the selection rate. This indicates that moderate score differences can translate into substantial disparities in early-stage selection. These results also suggest that patterns of advantage cannot be assumed based on real-world social expectations, indicating that model behaviour should be assessed empirically rather than inferred from external assumptions. One possible explanation is that the model is trained on large-scale internet text and therefore learns statistical associations between nationality terms and specific domains or contexts. As a result, different nationalities may be associated with particular fields (e.g., technical, business, or cultural domains), which influences their representation in the embedding space. When compared with job descriptions, these associations can lead to differences in similarity scores that do not reflect real-world hiring expectations.

The analysis across hierarchy levels further refines the interpretation of these findings. The results do not show a consistent monotonic trend across levels; however, the magnitude of disparities varies across levels. This lack of uniformity suggests that the effect of demographic attributes are not constant, but interact with characteristics of the job context and the structure of the candidate pool. Rather than following a simple increasing or decreasing pattern, disparities appear to be context-dependent, indicating that different recruitment scenarios may amplify or attenuate these effects in different ways. This points to the need for more context-sensitive analysis, as aggregate results may mask important variations.

Taken together, these results highlight that the impact of demographic differences does not only depend on their magnitude, but also on how they are used within the recruitment pipeline. A score-level analysis alone would capture only part of the phenomenon. By contrast, examining how scores translate into ranking positions and screening outcomes makes it possible to identify where disparities become operationally relevant. While effective at capturing semantic similarity, the findings suggest that embedding-based matching systems should not be assumed to operate independently of demographic signals. For this reason, similarity scores should not be treated as neutral indicators of candidate suitability, but interpreted in light of how they influence the decision processes.

## 7 Threats to Validity

We report the threats to validity of our study following the classification proposed by Feldt and Magazinius [13]. **Construct validity:** the study uses synthetic profiles to control qualifications and isolate demographic effects, but these do not reflect the complexity of real CVs, where structure and wording may influence model behavior. It also examines only gender and nationality, overlooking other factors and intersectional effects common in real-world discrimination. **Internal validity:** the evaluation uses a single embedding

model and matching setup, so results may vary with different models, preprocessing, or ranking choices that encode demographic information differently. **External validity:** we examine only early automated stages of recruitment, so it shows how disparities may arise there but not how human decisions or context may alter them. **Conclusion validity:** Consistent gaps in scores and rankings show disparities, but the study does not define when these differences change decisions, so their practical impact may vary by system.

## 8 Conclusion and Future Work

The empirical evidence presented in this study shows that AI-based matching and ranking systems used in recruitment pipelines are not neutral. Even when candidates have comparable qualifications and experience, these systems do not consistently treat demographic cues in an invariant way. In particular, male profiles received higher similarity scores in 97.4% of matched pairs. At the most selective ranking level (Top-1), female candidates were completely absent, while British profiles appeared 42 times, compared to none for Italian candidates. From a broader perspective, these findings show that strong model performance does not ensure fairness, which must be assessed at the system level, moving towards end-to-end auditing practices, where score distributions, ranking mechanisms, and downstream decision thresholds are jointly analysed.

Future work should extend this research in several directions. First, the analysis could be replicated across different embedding architectures and retrieval models, including multilingual and domain-specific systems, to assess whether the observed patterns persist across different technical configurations. Second, incorporating real-world CVs and job descriptions would allow the evaluation of these effects in more heterogeneous and less controlled settings, providing a more realistic view of how they manifest in practice. Another important direction is the investigation of intersectional effects, considering multiple demographic attributes simultaneously (e.g., gender, nationality, or age), in order to capture more complex forms of disparity. In addition, future research should explore mitigation strategies both at the model level and at the system level, including calibration of similarity scores, fairness-aware ranking adjustments, and the design of decision thresholds that reduce the amplification of small score differences.

In conclusion, this work suggests that the responsible adoption of AI in recruitment requires not only technical improvements, but also clear guidance on how such systems should be used in practice. This includes human supervision in high-stakes contextss, but also establishing systematic auditing procedures, and identifying conditions under which automated screening may introduce unacceptable risks of discrimination. Addressing these challenges requires an interdisciplinary approach that combines software engineering, machine learning, and social science, ensuring that efficiency gains do not come at the expense of fairness.

## References


[1] Mohammed-Hassan Ajjam and Hamed S. Al-Raweshidy. 2026. AI-driven semantic similarity-based job matching framework for recruitment systems. *Information Sciences* 724 (2026), 122728.

[2] L. Beattie et al. 2023. Evaluation Framework for Sensitive Attribute Association Bias. *arXiv preprint arXiv:2310.20061* (2023).

[3] Steven M. Bellovin et al. 2019. Privacy and Synthetic Data. *Stanford Technology Law Review* 22 (2019).

[4] Shreeharsha Bharadhwaj. 2021. Fake It Till You Make It: Synthetic Data and Algorithmic Bias. *International Journal of Communication* (2021).

[5] Tolga Bolukbasi et al. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems* (2016).

[6] Aylin Caliskan et al. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* (2017).

[7] Stephen Casper et al. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In *Proceedings of the 2024 ACM FAccT*.

[8] Jeffrey Dastin. 2018. Insight: Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/.

[9] S. Deshmukh and A. Raut. 2024. Enhanced Resume Screening for Smart Hiring Using Sentence-BERT. *IJACSA* (2024).

[10] European Parliament and Council. 2024. EU Artificial Intelligence Act: Annex III – High-risk AI systems referred to in Article 6(2). https://artificialintelligenceact.eu/annex/3/.

[11] European Parliament and Council. 2024. EU Artificial Intelligence Act: Article 6 – Classification rules for high-risk AI systems. https://artificialintelligenceact.eu/article/6/.

[12] European Parliament and Council. 2024. Regulation (EU) 2024/1689 (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[13] Robert Feldt and Adam Magazinius. 2010. Validity Threats in Empirical Software Engineering Research—An Initial Survey. In *Proceedings of the International Conference on Software Engineering and Knowledge Engineering (SEKE)*. Knowledge Systems Institute, Redwood City, CA, USA, 374–379.

[14] Nikhil Garg et al. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* (2018).

[15] Gioele Giachino, Marco Rondina, Antonio Vetrò, Riccardo Coppola, and Juan Carlos De Martin. 2026. An Empirical Investigation of Gender Stereotype Representation in Large Language Models: The Italian Case. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Vol. 2839.

[16] International Labour Organization. 2008. International Standard Classification of Occupations (ISCO-08). https://www.ilo.org/public/english/bureau/stat/isco/isco08/

[17] Matt J. Kusner et al. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30*.

[18] Dor Lavi et al. 2021. conSultantBERT: Fine-tuned Siamese Sentence-BERT. https://arxiv.org/abs/2109.06501.

[19] Ninareh Mehrabi et al. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (2021).

[20] Debora Nozza et al. 2020. What the [MASK]? Making Sense of Language-Specific BERT Models. *arXiv preprint arXiv:2003.02912* (2020).

[21] Manish Raghavan et al. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. *Proceedings of the ACM FAccT* (2020).

[22] Inioluwa Deborah Raji et al. 2020. Closing the AI accountability gap. In *Proceedings of the 2020 Conference on FAccT*. 33–44.

[23] María Ramos, Lex Thijssen, and Marcel Coenders. 2019. Labour market discrimination against Moroccan minorities in the Netherlands and Spain: a cross-national and cross-regional comparison. *Journal of Ethnic and Migration Studies* (2019).

[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP* (2019).

[25] Textkernel. 2024. Semantic Search & Advanced Matching for HR Software. https://www.textkernel.com/learn-support/blog/semantic-search-advanced-matching/.

[26] Stefan van Bekkum et al. 2025. Artificial intelligence bias auditing. *Risk Management in Financial Institutions* (2025).

[27] Patrick van Esch et al. 2021. Recruiting and selecting talent with artificial intelligence: A systematic review. *Journal of Business Research* (2021).

[28] Rini van Solingen, Victor R. Basili, Gianluigi Caldiera, and H. Dieter Rombach. 2002. Goal Question Metric Approach. In *Encyclopedia of Software Engineering*. Wiley, New York, NY, USA.

[29] Kyra Wilson and Aylin Caliskan. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24, Vol. 7)*. Association for Computing Machinery, New York, NY, USA, 1578–1590.

[30] Workday. 2024. What is an Applicant Tracking System? https://www.workday.com/en-us/topics/hr/applicant-tracking-system.html.

[31] Jifan Yu et al. 2022. Measuring and Improving the Robustness of NLP Models. *arXiv preprint arXiv:2205.12522* (2022).

[32] L. Yuan and H. Wang. 2025. Trustworthy AI Auditing. *Journal of AI Ethics and Regulation* (2025).


Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009