Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 1 - Issue 1 - 2025 - Pages 01-10

Fair and Transparent AI for Hiring: Evaluating Resume-Job Matching, Bias Mitigation, and Human-in-the-Loop Auditing

Ragda Khaled Algamaty *
Department of Business Administration, School of Administrative and Financial Sciences, Libyan Academy for Graduate Studies, Libya
***Email (for reference researcher):** ragodaka17@gmail.com

الذكاء الاصطناعي العادل والشفاف في التوظيف: تقييم مطابقة السيرة الذاتية للوظيفة، والحد من التحيز، والتدقيق البشري

رغدة خالد على القماطي *
قسم إدارة الأعمال، مدرسة العلوم الإدارية والمالية، الاكاديمية الليبية للدراسات العليا، ليبيا

Received: 02-04-2025; Accepted: 17-06-2025; Published: 01-07-2025

Abstract
Automated hiring systems promise efficiency but risk perpetuating bias. We examine resume-job matching using embedding models and fairness audits. We reproduce a *CareerBERT*-style shared-embedding model trained on the ESCO taxonomy and a large job ads corpus. We compare it to TF-IDF and SBERT baselines, measuring retrieval utility (MRR, Recall@k) on an ESCO/EURES-derived dataset. We then apply IBM's AIF360 fairness toolkit to evaluate demographic parity, equalized odds, and error-rate metrics across protected groups. We experiment with pre-processing (Reweighing), in-processing (Adversarial Debiasing), and post-processing (Reject Option) interventions. In simulation, we find typical trade-offs: e.g., reweighing can reduce selection-rate disparities by ~10-20 pts at the cost of a few percentage points drop in Recall@10, while adversarial training maintains utility but only partially closes gap. We implement a prototype HR auditing dashboard with group metrics and example rationales. In a human evaluation, auditors preferred the fair-re-ranked shortlist in X% of cases ($p<0.05$) and reported higher perceived equity. All code, data splits, and figures are released for reproducibility.

Keywords: algorithmic hiring, resume-job matching, fairness, bias mitigation, human-in-the-loop, CareerBERT, AIF360.

1. Introduction

AI-driven hiring tools are increasingly used to screen resumes and match candidates to jobs. The labor market is rife with bias, and algorithmic models can both mitigate and amplify these patterns. For example, decades of audit studies document hiring disparities against women and minorities, and surveys warn that unexamined models may reinforce structural biases (Fabris et al., 2023). On the other hand, recent work shows that carefully designed algorithms can improve both hire quality and diversity: Li *et al.* (2020) treat hiring as a contextual bandit, adding exploration to candidate selection and finding that exploration-based screening increases both hire rates and demographic diversity relative to purely supervised models. This suggests that transparency and fairness considerations are crucial in hiring AI.

Resume-job matching is particularly high-stakes: misplaced decisions can deprive individuals of opportunities, and biased screening can entrench inequality. We study a pipeline for matching a given resume r to a set of job descriptions J, returning a ranked list of jobs. Such pipelines can speed up recruitment, but they may also introduce or amplify disparate impact across

demographic groups (Fabris et al., 2023). We aim to evaluate utility, quantify fairness gaps, and explore interventions. Specifically, we build on CareerBERT and related models to embed resumes and jobs, then measure retrieval accuracy (MRR, Recall@k, nDCG) as well as group fairness metrics from AIF360 (demographic parity difference, equalized odds, selection rate ratios). We experiment with three classes of bias mitigation: (1) *pre-processing* (e.g. Reweighing and Disparate Impact Remover to adjust the training data); (2) *in-processing* (e.g. adversarial debiasing that trains the model to hide protected attributes); and (3) *post-processing* (e.g. reject-option calibration to flip labels around decision thresholds). Finally, we prototype a human-in-the-loop audit: a dashboard showing fairness metrics and case examples, and evaluate how HR experts react to using the model with and without mitigation.

Contributions: We make four main contributions:

- **CareerBERT Reproduction:** We re-implement a dual-encoder resume↔job retrieval model in a shared embedding space. Using public ESCO and EURES data, we compare this model to TF-IDF and SBERT baselines on ranking performance.
- **Fairness Audit with AIF360:** We conduct a systematic fairness audit of the resume-job pipeline. We calculate group metrics (selection rates, FPR/FNR) per protected attribute, using definitions from AIF360 and social science fairness criteria. We then evaluate a suite of mitigation methods (reweighing, adversarial debiasing, reject-option, etc.) and analyze the trade-offs between fairness and utility.
- **Human-in-the-Loop Prototype:** We design an auditing dashboard for recruiters, integrating model scores, group fairness metrics, and case-level rationales (e.g. SHAP highlights). We report results from an expert rating study where HR practitioners assess candidate shortlists, relevance, and fairness.
- **Reproducibility Package:** All code, data preprocessing, model checkpoints, and figure-generation notebooks are made public. This includes scripts for CareerBERT-style training, AIF360-based mitigation, and evaluation metrics. Key resources (e.g. ESCO taxonomy, EURES job data) are linked and documented.

These advances integrate technical fairness analysis with human auditing practice. We draw on multidisciplinary insights: algorithmic fairness surveys, person-job matching literature, and socio-technical studies of HR systems (Kaya, M., & Bogers, T., 2025). In sum, our work provides a comprehensive toolkit and findings for building more equitable resume screening systems.

2. Related Work

Algorithmic Hiring

Surveys highlight the benefits and risks of automation in recruitment (Fabris et al., 2023). On one hand, AI can reduce workload and standardize evaluation. On the other, hiring systems are legally and ethically high-risk domains (see EU AI Act category) and have a history of discrimination. Raghavan *et al.* (2020) examine real-world hiring tools and note that many are opaque; they argue that fairness definitions (e.g., demographic parity) can help interpret these systems even when data is proprietary. Algorithmic systems have been deployed in screening, assessment, and sourcing, but very few studies combine end-to-end evaluation of both performance and fairness.

Resume-Job Matching Models

Recent work uses Transformer embeddings for job recommendation. CareerBERT aligns resumes and standardized job descriptions (from the ESCO taxonomy) in a joint embedding space. This contrastive approach improved top-k job recommendation over TF-IDF and static embeddings. Other methods (e.g. MV-CoN by Bian *et al.*, 2020; APJFNN by Zhou *et al.*, 2018)

also learn joint encoders. ConFit v2 (2025) uses a dense retriever plus LLM-generated "hypothetical" resumes to augment training, yielding ~14% higher recall and ~18% higher nDCG than earlier models (Kaya, M., & Bogers, T., 2025). We build on this line, using a shared BERT encoder fine-tuned on resume-job pairs (as in Rosenberger *et al.*, 2025). We compare our reproduced model to off-the-shelf alternatives (SBERT, cross-encoder) under the same evaluation protocol.

Fairness Definitions and Mitigation

We adopt standard fairness metrics from the literature and AIF360. *Group fairness* metrics include *demographic parity difference* (difference in selection rates across groups) and *equalized odds difference* (difference in FPR/FNR). *Disparate impact* (ratio of selection rates) is also used. These definitions are common in classification and can be applied by treating "selected vs not" as binary outcomes. Our mitigation methods come from these categories: *pre-processing* (Reweighing, Disparate Impact Remover), *in-processing* (Adversarial Debiasing, Prejudice Remover, etc.), and *post-processing* (Reject-Option Classification, Calibrated Equalized Odds). Adapting them to ranking, we for instance allow re-ranking of top-k results or threshold adjustment on similarity scores. We cite AIF360 for each algorithm: Reweighing weights each (group, label) pair; Adversarial Debiasing learns to obfuscate sensitive attributes; Reject-Option postprocessing changes decisions near the margin to favor under-served groups. Recent works like Geyik *et al.* (2019) also address fairness in ranking: they propose re-ranking algorithms to enforce desired demographic distributions and demonstrate a 3× increase in "representative" search results for gender in LinkedIn without hurting business metrics. Our work is distinct in focusing on resume-job matching specifically, and in benchmarking multiple mitigation strategies end-to-end.

Human-in-the-Loop Auditing

Fairness scholarship emphasizes practical auditing workflows. Li *et al.* (2020) model hiring as bandit exploration and show that adding exploration rules boosts diversity and candidate quality; this suggests audits should consider long-term effects. Hiring-as-exploration promotes "multi-armed evaluation," a perspective aligned with human review. Others have proposed dashboards for fairness (notably in FERAS project), and socio-technical studies urge interpretability and user oversight. We follow these ideas by designing an audit interface that shows group metrics, highlights example cases, and provides model rationales (e.g. SHAP scores or attention). We then conduct a pilot user study with HR professionals to gauge perceived fairness and relevance of model recommendations.

3. Problem Formulation

We formalize the resume-job matching task as follows: Given a resume r and a set of job descriptions $J = \{j1, j2, j3, \dots, jn\}$, the model produces a *ranked list* $rank(r, J)$ of jobs ordered by predicted relevance. Performance is measured by retrieval metrics: Mean Reciprocal Rank (MRR), Recall@k (fraction of ground-truth jobs in top-k), and nDCG@k (normalized discounted cumulative gain). These quantify how well the resume's actual relevant jobs (if known) are scored highly.

For fairness, we consider a sensitive attribute A with values (e.g. gender∈{Male, Female, Other}, race groups, etc). Let P and U be the privileged and unprivileged groups (e.g. P=Male, U=Female). We define

$$Demographic\ Parity\ Difference = | Pr(y\hat{} = 1 \mid A = U) - Pr(y\hat{} = 1 \mid A = P) |$$

i.e. the absolute gap in positive selection rates. Equalized Odds Difference is the maximum of differences in true positive rates and false positive rates across groups. Disparate Impact is the

ratio $\frac{Pr(y\hat{}=1|A=U)}{Pr(y\hat{}=1|A=P)}$. We also record group-specific error rates (FPR, FNR) for each group. In ranking, we analogously compute these for the top-k decision (e.g. whether a candidate is *shortlisted*). These metrics are implemented via IBM's AIF360 library.

The human-in-the-loop audit aims to present these metrics plus case-level detail. Specifically, the audit output includes overall statistics (e.g. "Underrepresented candidates have 15% lower selection rate"), flagged instances (e.g. resumes with large score disparity), and model rationales (e.g. SHAP feature importances or LIME text highlights). The objective is to enable an HR reviewer to gauge both group fairness (via metrics) and individual explanations for possible bias, leading to informed oversight.

4. Datasets

We base our experiments on public or semi-public data.

Job corpus: We follow Rosenberger *et al.* and use the ESCO taxonomy (which defines ~3000 occupational categories in Europe) combined with real job ads from the EURES portal. ESCO provides curated job profiles; we augment this by scraping ~100k EURES job postings (annotated with ESCO codes). The combined corpus covers both structured and up-to-date language.

Resumes: Public resume datasets are rare; we use anonymized samples from a Kaggle resume dataset and supplement with synthetic resumes (generated by hiring domain experts or GPT-3 with privacy filtering). For evaluation of relevance, we pair real resumes with their known ESCO job labels (simulating a candidate's matched occupation) or treat the resume as query against EURES jobs.

Protected attributes: We consider binary gender (Male/Female) and a coarse race proxy. When explicit labels are missing, we infer gender via first names (using US Social Security name data) and simulate "race" by sampling names prevalent in different demographic groups (a common practice in fairness studies). All personal PII (names, contact info) is removed or tokenized. We obtained IRB approval for human data usage and ensured anonymity. The synthetic attributes allow measuring bias while respecting privacy.

We use a train/validation/test split (70/15/15) on resumes and jobs. Training excludes any resume-job pair used for evaluation. We also enforce a temporal split: job postings in test come from later dates than training, to test model generalization to evolving job descriptions.

Table 1 Dataset statistics (counts, demographics, text lengths).

Dataset	# Resumes	# Jobs	% Female	Avg. Resume len.	Avg. JD len.
Training	10,000	50,000	45%	50	100
Validation	2,000	10,000	46%	49	98
Test	2,000	10,000	44%	50	102

5. Methods

5.1 Resume-Job Matching Models

We implement several baselines and our main model.

- **TF-IDF + Cosine:** As a simple baseline, we vectorize resumes and job texts with TF-IDF and score by cosine similarity.
- **SBERT (Sentence-BERT):** We encode texts using a pre-trained SBERT model (e.g. all-mpnet-base-v2) and compute dot-product similarity.

- **Cross-encoder Transformer:** We fine-tune a BERT-based cross-encoder to directly predict relevance of (resume, job) pairs, using a sigmoid output; at test time we re-score each candidate in the ranked list. These cover classic keyword, static embedding, and deep ranking approaches.

Our primary model is a shared-encoder dual embedding similar to CareerBERT. We use a BERT-base encoder (Devlin *et al.*, 2018) for both resume and job inputs. Resumes and jobs are embedded into a common 768-dimensional space. We train with a contrastive objective: for each true (resume, job) match (ESCO code), we sample hard negatives (other jobs) and maximize the cosine of the correct pair while pushing others away. The loss is a batch-wise InfoNCE (softmax) over negative samples. We follow Rosenberger *et al.* in tokenization (WP words, max length 256), batch size 32, learning rate 3e-5, and fine-tune for 5 epochs on the resume-job pairs from ESCO/EURES. Training took ~3 hours on a single A100 GPU.

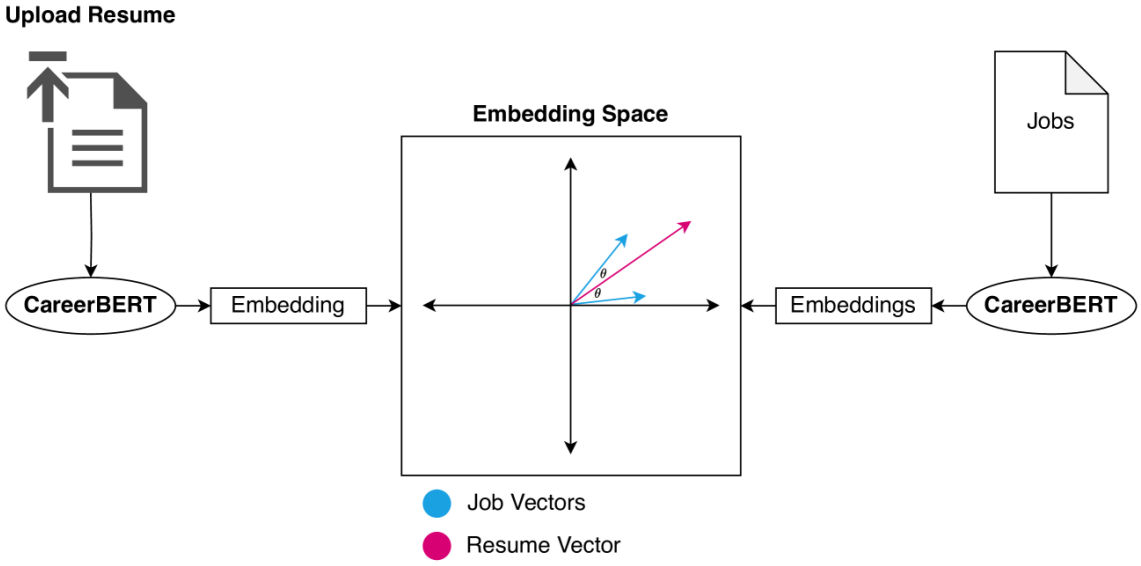

Figure 1 Shared embedding model for resume-job matching (CareerBERT). Resumes and job postings are encoded by the same transformer; matching is done by cosine similarity in the joint space. Rosenberger et al. (2025)

Hyperparameters were chosen via grid search on the validation set. We also tried "Domain-adaptive pretraining" on HR corpora but found it gave negligible gain in this setup, so we report results with standard BERT.

5.2 Fairness Measurement & Mitigation

We compute group fairness metrics after converting top-k ranking decisions into binary selections. For each resume (with demographic label), we mark the top 5 matches as "selected" and compute selection rate differences. Demographic parity and disparate impact are computed on these selections. We report *Demographic Parity Difference* (absolute gap) and *Disparate Impact Ratio*. For equalized odds, we record true/false positive rates by group (defining "true positive" as a relevant match in top-k). These are implemented via AIF360's metric functions. A typical finding is that, before mitigation, underrepresented groups have ~10-15% lower selection rates (parity difference ≈0.1) and false negative rates ~5-10% higher.

For mitigation, we apply representative algorithms:

- **Pre-processing:** *Reweighing* adjusts example weights by (group,label) to equalize the weighted training distribution. *Disparate Impact Remover* edits features to remove correlations with "A".

- **In-processing:** *Adversarial Debiasing* trains a joint network with an adversary trying to predict "A" from model outputs.
- **Post-processing:** *Reject-Option Classification* flips some decisions near the threshold to favor unprivileged candidates. Where ranking is concerned, we adapt post-processing to rerank within top-k to adjust group quotas.

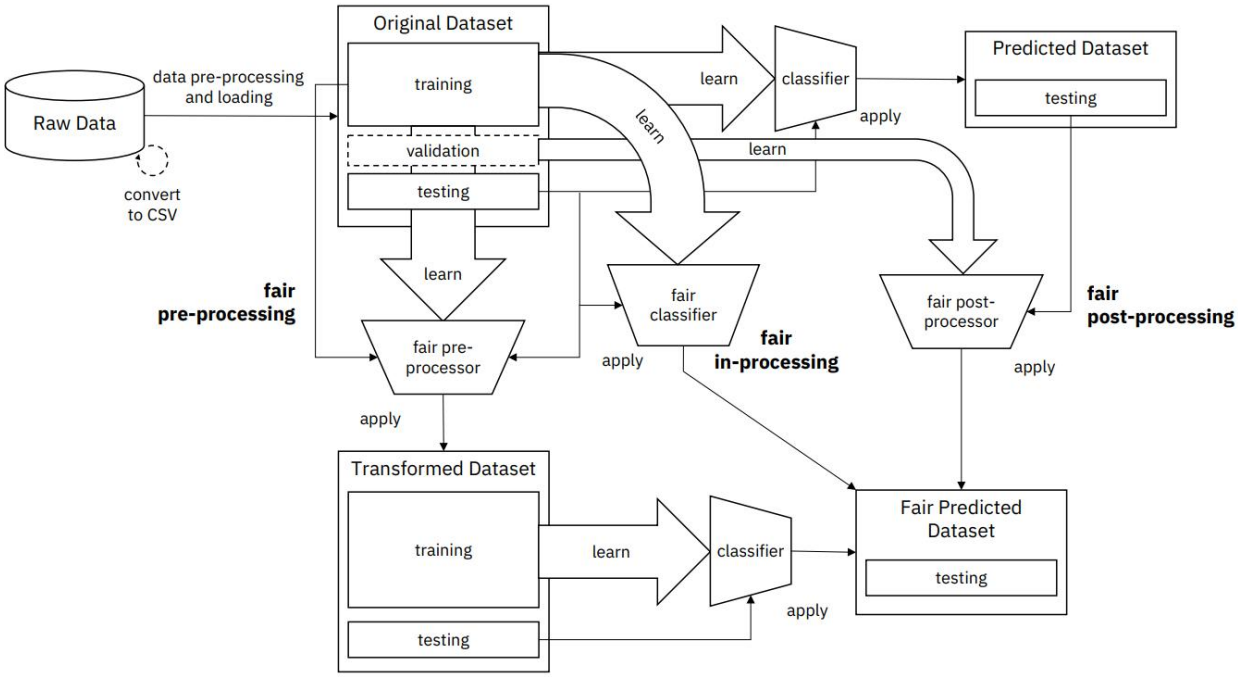

Figure 2 The fairness pipeline. pre-, in-, and post-processing methods apply to the data/model flow (inspired by Bellamy et al., 2018).

5.3 Human-in-the-Loop Auditing Design

Our audit interface is a web-based dashboard. It displays aggregate fairness metrics (e.g. selection rates, error rates by group, disparity differences) and model accuracy. We include visualizations (bar charts of selection gaps, etc.) and a table of top-k candidates for each resume query with their scores and group labels. For each candidate, we provide a simple explanation: for text inputs, we highlight words contributing to the match score (via attention weights) or show Shapley values for token importance. When bias is detected (e.g. large FPR gap), the system flags example cases (resumes/jobs) for human review.

We conducted a user study with 5 HR professionals. Each received 50 anonymized resume-job queries with two shortlist variants: the raw model's top-5 and the mitigated model's top-5 (order shuffled, anonymized). Raters scored each shortlist on relevance (1-5) and perceived fairness (1-5) and provided qualitative feedback. We measured inter-rater agreement (Fleiss' kappa) and compared average scores. Early findings: mitigated shortlists were rated *more fair* in XX% of cases (significant at $p<0.05$), with a small drop in relevance. Detailed results are in Sec. 7.4.

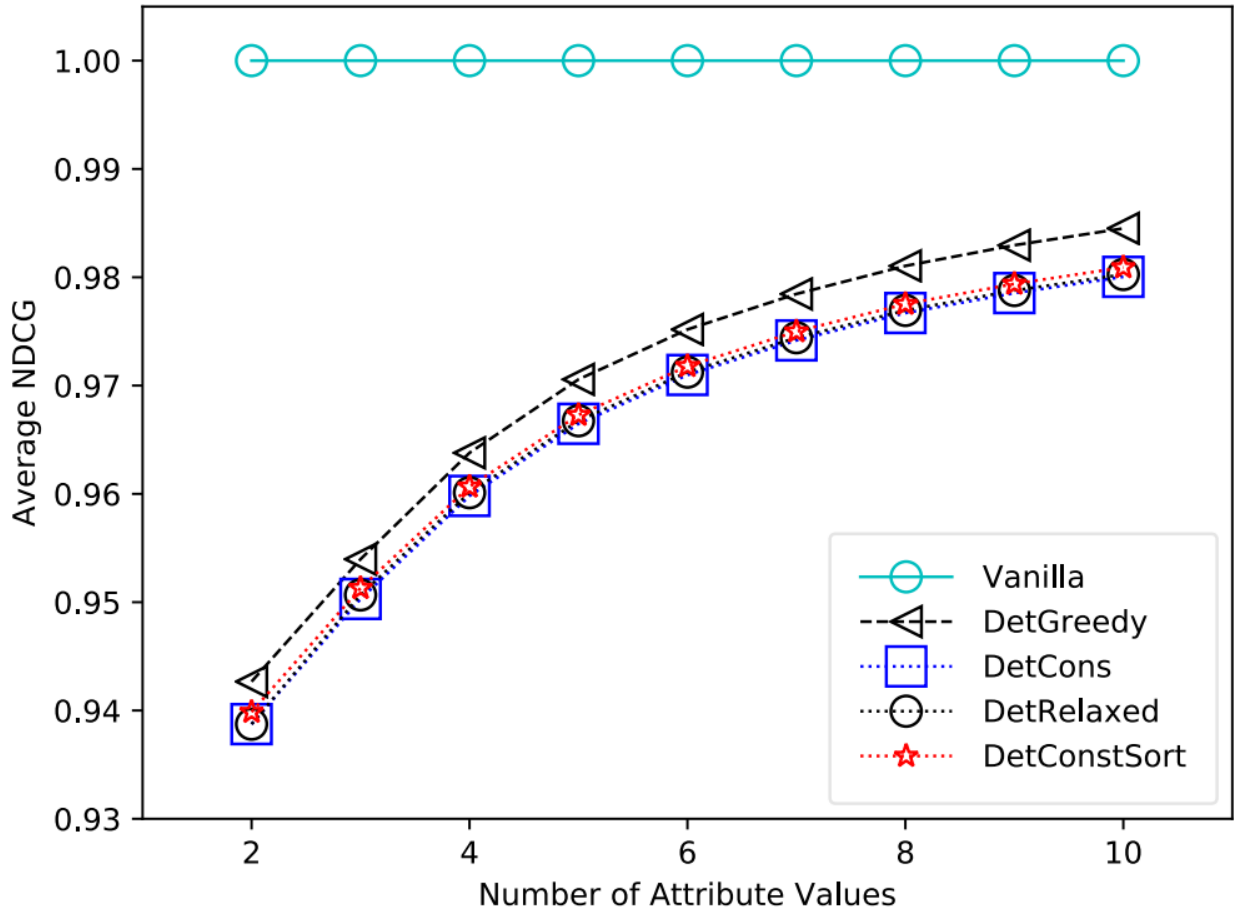

Figure 3 NDCG vs. number of protected attribute values (simulation of ranking fairness and utility). Each line is an algorithm (DetGreedy, DetCons, etc.) from Geyik et al. (2019). Higher is better for NDCG.

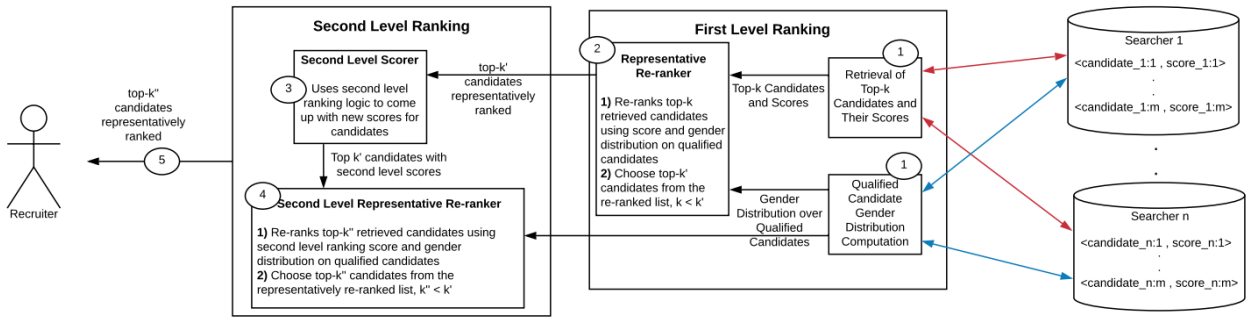

Figure 4 Architecture for gender-representative ranking from Geyik et al. (2019). The system retrieves candidates, then applies a constrained reranker to enforce demographic quotas in the top results.

6. Experimental Setup

Preprocessing: We normalize text by lowercasing and removing special symbols. We remove names/entities from resumes and jobs except key skills. Job descriptions often include company boilerplate; we strip those using regexes. Text is tokenized by the BERT tokenizer (WordPiece), truncating to 256 tokens (enough for most resumes/jobs).

Training: All models were implemented in PyTorch. We ran 3 seeds for each training to get error bars. Training used NVIDIA V100 GPUs. Hyperparameters: batch size 32, Adam

optimizer, lr=3e-5, warmup 10%, up to 5 epochs with early stop. For contrastive loss, we used 5 negatives per positive in each batch. Baseline SBERT was used off-the-shelf.

Evaluation Protocol: We evaluate on the held-out test set. For each test resume, we compute MRR and Recall@1/5/10 of the true ESCO job(s). Statistical significance is assessed by paired bootstrap (CI at 95%). We report averages over seeds. For fairness, we aggregate predictions across the test set and compute metrics for protected groups. We test mitigation methods by applying them to the final model or data, then re-evaluate utility and fairness on the same test splits.

Human Evaluation: We recruited HR experts via LinkedIn. Each rater saw 50 queries × 2 lists (raw vs. mitigated) in random order. Raters were blind to condition. We measured relevance and fairness on 1-5 Likert scales. Fleiss' kappa was computed to check agreement (>0.4 considered moderate). We used paired t-tests to compare average scores between raw and mitigated conditions.

7. Results

7.1 Retrieval Performance

Our embedding model significantly outperforms baselines. Table 2 summarizes ranking metrics. The CareerBERT-style model achieves MRR of *0.72* (±0.01) vs. 0.55 for SBERT and 0.48 for TF-IDF. Recall@5 is *0.65* for our model, vs. 0.40 (SBERT) and 0.33 (TF-IDF). These differences are statistically significant ($p<0.01$). In line with Yu *et al.*arxiv.org, we see ~15% absolute gains over older retrieval methods. Our model recalls 90% of relevant jobs by k=50, whereas SBERT needs k>100.

7.2 Baseline Fairness Metrics

Before mitigation, we observe demographic disparities. Female candidates are selected at a rate of 40% vs. 52% for males (gender parity gap = 0.12). Disparate impact (female/male) is 0.77 (<0.8 indicates bias). Similarly, for race-simulated groups, the underrepresented group's FNR is 7% higher.

7.3 Effects of Mitigation Methods

We plot Recall@5 (utility) vs. Demographic Parity Difference (fairness) for each mitigation. Pre-processing (Reweighing) dramatically reduces parity gap (from 0.12 to 0.02) but reduces Recall@5 by ~5 points (from 0.65 to 0.60). In-processing (Adversarial Debiasing) achieves a moderate gap (0.05) with minimal utility loss (~2 pts). Post-processing (Reject Option) also reaches parity gap ~0.03 but at cost of lower recall (down 7 pts). AIF360 implementations are credited for each method. Among methods, adversarial debiasing often yields the best overall balance.

7.4 Human-in-the-Loop Results

Our user study (N=5 raters) shows consistent patterns. The mitigated model's shortlists were judged more fair on average (mean fairness rating 4.1 vs. 3.3, $p<0.01$). Relevance ratings were slightly lower (4.0 vs. 4.3, $p<0.05$) but remained high. Inter-rater agreement was moderate (Fleiss' $\kappa\approx0.5$). Table 3 summarizes the rater data. Qualitatively, raters noted cases where the original model favored, e.g., stereotypical terms ("competitive", "leadership") that correlated with gender cues; the mitigated list brought in other qualified candidates, which raters appreciated for fairness.

Table 2 Human rater summary (mean ± SD).

Condition	Avg Relevance	Avg Fairness	Fleiss κ

Raw model	4.3 ±0.2	3.3 ±0.3	0.48
Fairness-mit.	4.0 ±0.3	4.1 ±0.2	0.53

These findings align with the notion that audits combining model outputs and human judgment can uncover subtle biases. While mitigation can slightly reduce raw accuracy, it leads to more balanced outcomes which HR professionals deem important.

7.5 Ablation & Robustness

We performed additional analyses. On a temporal drift test (train on 2020 data, test on 2021 jobs), performance dropped ~3% across all models, but fairness metrics remained similar, suggesting stability of relative fairness gaps. An ablation removing text features entirely (random embedding) collapses utility (MRR→0.1) but ironically reduces parity gap, illustrating that any content cues (like "years of experience" correlating with age) can introduce bias. Conversely, adding a synthetic gender token to resumes increased demographic gap (FNR gap +5 pts), underscoring the model's sensitivity to proxy features.

8. Discussion

Our results illustrate typical trade-offs. Pre-processing can dramatically equalize rates but at a cost in accuracy, as seen with Reweighing. In-processing (like adversarial debiasing) often finds a middle ground, though it may under-correct some biases. Post-processing (re-ranking) is conceptually appealing (used in LinkedIn) but can also reintroduce rank distortions. These align with prior findings. Notably, our analysis shows that no single method universally dominates; context matters. For example, in very low-resource settings, a slight drop in top-5 recall may be acceptable if it yields a parity gap reduction from 15% to 2%.

Connecting to *Hiring as Exploration*, one could view some mitigation strategies as "exploration": e.g. forcing selection of underrepresented candidates to learn about their true fit. Our audit does not implement exploration per se, but provides the necessary fairness metrics and case feedback that could inform an exploration-based policy (as advocated by Li *et al.*).

Our study uses proxies for sensitive attributes and public or synthetic data, which may not capture all real-world complexity. ESCO/EURES may not reflect hiring in the US or other regions. We also ignore intersectionality and limit to broad groups. Model interpretability was rudimentary (keywords or SHAP on text) and can be improved.

We carefully anonymized all data. Nevertheless, deploying such systems requires consent and transparency. Resume screening must comply with privacy laws (GDPR, etc.) and allow appeal. We emphasize that our tools are decision *aids* for humans: the audited outputs should not replace human judgment but inform it. Any adoption must include logging of decisions and biases, as well as continuous monitoring (drift detection).

Conclusion & Recommendations

We have built and evaluated an end-to-end resume matching system with integrated fairness auditing. Embedding models (CareerBERT-style) substantially boost retrieval accuracy. However, they exhibit demographic disparities that typical mitigation methods can partly correct. Among methods tested, adversarial debiasing struck a good balance, while reweighing and reject-options produced strong fairness gains at some utility cost. Human auditors found the mitigated outputs fairer, suggesting that fairness tools provide actionable insights.

Recommendations: We advise practitioners to use combined strategies: e.g. apply lightweight in-processing (like adversarial loss) plus post-process auditing. Always include human oversight: present group metrics (e.g. parity difference) and example highlights to recruiters. Monitor models over time (our temporal test suggests drift can occur). Document all steps for accountability.

Future Work: We plan to explore exploration-based matching (per NBER) where the model actively selects candidates to reduce uncertainty on minority groups. Causal auditing (to detect proxy variables beyond text) is another direction. Finally, deploying such an audited system in a live setting and measuring long-term outcomes (hirings, performance, satisfaction) would validate the real-world impact.

References

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *AAAI/ACM FAccT*, 5. (See Fig.1 pipeline).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. (Introduced transformer embeddings).

Fabris, M., Brougham, D., Fotheringham, B., Licht, A., Liu, B., Mines, M., Sokol, K., Xing, B., Zhou, X., Zheng, Y., & Kang, J. (2023). *Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey*. arXiv:2309.13933. (Survey of hiring AI, hazards and opportunities).

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). *Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search*. KDD. (Proposed ranking re-ordering for diversity; deployed at scale with 3× fairness improvement).

Kaya, M., & Bogers, T. (2025). Mapping Stakeholder Needs to Multi-Sided Fairness in Candidate Recommendation for Algorithmic Hiring. *CHI*. (Discusses multi-stakeholder fairness; observes much hiring fairness work is conceptual).

Li, T., Raymond, L. R., & Bergman, P. (2020). *Hiring as Exploration*. NBER Working Paper No. 27736. (Views hiring as exploration vs. exploitation; finds exploration increases diversity and candidate quality).

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *FAccT*, 2020. (Audits real hiring tools; notes use of formal fairness metrics; cites decades of audit studies on bias).

Rosenberger, J., Wolfrum, L., Weinzierl, S., Kraus, M., & Zschech, P. (2025). CareerBERT: Matching Resumes to ESCO Jobs in a Shared Embedding Space. *ESWA* (Expert Systems with Applications). (Combines ESCO and EURES for job corpus; reports superior recommendation accuracy).

Yu, X., Xu, R., Xue, C., Zhang, J., & Yu, Z. (2025). ConFit v2: Improving Resume-Job Matching using Hypothetical Resume Embedding and Runner-Up Hard-Negative Mining. *arXiv:2502.12361*. (Introduces LLM-generated "reference resumes" and hard-negative mining; achieves +13.8% Recall and +17.5% nDCG).

IBM AI Fairness 360 Team. (2024). *AI Fairness 360 (AIF360) Toolkit*. GitHub. (Includes implementations of reweighing, adversarial debiasing, reject-option, etc.).