

Section-Level Bias Analysis and Fairness-Aware Pipeline in AI Hiring

Priansu Koirala

The British College
(Leeds Beckett University),
Leeds, United Kingdom
kpriansu25@tbc.edu.np

Roshan Chitrakar

Nepal College of Information Technology
(Pokhara University)
Lalitpur, Nepal
roshanchi@ncit.edu.np

Abstract—Automated resume screening systems are typically audited at the level of the whole document, which reveals that disparities exist but not where in a candidate's resume they originate. This paper reports a controlled experiment that decomposes the resume into four functional sections viz. skills, education, experience and certifications and then measures how each section contributes to both relevance prediction and demographic disparity. Using 1,730 synthetic job–resume pairs have been generated with the FINDHR CV generator across five occupational roles, each section and each job description has been embedded into a sentence-transformer model and reduced to a cosine similarity score, yielding five interpretable features per pair. A logistic regression relevance scorer over these features reaches a test accuracy of 0.705 and an F1 of 0.777. An ablation study establishes that the four section similarities alone outperform full-document similarity (test accuracy 0.711 versus 0.688), showing that the decomposition retains the relevance signal rather than merely supplementing it. A fairness audit of top-10 shortlists reveals substantial disparity in the cosine-similarity baseline: a demographic parity difference (DPD) of 0.151 for gender and 0.442 for race. Diagnostic probing shows a sharp asymmetry between the two attributes. Gender is recoverable from full-resume embeddings with an AUC of 0.998, and section-level attribution localizes that signal in the education and skills sections. Race, by contrast, is not linearly recoverable: the probe scores 0.751 accuracy, exactly the majority-class floor on the test partition, and the White versus non-White AUC of 0.461 is no better than chance. Decomposing the pipeline shows that the supervised section-aware scorer accounts for most of the disparity reduction, lowering gender DPD from 0.151 to 0.072 and race DPD from 0.442 to 0.189; a subsequent group-aware score adjustment reduces gender DPD further to 0.048 but left race DPD unchanged. Notably, the fitted relevance scorer independently down-weights the sections that the probes have flagged as demographically informative. These results support section-level decomposition as a practical diagnostic layer for hiring pipelines, and show that substantial selection disparity can coexist with a protected attribute that is not linearly encoded in the underlying representation.

Keywords: *algorithmic fairness; resume screening; sentence embeddings; demographic parity; fair ranking; explainable AI; synthetic data*

I. INTRODUCTION

Organisations increasingly delegate the first pass of candidate screening to automated systems. Applicant tracking software, and more recently pipelines built on transformer language models, now mediate the point at which a large applicant pool is reduced to a shortlist a human will actually read [1], [2]. The appeal is straightforward: manual screening

at scale is slow, expensive and inconsistent, and an automated ranker applies the same function to every applicant. That consistency, however, is not the same thing as fairness. A model trained on or matched against historical patterns can reproduce those patterns systematically rather than idiosyncratically, and at a volume no individual recruiter could match [3], [4].

Empirical work has established that this concern is well founded. [5] showed that massive text embedding models retrieving candidates against job descriptions favour resumes associated with White and male identities across nine occupations, and that the magnitude of the effect varies with resume length and structure. [6] report comparable gender, race and education effects in large language model matching. What these studies share is a unit of analysis: the resume is treated as a single document, and bias is measured at the level of the final decision. That framing answers whether a system is biased. It does not answer where the bias enters.

The distinction matters operationally. A resume is not undifferentiated text. It is a structured document whose sections carry different kinds of information and, plausibly, different demographic signals. An education section encodes institution names that correlate with geography, socioeconomic background and migration history. An experience section encodes employer names, career continuity and gaps that may correlate with caregiving or visa status. A skills section is closer to a direct statement of capability. If these sections differ in how much protected-attribute signal they carry, then a mitigation strategy applied uniformly across the whole document is applying a blunt instrument, and one that the debiasing literature suggests may suppress the surface form of a signal while leaving its structure intact [7].

This paper takes the section as the unit of analysis. The contribution is threefold. First, a screening pipeline is constructed in which resume sections are embedded and scored separately, producing a relevance model whose coefficients state directly how much each section contributed to a shortlisting decision. Second, that pipeline is audited with embedding probes and section-level attribution, asking not only whether demographic information is recoverable but from which part of the document. Third, a two-part mitigation strategy group-aware score adjustment followed by soft-penalty re-ranking is applied and evaluated, and the fairness improvement is decomposed to determine which component is responsible for it.

Two findings are worth flagging in advance because they shape how the rest of the paper should be read. The first is an asymmetry: gender is strongly and locally encoded in these

representations, whereas race is not linearly recoverable at all under the class distribution used here, despite race showing the larger selection disparity. Disparity in outcomes and recoverability of the attribute are separable properties, and conflating them would lead an auditor to the wrong conclusion. The second concerns where the fairness gain actually originates: decomposing the pipeline attributes most of the disparity reduction to the change in scoring model rather than to either mitigation stage. The soft re-ranking stage, which the fair-ranking literature would lead one to expect to do substantial work [8], left both parity metrics unchanged, and the group-aware score adjustment produced no change in race parity. Both results are reported as they emerged.

II. RELATED WORK

Early automated screening relied on exact term matching between resumes and job descriptions. Such systems were computationally cheap but semantically brittle, rejecting qualified candidates whose vocabulary diverged from the parser's expectations [4]. Statistical representations such as TF-IDF improved robustness without resolving the underlying problem: two descriptions of the same competence written differently remained distant in feature space.

Dense sentence representations changed this. Sentence-BERT and its distilled successors produce fixed-length vectors in which semantically related expressions are close, allowing resume-job matching to be framed as retrieval with cosine similarity as the scoring function [9]. This is now the dominant architecture in the applied literature [10], [11]. A parallel line of work argues for hybrid scoring, combining embeddings with structured attributes or treating sections as separate semantic signals before pooling them into a composite score [6]. The pipeline described here sits in this second tradition, but uses the decomposition diagnostically rather than solely to improve ranking accuracy.

Evidence of demographic bias in automated screening is consistent across model families and experimental designs. [12] established that distributional semantics absorb human-like associations from training corpora, providing the mechanism. [5] demonstrated the downstream consequence in a retrieval setting designed to mirror screening. [13] surveyed commercial vendors and found that fairness claims frequently outrun the evidence supporting them. [14] situate the problem in its legal and organisational context, arguing that technical fixes decoupled from governance are unlikely to succeed.

A recurring limitation across this body of work is the treatment of the resume as an atomic unit. Studies establish that disparity exists and characterise its magnitude, but the experimental design cannot attribute it to any component of the input. [15] move toward finer granularity in the context of large language models, but section-level attribution in embedding-based pipelines remains largely unexamined.

Group fairness criteria dominate the hiring literature. Demographic parity requires equal selection rates across protected groups; equalised odds additionally constrains error rates [16]. They argue that for allocative decisions such as shortlisting, where the harm is the denial of an opportunity, demographic parity is the appropriate framing. [17] counter that parity is sensitive to base rates and can mislead when groups differ substantially in size or qualification distribution, a caveat that applies directly to the data used here and one revisited in Section V-C.

Mitigation techniques divide into pre-processing, in-processing and post-processing families [18]. Pre-processing

removes or masks sensitive information, but proxy signals survive masking [19]. In-processing embeds fairness constraints in the objective at the cost of retraining and additional hyperparameters. Post-processing modifies outputs of a fixed model and is attractive where retraining is impractical; the canonical hiring application is the re-ranking framework [8] deployed in LinkedIn Talent Search. [20] survey fairness in ranking specifically, noting that most mitigation research still targets binary classification rather than the ranked shortlists that screening actually produces.

Probing classifiers provide the diagnostic counterpart. A low-capacity model trained to predict a protected attribute from a representation tests whether that attribute is linearly encoded; keeping the probe simple ensures the result reflects the information in the representation rather than the probe's expressive power [21]. This convention is adopted throughout.

Three gaps motivate this study. First, no existing work systematically compares how much protected-attribute signal different resume sections carry, so mitigation cannot be targeted. Second, fairness auditing and utility evaluation are frequently reported separately, obscuring the trade-off. Third, the relative contribution of score adjustment versus re-ranking within a combined pipeline is rarely decomposed, so it is unclear which intervention is doing the work. This paper addresses all three within a single controlled experiment.

III. METHODOLOGY

The experiment follows a three-stage design: representation, diagnosis, mitigation. Figure 1 gives the overall architecture. Each stage emits structured outputs similarity scores, probe metrics, fairness metrics, ranked shortlists that feed the next, so that any reported difference can be attributed to a specific intervention.

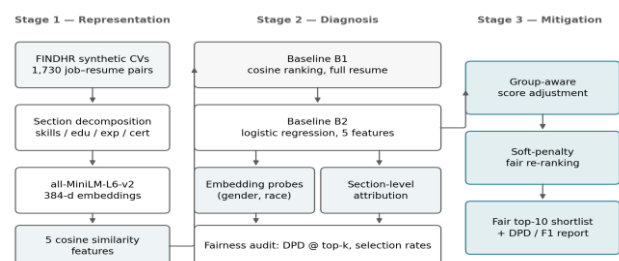


Fig. 1. Three-stage experimental pipeline. Stage 1 converts structured CVs into five interpretable similarity features. Stage 2 establishes two baselines and audits them. Stage 3 applies group-aware score adjustment followed by soft-penalty re-ranking.

A. Data

Access to real recruitment data for fairness research is constrained by data protection law and by the commercial and reputational exposure organisations face if historical discrimination is surfaced [14]. This study therefore uses the FINDHR Synthetic CV Generator [22], developed within the EU-funded FINDHR project and released specifically as a benchmark for algorithmic hiring discrimination research. The generator samples from distributions derived from a GDPR-compliant data donation campaign and emits structured JSON CV objects with explicit educational episodes, professional history, skills and language competences, along with sector, age band, gender and nationality attributes.

Each JSON object was flattened into a row representing one job-resume pair. Section fields were built by concatenating the corresponding sub-components into plain

text (education_text, experience_text, skills_text, certifications_text), and resume_text was formed by concatenating all four. This preserves section boundaries, which the rest of the analysis depends on. Race is recorded as a categorical attribute emitted directly by the FINDHR generator, with values White, Asian, Black, Hispanic and Other. It is not derived from nationality, country of origin or any regional grouping. Protected attributes were used exclusively for fairness evaluation and were never supplied as features to any predictive model.

The resulting dataset comprises 1,730 job–resume pairs across five occupational roles, each carrying a binary relevance label indicating whether the candidate is a suitable match. Binary relevance labels were assigned by comparing each candidate's skills section against the requirements stated in the job description, with pairs exceeding a match criterion labelled relevant. The dataset was constructed by the author for this study using the FINDHR synthetic CV generator. Text normalisation was deliberately minimal lowercasing and whitespace collapse only. Stopword removal, stemming and punctuation stripping were avoided because the embedding model encodes contextual meaning rather than token frequency, and such operations discard information the encoder uses.

B. Section decomposition and feature construction

Each job description and each resume section was encoded independently into a 384-dimensional vector using sentence-transformers/all-MiniLM-L6-v2 [9]. This model was selected for three reasons: it is substantially lighter than BERT-large while remaining competitive on semantic textual similarity benchmarks; it tolerates the informal and heterogeneous register of resume text without task-specific fine-tuning; and holding the encoder fixed across all conditions ensures that observed differences are attributable to the experimental variables rather than to representation drift.

For every pair, the cosine similarity between the job embedding and each of five resume embeddings was computed, giving the feature vector [full, skills, education, experience, certifications]. Cosine similarity was chosen over Euclidean distance because it is invariant to vector magnitude, which matters when comparing a terse skills list against a long experience narrative within the same feature vector.

C. Two baselines

Two baselines are distinguished explicitly, because they answer different questions and conflating them obscures where each observed effect originates.

B1: cosine ranking. Candidates are ranked per job by full-resume similarity alone, with no supervision. This represents the naive embedding-based screener common in deployed systems and supplies the reference point for the fairness audit.

B2: section-aware relevance scorer. A logistic regression classifier is trained on the five similarity features to predict the relevance label. This measures what section decomposition buys in ranking utility, before any fairness intervention is applied.

Logistic regression was chosen over a fine-tuned neural scorer deliberately. Its coefficients state directly how much each section contributed to a score, which is the quantity of interest here; a five-dimensional feature space over a dataset of this size offers a neural model little room to add value while

forfeiting interpretability; and the same architecture serves as the probe elsewhere in the study, keeping capacity constant across analyses.

D. Demographic leakage probes

To test whether protected attributes are recoverable from the representations, multinomial logistic regression probes were trained to predict gender and race from full-resume embeddings, and separately from each section embedding. Probe accuracy is interpreted against the majority-class baseline rather than against chance, since a probe that merely learns the dominant class conveys no information about encoding. Binary ROC-AUC was computed for Female versus non-Female and White versus non-White splits as a threshold-independent complement.

E. Section-level attribution

Probing over section embeddings establishes which sections carry signal but not the direction of the relationship. A second set of logistic regressions was therefore fitted, predicting each protected attribute from the four section similarity features jointly, with signed coefficients and odds ratios reported. This identifies which sections push a candidate toward or away from a demographic group conditional on the others.

F. Group-aware score adjustment

The trained relevance scorer emits a calibrated probability for each pair. Let s_i denote the predicted relevance probability for candidate i , $g(i)$ and $r(i)$ the candidate's gender and race group, and μ the mean predicted score across the candidate pool. Writing μ_g and μ_r for the mean predicted score within gender group g and race group r , the adjusted score is

$$s'_i = \text{clip}(s_i + \lambda_G(\mu - \mu_{g(i)}) + \lambda_R(\mu - \mu_{r(i)}), 0, 1) \quad (1)$$

with $\lambda_G = 0.10$ and $\lambda_R = 0.05$. These coefficients were set by inspection rather than tuned, and no sensitivity analysis over λ was performed.

Both attributes are adjusted additively on the same score. The sign of each term raises scores for groups scoring below the pool mean and lowers them for groups above it, in proportion to the gap. Clipping keeps the result in the valid probability range. These group means are computed over the candidate pool being ranked, using predicted scores only; no relevance labels enter the adjustment. This reflects deployment conditions, in which the full applicant pool for a given vacancy is available at scoring time. The adjustment is intentionally partial: it nudges group means toward parity rather than equalising them, which preserves within-group ordering by relevance. This is a lightweight heuristic analogue of the post-processing operators in toolkits such as Fairlearn, chosen because its effect on any individual score can be stated in one sentence to a recruiter or an auditor.

G. Soft-penalty fair re-ranking

The re-ranker operates on adjusted scores per job and modifies only the top- k prefix. Building the shortlist greedily, it compares each group's representation in the prefix so far against its representation in the candidate pool, and subtracts a small penalty from candidates whose group is already over-represented. The highest penalised score is selected, group counts are updated, and the procedure repeats. At each selection step t , let $\hat{p}_G(t)$ be the proportion of already-selected candidates belonging to the current candidate's gender

group, and p_G that group's share of the full candidate pool for the vacancy. The penalised score is

$$\tilde{s}_i(t) = s'_i - \alpha_G \max(0, \hat{p}_G(t) - p_G) - \alpha_R \max(0, \hat{p}_R(t) - p_R) \quad (2)$$

with $\alpha_G = 0.05$ and $\alpha_R = 0.10$.

The penalty applies only when a group is over-represented relative to its pool share, and scales with the size of that excess.

Remaining candidates are appended in descending adjusted score order. Soft penalties were preferred over hard quotas or constrained optimisation because quota-based methods require assumptions about acceptable group proportions that are difficult to justify outside a specific deployment context, and because each greedy step remains individually explainable.

H. Evaluation

Utility is reported as accuracy and F1 on the relevance label, with F1 given primacy because the classes are imbalanced. Ranking quality is reported as precision@10 and recall@10 per job. Fairness is reported as demographic parity difference at top-10, the spread between the highest and lowest group selection rates, computed with the Fairlearn MetricFrame construct, separately for gender and race. A shortlist threshold of $k = 10$ reflects a realistic recruiter workload per vacancy and matches the convention in the fair-ranking literature [8].

Accuracy is the proportion of correctly classified pairs. F1 is the harmonic mean of precision and recall on the positive (relevant) class. Precision@10 is the proportion of the top ten ranked candidates for a vacancy that are relevant; recall@10 is the proportion of all relevant candidates for that vacancy appearing in the top ten. Demographic parity difference is defined as $DPD = \max_a P(\hat{y} = 1 | A = a) - \min_a P(\hat{y} = 1 | A = a)$, where the selection rate for group a is the proportion of candidates appearing in a top-10 shortlist.

I. Reproducibility

The dataset was split 80/10/10 into training, validation and test partitions by stratified sampling on the relevance label, giving 1,384 / 173 / 173 pairs. All stochastic operations used a fixed seed (42). The relevance scorer uses L2-regularised logistic regression at the scikit-learn default regularisation strength ($C = 1$); no hyperparameter tuning was performed. A post-hoc sensitivity analysis over C is reported in Section 4.2. All experiments ran in Python 3.11 using sentence-transformers, scikit-learn and Fairlearn. The test partition was held out until final evaluation.

IV. RESULTS

A. Dataset composition

TABLE I. COMPOSITION OF THE 1,730 JOB-RESUME PAIRS ACROSS FIVE OCCUPATIONAL ROLES

Attribute	Category	n	%
Relevance label	Relevant (1)	1,050	60.7
	Not relevant (0)	680	39.3
Gender	Female	849	49.1
	Male	805	46.5
	Non-binary	76	4.4

Race	White	1,397	80.8
	Asian	127	7.3
	Black	77	4.5
	Hispanic	74	4.3
	Other	55	3.2

Table 1 summarises the distribution of labels and protected attributes. The relevance label is moderately imbalanced. The gender distribution is close to balanced between Female and Male with a small non-binary group. The race distribution is heavily skewed toward the White category, which constitutes 80.8% of the sample. This skew is consequential for interpreting the race results and is revisited repeatedly.

B. Relevance scoring utility

The section-aware scorer (B2) achieved a validation accuracy of 0.717 and F1 of 0.786, and a test accuracy of 0.705 with F1 of 0.777. The narrow validation-to-test gap indicates the model generalises rather than fitting the validation split. That five cosine similarities support this level of performance is the first substantive result: section-level alignment captures the relevance signal in full, as the ablation in Table 3 confirms, and does so in a form whose contribution can be read off directly.

TABLE II. SENSITIVITY OF THE RELEVANCE SCORER TO REGULARISATION STRENGTH

Configuration	C	Val acc	Val F1	Test acc	Test F1
C=0.1	0.1	0.665	0.775	0.676	0.788
C=1(used)	1	0.717	0.786	0.705	0.777
C=10	10	0.740	0.796	0.717	0.778

A post-hoc sensitivity analysis over the regularisation strength C is reported in Table 2. The deployed scorer uses the scikit-learn default ($C = 1$); the grid was run after the fact to establish whether the reported results depend on this choice. Test accuracy varies by approximately four percentage points across two orders of magnitude of C , and test F1 by roughly one point. Notably, the lowest setting ($C = 0.1$) yields the highest test F1 but the lowest accuracy, reflecting a drift toward majority-class prediction that inflates F1 without improving discrimination. Accuracy and F1 are therefore reported together throughout.

TABLE III. FEATURE ABLATION FOR THE RELEVANCE SCORER

Feature set	n	Val acc	Val F1	Test acc	Test F1	P@10	R@10
Full-document similarity only	1	0.671	0.771	0.688	0.777	0.84	0.412
Section similarities only	4	0.717	0.788	0.711	0.783	0.88	0.431

All features	5	0.717	0.786	0.705	0.777	0.88	0.431
--------------	---	-------	-------	-------	-------	------	-------

Because the section-aware scorer includes full-document similarity alongside the four section similarities, its performance alone does not establish that section-level features carry the relevance signal. Table 3 isolates their contribution. Using only the four section similarities outperforms using full-document similarity alone on every measure: test accuracy rises from 0.688 to 0.711, test F1 from 0.777 to 0.783, precision@10 from 0.84 to 0.88, and recall@10 from 0.412 to 0.431. Adding full-document similarity to the section features produces no improvement and a slight degradation on test (accuracy 0.705, F1 0.777), with identical ranking metrics. The section decomposition is therefore not merely compatible with full-document matching but sufficient on its own; the full-document feature is redundant given the sections. This also confirms that the interpretability gained from section-level decomposition costs nothing in predictive performance. The five-feature configuration is retained throughout for continuity with the pipeline as originally specified and because the difference falls within the variation expected at this sample size; the ablation is reported as evidence about feature sufficiency rather than as a model-selection result.

C. Ranking quality

Figure 2 reports precision@10 and recall@10 for each of the five jobs. Precision is consistently high, ranging from 0.70 to 0.90, so most candidates surfaced in a shortlist are genuinely relevant. Recall is markedly lower, between 0.33 and 0.53, meaning that roughly 40% to 67% of relevant candidates never reach the shortlist at all.



Fig. 2. Precision@10 and recall@10 by job for the baseline ranker. High precision coexists with substantial recall loss.

This precision–recall asymmetry is a property of top-k screening generally rather than a defect of this model, but it has a fairness implication that a precision-only evaluation would conceal. When a large majority of qualified candidates are excluded by the cut-off, the question of which qualified candidates are excluded becomes consequential, and that is precisely what the parity analysis in the next section examines.

D. Baseline fairness audit

Table IV reports top-10 selection rates under the cosine baseline B1. Gender rates are relatively close for Female and Male candidates, with the non-binary group selected at a substantially higher rate, producing a DPD of 0.151. Race

disparity is far larger: the Hispanic group is selected at 60.0% against 15.8% for the Asian group, a DPD of 0.442.

Two points about interpretation. First, the direction of the disparity does not follow the pattern most often reported in the literature: the numerically smallest race groups are not uniformly disadvantaged, and the majority White group sits near the middle of the range. This is a property of the synthetic generator's sampling rather than a claim about real labour markets. Second, and more importantly, these rates rest on very few individuals. The test partition contains 173 pairs across five jobs, so each shortlist draws from roughly 35 candidates and the full audit covers approximately 50 selections. A rate of 0.600 for the Hispanic group plausibly reflects three selections out of five candidates. The disparity is real in this sample; its precision is low, and no confidence intervals are reported.

TABLE IV. TOP-10 SELECTION RATES AND DEMOGRAPHIC PARITY DIFFERENCE UNDER THE COSINE-SIMILARITY BASELINE (B1)

Attribute	Group	Selection rate	DPD
Gender	Female	0.289	0.151
	Male	0.277	
	Non-binary	0.429	
Race	Asian	0.158	0.442
	Black	0.400	
	Hispanic	0.600	
	Other	0.333	
	White	0.285	

E. Demographic leakage: a sharp asymmetry

Probing full-resume embeddings produced markedly different results for the two attributes. The gender probe reached 0.94 accuracy and 0.92 weighted F1, with precision and recall above 0.94 for both dominant classes; the small non-binary class was not recovered, which reflects its support of seven rather than an absence of signal. The binary Female versus non-Female ROC curve in Figure 3 gives an AUC of 0.998. Gender is, for practical purposes, fully recoverable from these representations by a linear model.

ROC Curve for Gender Prediction (Female vs. Male)

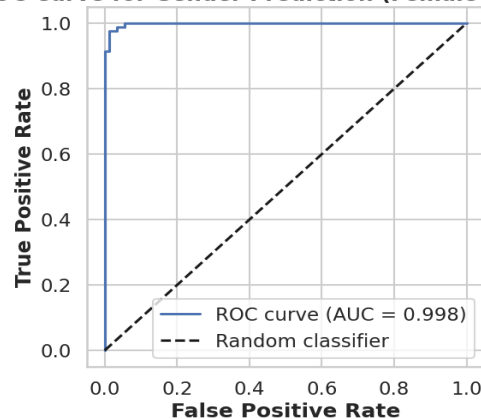


Fig. 3. ROC curve for the binary gender probe on full-resume embeddings (AUC = 0.998), indicating near-complete linear separability.

The race probe tells the opposite story, and reading it correctly requires care. Its raw accuracy of 0.751 appears moderate, but the majority-class floor on the test partition is also 0.751, the White share of 130 of 173 pairs; the probe therefore performs exactly at the floor, predicting the majority class and adding nothing. The per-class report confirms the

diagnosis: performance is concentrated entirely in the majority class, with precision and recall near zero for every minority category. The binary White versus non-White AUC of 0.461 falls below chance. The race probing result is therefore reported as inconclusive: under this class distribution, race is not linearly recoverable from the embeddings, and the data do not permit a determination of whether the signal is absent or merely undetectable at this sample size.

Section-wise probing in Table V reinforces the reading. Gender accuracy varies meaningfully across sections, from 0.457 for experience to 0.543 for education, indicating that no single section is sufficient but that education and skills carry more signal than the rest: gender cues are distributed, and become highly predictive only when sections are combined. Race accuracy, by contrast, is identical to three decimal places across all four sections. A constant of 0.751 regardless of input is the signature of a probe predicting the majority class and ignoring the features entirely.

TABLE V. SECTION-WISE PROBE PERFORMANCE

Section	Gender acc.	Gender F1	Race acc.	Race F1
Skills	0.526	0.509	0.751	0.645
Education	0.543	0.532	0.751	0.645
Experience	0.457	0.441	0.751	0.645
Certifications	0.491	0.457	0.751	0.645

F. Section-level attribution

Figure 4 reports signed coefficients from the joint attribution models. For gender, education similarity is the strongest positive predictor (coefficient +0.68, odds ratio ≈ 1.98), followed by skills (+0.49, OR ≈ 1.64); experience and certifications carry small negative coefficients (OR ≈ 0.88). For race, experience similarity dominates (+0.88, OR ≈ 2.42), with education contributing modestly (+0.24, OR ≈ 1.27) and skills and certifications negative. These coefficients are subject to the labelling dependency noted in Section 5.3.

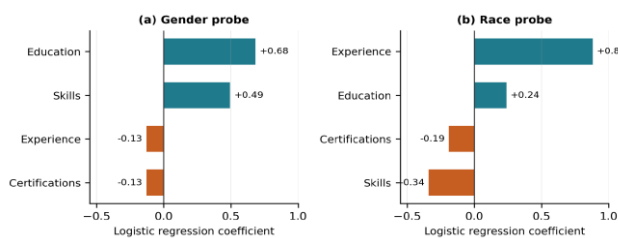


Fig. 4. Signed section-similarity coefficients for predicting each protected attribute. Positive coefficients (teal) increase the odds of the probed class; negative coefficients (orange) decrease them.

The gender attribution is consistent with the probing results and can be read at face value: education and skills content is where gender-associated signal concentrates. The race attribution requires the qualification established in Section 4.5. Because the underlying race probe does not exceed its majority-class floor, the experience coefficient should be treated as a descriptive association within this sample and not as evidence of recoverable race encoding. It is reported because it is a plausible hypothesis worth testing on balanced data: employer names, career continuity and geographic markers concentrated in experience sections, and there are substantive reasons to expect a signal there but the present experiment cannot confirm it.

G. Fairness-aware mitigation

Table VI decomposes the contribution of each pipeline stage. Moving from cosine ranking (B1) to the supervised section-aware scorer (B2) reduced gender DPD from 0.151 to 0.072 and race DPD from 0.442 to 0.189. The subsequent group-aware score adjustment reduced gender DPD further, from 0.072 to 0.048, but left race DPD unchanged at 0.189, with per-group selection rates identical to three decimal places. Test F1 is unchanged at 0.777 throughout, since the adjustment modifies ranking scores rather than classification decisions. The majority of the observed disparity reduction is therefore attributable to the change in scoring model rather than to the fairness intervention applied on top of it.

Figure 5 shows the mechanism at group level. The shifts occur almost entirely at the B1 to B2 transition: the non-binary selection rate falls from 0.429 to 0.286, the Hispanic rate from 0.600 to 0.400, and the Asian rate rises from 0.158 to 0.211. The subsequent adjustment leaves the race rates untouched and makes only a small further change on gender, raising the Female rate and lowering the Male rate. Residual race disparity remains substantially larger than residual gender disparity.

TABLE VI. UTILITY AND FAIRNESS ACROSS PIPELINE CONFIGURATIONS

System	Test F1	Gender DPD	Race DPD
B1 — cosine ranking	—	0.151	0.442
B2 — section-aware LR scorer	0.777	0.072	0.189
Fair scorer (adjusted)	0.777	0.048	0.189
Fair scorer + soft re-ranking	0.777	0.048	0.189

The soft-penalty re-ranking stage left both DPD values exactly unchanged at 0.048 and 0.189. Direct comparison of the selected sets confirms that re-ranking altered the ordering within each job's top ten but not its membership: no candidate crossed the selection boundary. The penalty applies only to candidates whose group is over-represented relative to its share of the candidate pool, and scales with that excess; at the values tested it did not exceed the score gaps at the cut-off, which ranged from 0.003 to 0.018 across the five jobs. The entire fairness gain in this pipeline comes from the scoring adjustment. This is a negative result for the re-ranking component as configured, and is reported as such rather than presenting the combined system as though both parts contributed. Its practical value, if any, is as a guard against over-correction, not as a source of parity improvement.

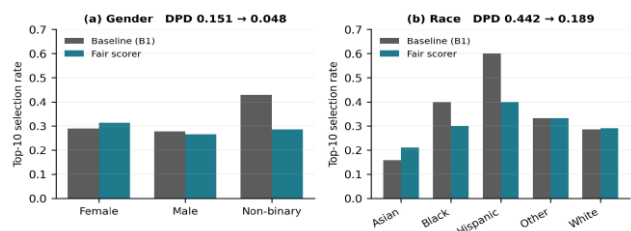


Fig. 5. Top-10 selection rates by group before and after group-aware score adjustment.

H. Alignment between utility and fairness objectives

The relevance scorer's own coefficients were examined against the attribution results. The scorer assigns its largest positive weights to skills similarity and full-document similarity; experience and education similarity receive negative weights, and certifications a small positive one. The large full-document coefficient coexists with the ablation finding that removing this feature does not degrade performance, indicating collinearity with the section features rather than independent contribution. The comparison is worth stating precisely: the two sections the attribution analysis associated most strongly with protected attributes (education for gender, experience for race) are the two the relevance model independently down-weights, while the section it relies on most heavily, skills, is the one least associated with gender in the probing results.

The scorer had no access to protected attributes and no fairness objective; it was fitted solely to predict relevance. The alignment therefore reflects a property of this dataset: the generator constructed relevance to depend primarily on skills and overall fit rather than a general law about resume screening. In a domain where the most predictive section were also the most demographically informative, the two objectives would pull apart and the trade-off this pipeline avoided would be unavoidable. The section-level decomposition is nonetheless what makes the relationship visible at all, and being able to see whether utility and fairness point the same way is useful regardless of which way they point. The large full-document coefficient coexists with the ablation finding that removing the feature does not degrade performance, indicating collinearity with the section features rather than independent contribution.

V. DISCUSSION

A. Interpretation

Three findings carry beyond this dataset. The first is that section-level cosine similarities are sufficient for competitive relevance prediction. Five interpretable numbers reproduce most of the signal available in the full document, which means the interpretability of this design is close to free rather than purchased with accuracy. For a domain where a rejected candidate may reasonably ask why, that matters [23].

The second is that disparity in outcomes and recoverability of a protected attribute are distinct properties that can move in opposite directions. Race produced the larger selection disparity by a wide margin, yet was the attribute the probes could not recover at all. An audit that used probe performance as its primary instrument would have concluded that race posed the smaller risk, which is the opposite of what the selection rates show. Selection-rate auditing and representational auditing answer different questions, and a pipeline audit needs both.

The third is the null result on re-ranking. The fair-ranking literature would suggest that a post-processing re-ranker should carry meaningful weight [8]. Here it carried none, because the penalty never exceeded the score gaps at the selection boundary, leaving the selected set unchanged. This is not evidence that re-ranking is ineffective in general; it is evidence that stacking interventions without decomposing their contributions can produce a system whose reported fairness gain is credited to components that did nothing. Reporting the decomposition is inexpensive and is encouraged here as standard practice.

The same pattern appears one stage earlier. The group-aware score adjustment, which the pipeline might plausibly be credited with, produced no change in race parity whatsoever, and only a modest further improvement in gender parity beyond what the scoring model had already achieved. Decomposition is what makes this visible; a two-point comparison between the cosine baseline and the fully mitigated system would have attributed the entire gain to the mitigation.

B. Relation to prior work

The gender findings reported here corroborate the established result that demographic signal survives the removal of explicit identifiers [5], [12]. The AUC of 0.998 is at the high end of what has been reported, which the synthetic generator's construction may partly explain: generated CVs are likely to encode gendered patterns more consistently than naturally written ones. The section-level localisation extends this work rather than replicating it. To the author's knowledge, no prior study reports which resume sections carry protected-attribute signal in an embedding-based screener, and the finding that gender cues are distributed across sections but concentrated in education and skills is new.

On mitigation, the result obtained here is consistent with the broader post-processing literature in achieving substantial disparity reduction at negligible utility cost [18], and with the observation that a single mechanism serves some attributes better than others [24]; residual race DPD remained roughly four times residual gender DPD. Where this study departs from much of that literature is in reporting the fairness and utility of every configuration side by side, including the configuration where an intervention did nothing.

C. Limitations

Sample size is the most consequential constraint. Fairness metrics were computed on a 173-pair test partition yielding roughly 50 top-10 selections, so group selection rates for smaller categories rest on a handful of individuals and are correspondingly unstable. No confidence intervals are reported. The reported DPD values describe this sample reliably; they should not be treated as precise population estimates, and replication with bootstrap intervals on a larger test partition would be a necessary step before drawing operational conclusions.

The race analysis is inconclusive rather than negative. With 80.8% of the sample in a single category and as few as 55 in the smallest, the design lacks the statistical power to detect race encoding even where it exists. The section-level race attribution should be read as hypothesis-generating.

Ecological validity is limited by the synthetic data. The FINDHR generator draws on real distributions and is purpose-built for fairness research [22], but synthetic corpora reproduce neither the linguistic variety nor the ambiguity of authentic resumes, and prior work has shown that models can achieve implausibly clean performance on synthetic screening data [25], [26]. The disparity directions observed here are properties of the generator, not claims about labour markets. Because the labels derive from skills matching, part of the weight the scorer assigns to skills similarity may reflect the labelling rule rather than an independent finding.

Finally, the fairness framing is narrow. Demographic parity is evaluated at a single threshold. Equal opportunity, calibration, individual fairness and intersectional criteria are not assessed, and parity is itself sensitive to base rates in the way [17] describe, a caveat with direct force given the class

distribution of this dataset. Candidate perceptions of procedural justice and jurisdiction-specific legal compliance are outside scope entirely [27].

VI. CONCLUSION

This paper treated the resume section rather than the resume as the unit of fairness analysis in AI-assisted screening. Five section-level cosine similarities supported competitive relevance prediction (test F1 0.777) while making each section's contribution directly readable. Auditing the baseline revealed substantial selection disparity, with a race DPD of 0.442 and a gender DPD of 0.151, and a group-aware score adjustment the supervised section-aware scorer reduced these to 0.072 and 0.189 without utility loss, with a subsequent group-aware adjustment lowering gender DPD to 0.048 while leaving race DPD unchanged.

The diagnostic results were asymmetric in an instructive way. Gender proved almost perfectly recoverable from the embeddings and localisable to the education and skills sections, while race (the attribute with by far the larger outcome disparity) was not linearly recoverable at all under the class distribution of this dataset. Large selection disparities can coexist with a protected attribute that no linear probe detects, which means an audit resting on either instrument alone will produce a misleading picture.

A null result is also reported: soft-penalty re-ranking, applied on top of the score adjustment, changed neither fairness metric. Decomposing a multi-stage intervention is inexpensive and prevents a pipeline from taking credit for components that are inert.

The immediate extensions are to replicate on demographically balanced data with bootstrap confidence intervals; and to test whether the section-level patterns observed here hold on authentic resumes across labour markets and languages. Section-level modelling is best understood as a diagnostic layer within a sociotechnical governance process, not as a standalone remedy for unfair hiring.

VII. ETHICAL STATEMENT

This research used exclusively synthetic CVs generated by the FINDHR Synthetic CV Generator and involved no human participants and no personal data. Institutional ethics approval was obtained from Leeds Beckett University. Demographic attributes were used solely to evaluate fairness and were never supplied as features to any predictive model. No model developed here was applied to any live recruitment process, and no real candidate's outcome was affected. The pipeline is a research prototype; practitioners adapting these methods would need to conduct further validation, including legal review and stakeholder impact assessment, for their specific deployment context.

REFERENCES

- [1] P. Budhwar, A. Malik, M. T. T. Silva, and P. Thevisuthan, "Artificial intelligence – challenges and opportunities for international HRM: a review and research agenda," *Int. J. Hum. Resour. Manag.*, vol. 33, no. 6, pp. 1065–1097, 2022, doi: 10.1080/09585192.2022.2035161.
- [2] P. van Esch, J. S. Black, and J. Ferolie, "Marketing AI recruitment: the next phase in job application and selection," *Comput. Hum. Behav.*, vol. 90, pp. 215–222, 2019, doi: 10.1016/j.chb.2018.09.009.
- [3] A. L. Hunkenschroer and C. Luetge, "Ethics of AI-enabled recruiting and selection: a review and research agenda," *J. Bus. Ethics*, vol. 178, no. 4, pp. 977–1007, 2022, doi: 10.1007/s10551-022-05049-6.
- [4] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: challenges and a path forward," *Calif. Manage. Rev.*, vol. 61, no. 4, pp. 15–42, 2019, doi: 10.1177/0008125619867910.
- [5] K. Wilson and A. Caliskan, "Gender, race, and intersectional bias in resume screening via language model retrieval," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2024, pp. 1578–1590, doi: 10.1609/aies.v7i1.31748.
- [6] H. Iso, A. Kumar, P. Pezeshkpour, and N. Bhutani, "Evaluating bias in LLMs for job-resume matching: gender, race, and education," in *Proceedings of NAACL-HLT*, 2025, pp. 672–683, doi: 10.18653/v1/2025.naacl-industry.55.
- [7] H. Gonen and Y. Goldberg, "Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them," in *Proceedings of NAACL-HLT 2019*, 2019.
- [8] S. C. Geyik, S. Ambler, and K. Kenthapadi, "Fairness-aware ranking in search and recommendation systems with application to LinkedIn Talent Search," in *Proceedings of KDD 2019*, 2019, pp. 2221–2231, doi: 10.1145/3292500.3330691.
- [9] N. Reimers and I. Gurevych, "Sentence-BERT: sentence embeddings using Siamese BERT-networks," in *Proceedings of EMNLP-IJCNLP 2019*, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
- [10] R. V. K. Bevara, N. R. Mannuru, S. P. Karedla, and T. Xiao, "Resume2Vec: transforming applicant tracking systems with intelligent resume embeddings for precise candidate matching," *Electronics*, vol. 14, no. 4, p. 794, 2025, doi: 10.3390/electronics14040794.
- [11] N. Firdaus, B. Kusuma Riasti, and M. Asri Safi'ie, "Comparative study of transformer-based models for automated resume classification," *JITK*, vol. 11, no. 2, pp. 372–381, 2025, doi: 10.33480/jitk.v11i2.7453.
- [12] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2016, doi: 10.1126/science.aal4230.
- [13] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: evaluating claims and practices," in *Proceedings of FAT* 2020*, 2020, doi: 10.1145/3351095.3372828.
- [14] J. Sánchez-Monedero, L. Dencik, and L. Edwards, "What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems," in *Proceedings of FAT**, 2020, pp. 458–468, doi: 10.1145/3351095.3372849.
- [15] A. Peña, I. Serna, A. Morales, and J. Fierrez, "Addressing bias in LLMs: strategies and application to fair AI-based recruitment," 2025.
- [16] H. Weerts, R. Xenidis, F. Tarissan, and H. P. Olsen, "Algorithmic unfairness through the lens of EU non-discrimination law: or why the law is not a decision tree," in *Proceedings of FAccT 2023*, 2023, pp. 805–816, doi: 10.1145/3593013.3594044.
- [17] D. F. Mujtaba and N. R. Mahapatra, "Fairness in AI-driven recruitment: challenges, metrics, methods, and future directions," 2024.
- [18] M. Hort, Z. Chen, J. M. Zhang, and M. Harman, "Bias mitigation for machine learning classifiers: a comprehensive survey," *ACM J. Responsible Comput.*, vol. 1, no. 2, pp. 1–52, 2022, doi: 10.1145/3631326.
- [19] S. Biswas and H. Rajan, "Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline," *ESEC/FSE*, pp. 981–993, 2021, doi: 10.1145/3468264.3468536.
- [20] M. Zehlike, K. Yang, and J. Stoyanovich, "Fairness in ranking: a survey," *ACM Comput. Surv.*, 2022, doi: 10.1145/3533379.
- [21] T. Pimentel and R. Cotterell, "A Bayesian framework for information-theoretic probing," in *Proceedings of EMNLP 2021*, 2021, pp. 2869–2887, doi: 10.18653/v1/2021.emnlp-main.229.
- [22] J. Saldivar, A. Gatzoura, and C. Castillo, "Synthetic CVs to build and test fairness-aware hiring tools," *ACM Transactions on Intelligent Systems and Technology*, 2025.
- [23] S. Vaishampayan, S. Farzanehpour, and C. Brown, "Procedural justice and fairness in automated resume parsers for tech hiring: insights from candidate perspectives," *VL/HCC*, pp. 103–108, 2023, doi: 10.1109/VL-HCC57772.2023.00019.
- [24] F. Hu, P. Ratz, and A. Charpentier, "A sequentially fair mechanism for multiple sensitive attributes," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 11, pp. 12502–12510, 2024, doi: 10.1609/aaai.v38i11.29143.
- [25] L. Casoria, F. Bertolotti, and A. Locoro, "Evaluating LLMs for synthetic personas generation: a comparative analysis of personality

- representation and censorship effects,” in *CHIItaly 2025*, 2025. doi: 10.1145/3750069.3750142.
- [26] A. Haughey, B. Mann, and S. Oca, “*Case study: using synthetic datasets to examine bias in machine learning algorithms for resume screening.*” American Society for Engineering Education, 2025. doi: 10.18260/1-2-56071.
- [27] [27] D. Kumar, T. Grosz, N. Rekabsaz, and E. Greif, ““Fairness of recommender systems in the recruitment domain: an analysis from technical and legal perspectives,”” *Front. Big Data*, vol. 6, p. 1245198, 2023, doi: 10.3389/fdata.2023.1245198.