International Journal of Instruction, Technology, and Social Sciences
Vol. 5, 86-105, 2026
ISSN: 2716-6546
www.ijitsc.net

Bias-Audited Resume Screening with Calibrated Matching Scores, Selective Review, and Evidence-Grounded LLM Explanation Cards

Lucas Cui
Computer Science, University of Florida, Gainesville, FL, USA
lucas.cui_dev@yahoo.com

Abstract

Automated resume screening requires a valid target, calibrated uncertainty, a defensible review policy, and explanations anchored to candidate evidence. This study evaluates those requirements on the datasetmaster resume collection. The 4,817-record snapshot contains a heterogeneous block of 217 records and a standardized block of 4,600 records balanced across 46 technical roles. Because the collection has neither job descriptions nor hiring outcomes, the target is internal proxy role retrieval, not employability or candidate–vacancy fit. A leakage audit removed identity fields, summaries, job titles, education majors, and project prose. The standardized cohort was divided into 2,760 training, 920 calibration, and 920 test profiles. The final temperature-scaled linear support vector machine obtained 0.2489 accuracy, 0.2325 macro-F1, 0.3978 top-3 accuracy, and 0.0401 expected calibration error. A logistic ablation rose from 0.2348 accuracy on strict qualifications to 0.9076 with project descriptions and 1.0000 with direct role text, demonstrating severe leakage. At the 80%-coverage threshold, test coverage was 0.8130 and accepted accuracy was 0.2955. Accuracy fell to 0.1346 on 104 mapped heterogeneous profiles. A lexical support gate reviewed 99.08% of all 217 heterogeneous records while retaining 94.35% of core records. All 920 structured cards passed source checks; an LLM then verbalized one card per role, and all 46 summaries preserved the role, route, safety constraints, and at least one exact evidence span. The results support human-supervised role triage, not autonomous hiring.

Keywords

resume screening; role retrieval; algorithmic hiring; calibration; selective classification; counterfactual audit; explanation cards; human review

International Journal of Instruction, Technology, and Social Sciences
Vol. 5, 86-105, 2026
ISSN: 2716-6546
www.ijitsc.net

Introduction

Resume screening is an information-retrieval problem inside a consequential social process. Speed does not validate the target, score, error distribution, or explanation. Audits identify gaps between hiring-system claims and setting-specific evidence, including biased labels, proxy features, distribution shift, opaque explanations, and automation bias (Fabris et al., 2025; Raghavan et al., 2020). Fluent LLM reasons can also make weak matches appear authoritative and alter human judgment (Wilson et al., 2025).

This study audits the Advanced Resume Parser & Job Matcher collection (datasetmaster, 2023), comprising 4,817 structured JSON records. Because it contains no vacancies, interview decisions, performance evaluations, or suitability judgments, we formulate a narrower 46-way task: retrieve the role encoded by a profile's first recognized experience title from qualification text. The title supplies the reference label but never enters the decision view. The output is *role-retrieval confidence*, not a probability of hire, performance, or superiority to another candidate.

Target validity is the first design issue. Language-model screening can react to demographic markers, culturally associated terms, and small wording changes (Tan et al., 2026; Wilson & Caliskan, 2024), while internal labels can reward template recognition instead of occupational reasoning. Construct-validity and human-comparison studies caution against inferring hiring validity from a generic NLP score (Castleman et al., 2026; Vaishampayan et al., 2025; Zhou et al., 2025). We therefore exclude role-bearing and identity-bearing fields, check exact duplicates before splitting, and withhold the heterogeneous records from model selection.

The second issue is uncertainty. Calibration links scores to observed correctness, while selective classification sends low-confidence profiles to review. Temperature scaling corrects score sharpness on held-out data (Guo et al., 2017), and selective prediction exposes the coverage–risk trade-off (El-Yaniv & Wiener, 2010). Abstention still needs auditing because group-dependent confidence can concentrate review burden (Jones et al., 2020).

The third issue is explanation fidelity. Local explanations aid inspection (Lundberg & Lee, 2017; Ribeiro et al., 2016), while model cards structure purpose and limitations (Mitchell et al., 2019). Here, a deterministic card holds scores, terms, and exact spans; the LLM verbalizes only that packet.

The study compares strict-view models, calibration, selective review, bias probes, heterogeneous transfer, and grounded LLM cards. Figure 1 summarizes the workflow.

International Journal of Instruction, Technology, and Social Sciences
Vol. 5, 86-105, 2026
ISSN: 2716-6546
www.ijitsc.net

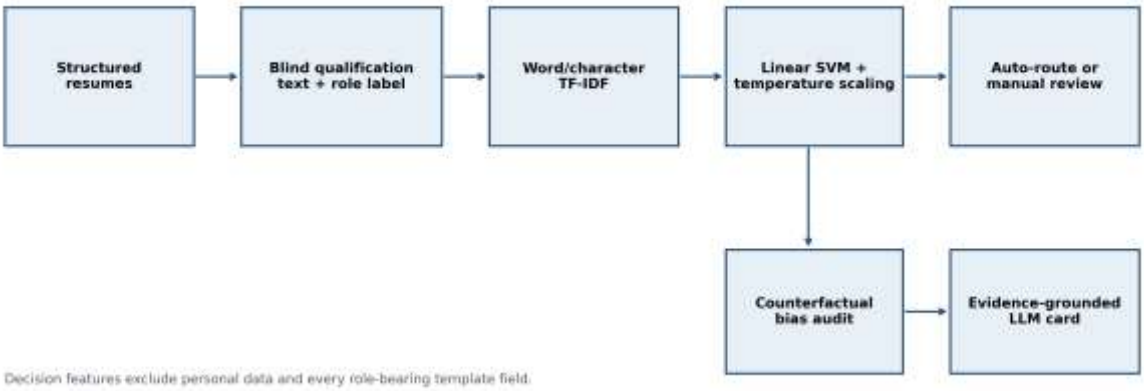

Figure 1. Leakage-controlled role retrieval, calibration, selective review, counterfactual audit, and evidence-grounded LLM-card workflow.

Literature Review

Algorithmic hiring, validity, and bias

Institutions define hiring labels, screening stages, and acceptable error. Bias-mitigation claims often rest on limited training-data access, job analysis, or independent audits (Raghavan et al., 2020), and fairness cannot be reduced to one parity statistic (Fabris et al., 2025). The NIST AI Risk Management Framework likewise connects validity, reliability, transparency, accountability, and harmful bias (Tabassi, 2023). We therefore evaluate target validity, prediction, calibration, counterfactual sensitivity, and transfer separately.

Identity-bearing fields require special treatment. Correspondence testing found different callbacks after changing racially associated names on comparable resumes (Bertrand & Mullainathan, 2004). Language-model rankings can preserve gender, race, and intersectional bias, and JobFair formalizes related benchmark procedures (Wang et al., 2024; Wilson & Caliskan, 2024). Sociocultural markers may remain influential after overt identifiers are removed (Tan et al., 2026). Name swaps are therefore sensitivity probes, not demographic identification or estimates of real-world discrimination.

Automation changes decision-making beyond the model's direct predictions. Wilson et al. (2025) show that biased LLM recommendations can shift human hiring judgments and limit autonomy. At system level, shared models and datasets can create an algorithmic monoculture in which employers make correlated errors and candidates face the same failure repeatedly (Bommasani et al., 2026). These findings favor review interfaces that disclose uncertainty and alternatives instead of providing a single fluent recommendation. They also support the present decision to describe the output as internal role triage, not a recommendation to interview, reject, or hire.

Validity evidence must match the claimed use. Castleman et al. (2026) examine how LLM resume-screening validity is measured, emphasizing the distance between benchmark tasks and employment constructs. Vaishampayan et al. (2025) compare human and LLM resume matching directly and show that observational agreement patterns offer information unavailable from accuracy alone. The current corpus lacks paired jobs and human judgments, so it cannot reproduce those designs. It supports a narrower experiment on whether qualifications recover an internally assigned technical role and whether that retrieval process is calibrated and auditable.

Representation and calibrated retrieval

Sentence-BERT demonstrates efficient semantic search with paired sentence embeddings (Reimers & Gurevych, 2019), but pretrained representations add provenance and bias questions. We use sparse word and character TF–IDF to isolate in-corpus evidence, comparing skill overlap, role centroids, a probabilistic bag-of-terms classifier, logistic regression, and a linear support vector machine.

Classification accuracy and score calibration answer different questions (Xin, 2025). Guo et al. (2017) show that temperature scaling can substantially improve the reliability of modern classifier confidence while preserving class order. Kull et al. (2019) extend calibration beyond a single temperature through multiclass Dirichlet transformations, underscoring that calibration methods have different flexibility and overfitting risks. A scalar temperature is appropriate here because the calibration set contains only 20 profiles per role; higher-dimensional calibration would estimate many more parameters from limited class-wise data. Expected calibration error, negative log-likelihood, and multiclass Brier score accompany accuracy and macro-F1.

Overall calibration can conceal subgroup deviations. Multicalibration covers computationally identifiable subsets (Hébert-Johnson et al., 2018), while equality of opportunity requires outcome-conditioned protected-group comparisons (Hardt et al., 2016). Neither criterion can be established without hiring outcomes and verified protected labels. We instead report operational slices and controlled perturbations, without equating input invariance with demographic parity.

Selective review and accountable explanations

Selective classification joins a predictor with a selection function that accepts some cases and abstains on the rest. El-Yaniv and Wiener (2010) formalize the coverage–risk relationship, and Geifman and El-Yaniv (2017) show how confidence-based selection can control error in modern classifiers. In resume triage, abstention translates into human review rather than a missing output. The operational question is therefore not whether the model classifies every profile, but which coverage level produces a defensible accepted-case error and review workload.

Confidence ranking alone does not make abstention fair. Jones et al. (2020) show that selective classification can magnify group disparities, and Lee et al. (2021) develop sufficiency-based methods for fair selective classification. The present analysis reports review coverage across remote preference,

seniority, and employment type because those fields exist in the standardized cohort. These are operational descriptors, not substitutes for race, gender, disability, or age. The audit consequently identifies where review burden differs inside the corpus but makes no protected-class fairness certification.

Explanations also require scope discipline. LIME explains a prediction through a local surrogate (Ribeiro et al., 2016), and SHAP unifies additive feature-attribution approaches (Lundberg & Lee, 2017). Both motivate instance-level evidence, but a production-facing explanation also needs traceability to the source (Jin, 2025a, 2025b). Model cards document model purpose and limitations (Mitchell et al., 2019); datasheets document dataset composition, collection, and appropriate use (Gebru et al., 2021). Our explanation card combines these accountability ideas at the decision level: it states the role taxonomy, calibrated score, review route, runner-up, supporting and missing terms, exact source spans, and a safety note about excluded fields.

Methodology

Research questions and task definition

The experiment answers four questions. RQ1 asks how well leakage-controlled qualification text retrieves 46 internal technical-role labels. RQ2 asks whether held-out temperature scaling improves the interpretation of the top-role score. RQ3 asks how selective review changes coverage, accepted-case accuracy, and error capture. RQ4 asks whether the system remains stable under identity-cue perturbations, operational subgroup slicing, and transfer to heterogeneous records, and whether every explanation-card claim remains grounded.

The reference role was the first experience title that mapped to the 46-role taxonomy. The input contained job-relevant qualification fields but not the title itself. Because no job description or hiring outcome exists, the model estimates P(reference role | qualification text). It does not estimate person–vacancy compatibility. The term *screening* denotes routing within this proxy role index.

Data provenance and audit

The dataset card attributes the collection to 2023, and repository history records continued file activity during 2025; the analyzed 2026 snapshot contains 4,817 parseable JSON records (datasetmaster, 2023). The card describes a merged real-and-synthetic collection normalized to one schema. Empirical inspection identified a first block of 217 heterogeneous resumes and a second block of 4,600 standardized resumes. The second block contains exactly 100 profiles for each of 46 technical roles. The standardized cohort is the primary benchmark because its complete balanced taxonomy supports stratified comparison; the heterogeneous block is reserved for transfer testing.

The audit in Table 1 found no invalid JSON, empty standardized qualification text, conflicting label duplicates, or same-label exact duplicates after blinding. Hashes were computed from whitespace-normalized, case-folded decision text. The split assigned 60 profiles per role to training and 20 each to calibration and test, producing 2,760/920/920 records with seed 2026. Model and vocabulary fitting used training data only; hyperparameter and temperature selection used calibration data only; the test split remained untouched until final evaluation.

Table 1. Dataset audit, cohort boundary, and stratified split.

Audit item	Count
Raw JSON records	4817
Standardized cohort	4600
Heterogeneous cohort	217
Mapped standardized profiles	4600
Unmapped or empty standardized profiles	0
Conflicting duplicate rows	0
Same-role duplicates removed	0
Analysis set	4600
Training split	2760
Calibration split	920
Test split	920

The 46 labels were grouped into seven interpretable families for error analysis. Table 2 shows that Software Development is largest because it contains 18 labels, while every individual label remains balanced. Median decision-text length stays between 101.5 and 105.5 words across families. Figure 2 visualizes family composition across the three splits.

Table 2. Standardized analysis set by role family.

Role family	Resumes	Role labels	Median words
Architecture	200	2	101.5000
Cloud and Platform	700	7	102
Data and AI	800	8	103
Database	400	4	102.5000
QA and Embedded	300	3	105.5000
Security	400	4	102
Software Development	1800	18	103

International Journal of Instruction, Technology, and Social Sciences
Vol. 5, 86-105, 2026
ISSN: 2716-6546
www.ijitsc.net

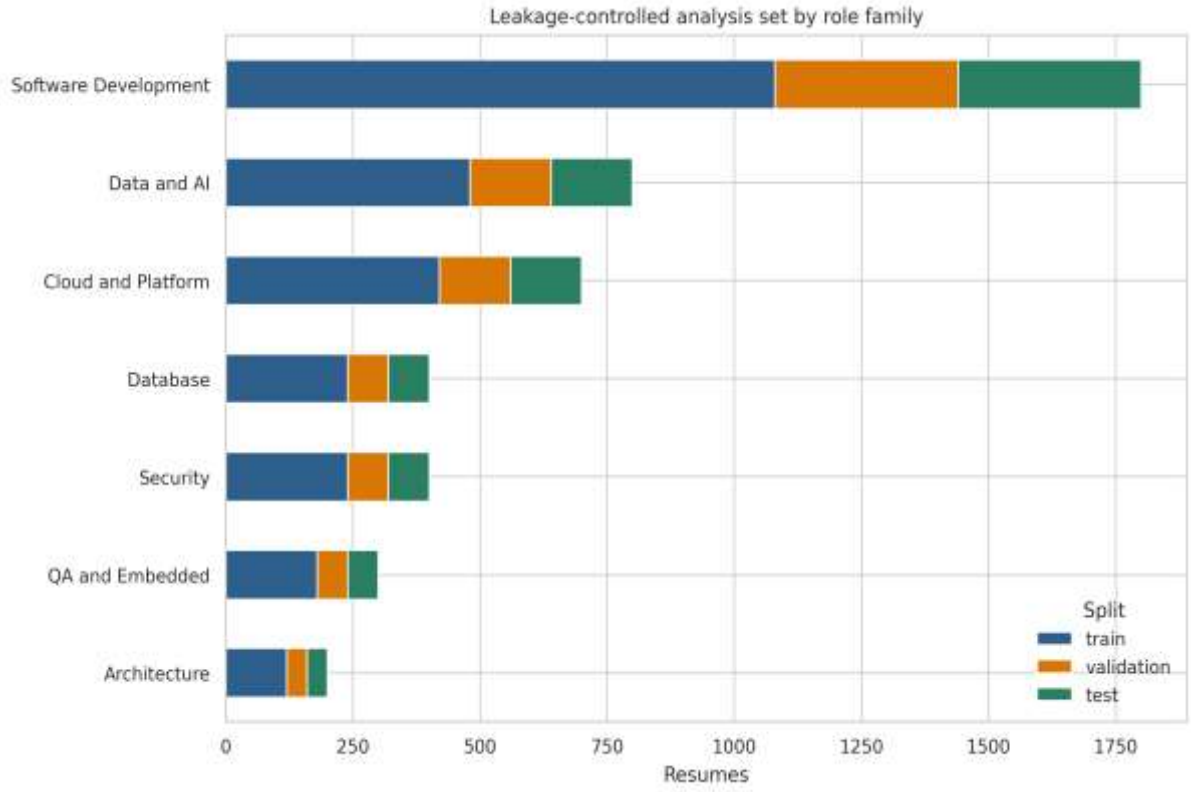

Figure 2. Standardized cohort by role family and train, calibration, and test split.

Leakage-controlled representation

The target title appeared in several structured fields, so feature construction followed a conservative allowlist. Decision text retained experience responsibilities, technical environments, industry, generic degree level and field, honors, relevant coursework, skills, project technology lists, certifications, and publications. It excluded names, email addresses, phone numbers, links, locations, company names, institution names, dates, profile summaries, every job title, education majors, and project names, roles, descriptions, and impacts. The education-major exclusion was essential because that field reproduced the target label in the standardized block. Summary and project prose also contained direct or repeated role templates. Their removal converts the experiment from label recovery to qualification-based retrieval.

An ablation quantified leakage rather than silently discarding it. The strict representation is the only representation used for the main results, calibration, review threshold, subgroup analysis, counterfactual audit, and explanation cards. Broader representations were evaluated separately to show how direct role-bearing fields inflate apparent accuracy.

Models, calibration, and evaluation

The word vectorizer used lowercase, Unicode-normalized TF–IDF with word unigrams and bigrams, sublinear term frequency, minimum document frequency 2, maximum document frequency .98, L2 normalization, and at most 40,000 features. Linear models additionally used a 30,000-feature character-within-word vectorizer with 3–5 character grams and minimum document frequency 3. Table 3 records the search spaces and fixed choices. A skill Jaccard baseline compared binary active terms with each training-role centroid vocabulary. A TF–IDF centroid model used cosine similarity. Multinomial naive Bayes selected smoothing by calibration negative log-likelihood. Multinomial logistic regression selected inverse regularization strength by calibration negative log-likelihood. A class-balanced linear support vector machine selected C by calibration accuracy.

Table 3. Experimental configuration and selection rules.

Component	Setting
Word TF-IDF	1–2 grams; min_df=2; max_df=.98; 40,000 features
Character TF-IDF	3–5 char_wb grams; min_df=3; 30,000 features
Logistic regression	C=4.0; class-balanced; L-BFGS
Linear SVM	C=0.25; class-balanced
Calibration	Held-out temperature scaling (T=0.143196)
Selective review	threshold=0.087092
Randomness	seed=2026

The selected settings in Table 3 were smoothing α=.1, logistic C=4, SVM C=.25, and temperature T=.143196. The final 80%-coverage confidence threshold was .087092.

For an SVM score vector z, the temperature-scaled probability for class k is $\exp(z_k/T)$ divided by the sum of $\exp(z_j/T)$ over all classes j. The scalar T was chosen solely by calibration-set multiclass negative log-likelihood. The evaluation reports accuracy, macro precision, macro recall, macro-F1, top-3 accuracy, mean reciprocal rank (MRR), negative log-likelihood (NLL), multiclass Brier score, and 15-bin expected calibration error (ECE). Accuracy, macro-F1, and ECE uncertainty used 1,000 bootstrap resamples of the test set. An exact paired McNemar test compared the final SVM with logistic regression. These metrics separate top-choice discrimination, ranked retrieval, class balance, and probability quality.

Selective thresholds were calibration-set quantiles for target automatic coverage levels of 90%, 80%, 70%, 60%, and 50%. At each fixed threshold, the test analysis measured observed coverage, accuracy and macro-F1 among accepted profiles, review rate, and the fraction of all errors sent to review. The

risk–coverage curve orders test cases by confidence and plots cumulative error among accepted cases; area under that curve summarizes selection quality. A threshold is therefore a workload decision, not a new classifier.

Counterfactual, subgroup, transfer, and explanation audits

The counterfactual audit compared the blind view with a parallel, validation-calibrated C=.5 sensitive-fields SVM that also received conventional parser fields. Following correspondence-testing logic (Bertrand & Mullainathan, 2004), four female/male-associated name-and-pronoun pairs, four White/Black-associated name pairs, and one younger/older age-and-graduation cue pair were inserted into otherwise identical profiles. The audit measured signed and absolute reference-role score change, top-role flips, and review-route flips. These cues probe textual sensitivity; they are not verified demographic identities.

Operational slices used remote preference, seniority, and employment type, retaining only subgroups with at least ten test profiles. The analysis reports subgroup accuracy, macro-F1, ECE, and automatic-route coverage. It does not calculate equal opportunity because no outcome-conditioned protected attribute exists. The heterogeneous stress test mapped every eligible record in the 217-record block to the same taxonomy, removed duplicate decision texts, and applied the frozen model without adaptation.

A supplementary out-of-domain gate covered all 217 heterogeneous records, including records whose roles were outside the 46-class taxonomy. Its support score was the maximum cosine similarity to any training-role word-TF–IDF centroid. The support threshold was the fifth percentile of calibration support; a separate validation-calibrated C=2 SVM supplied the confidence gate. Core test profiles represented in-domain cases and heterogeneous profiles represented shifted cases for AUROC. Support-only and confidence-plus-support rules were evaluated alongside confidence alone.

Finally, every test profile received one structured explanation card containing the predicted role, calibrated score, route, runner-up, margin, three exact evidence snippets, present and missing role signals, and an identity-field safety note. Snippets were ranked by overlap with high-weight word features and filled from source evidence. Validation checked schema, taxonomy, score range, three-span coverage, and exact source membership. For an actual language-model realization, one median-confidence card per reference role was sent to a GPT-5-based Codex language model with a fixed one-to-two-sentence prompt. The model could use only card fields, had to qualify missing signals as “not observed in the retained fields,” and could make no hiring recommendation or absolute qualification claim. Its summaries had no effect on matching scores or routes; the prompt, 46 outputs, and validation records were retained with the run.

International Journal of Instruction, Technology, and Social Sciences
Vol. 5, 86-105, 2026
ISSN: 2716-6546
www.ijitsc.net

Results and Discussion

Leakage audit and strict model comparison

The leakage ablation in Table 4 establishes which performance result is valid. With strict qualification fields, the word-only logistic model achieved 0.2348 top-1 accuracy, 0.3913 top-3 accuracy, and 0.2271 macro-F1. Restoring project descriptions raised the same model to 0.9076 accuracy and 0.9071 macro-F1. Restoring summaries, titles, and other direct role-bearing text produced 1.0000 on every discrimination and ranking metric. The increase is not evidence that the richer model understood candidate suitability: project descriptions repeat role templates, and direct fields reveal the reference title. This ablation explains why field provenance matters more than a superficially strong benchmark score.

Table 4. Role-bearing feature leakage ablation.

Feature view	Temperature	Top-1 accuracy	Top-3 accuracy	Macro F1	NLL	ECE
Strict qualification fields	0.7261	0.2348	0.3913	0.2271	3.0183	0.0477
+ project descriptions	0.3432	0.9076	0.9630	0.9071	0.2502	0.0173
+ direct role-bearing text	0.0063	1.0000	1.0000	1.0000	2.22e-16	0.0000

Under the strict view, Table 5 and Figure 3 show a difficult 46-class task. Skill Jaccard obtained 0.0598 accuracy, the TF–IDF centroid 0.1989, multinomial naive Bayes 0.1511, and logistic regression 0.2413. The temperature-scaled SVM reached the highest accuracy (0.2489), top-3 accuracy (0.3978), MRR (0.3663), and lowest NLL (2.9329), while logistic regression had a slightly higher macro-F1 (0.2364 versus 0.2325). The exact paired test counted 24 cases correct only for SVM and 17 correct only for logistic regression (p=.3489); their accuracy difference is not statistically resolved. Bootstrap intervals were 0.2228–0.2783 for accuracy and 0.2139–0.2486 for macro-F1. These results demonstrate limited role recoverability after direct role signals are removed.

Table 5. Strict-view held-out model comparison.

Model	Accuracy	Macro F1	Top-3 accuracy	MRR	NLL	ECE
Skill Jaccard	0.0598	0.0396	0.1196	0.1408	3.8062	0.0266
TF-IDF centroid	0.1989	0.1754	0.3174	0.3049	3.2632	0.0647
Multinomial NB	0.1511	0.1596	0.2609	0.2543	3.5437	0.0807
Logistic regression	0.2413	0.2364	0.3859	0.3604	3.0172	0.0551
Linear SVM + temperature	0.2489	0.2325	0.3978	0.3663	2.9329	0.0401

International Journal of Instruction, Technology, and Social Sciences
Vol. 5, 86-105, 2026
ISSN: 2716-6546
www.ijitsc.net

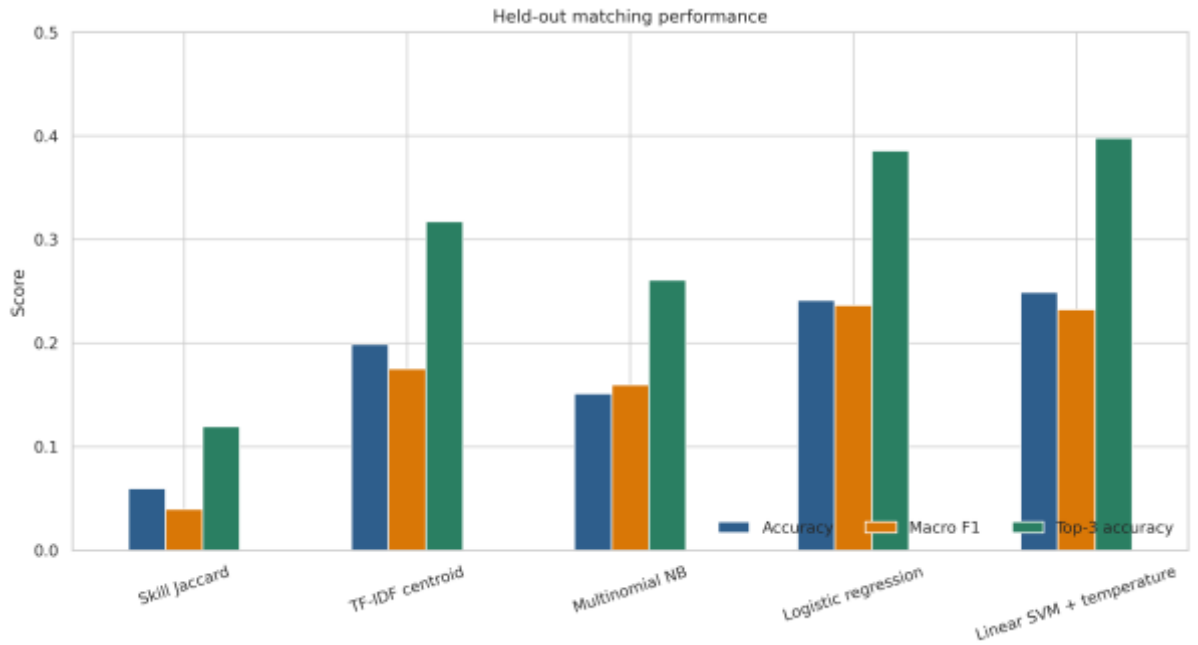

Figure 3. Accuracy, macro-F1, and top-3 accuracy for five strict-view models on 920 test profiles.

Calibration and selective review

Temperature scaling changed score interpretation without changing SVM predictions. Table 6 reports that T=.143196 reduced NLL from 3.5529 to 2.9329, Brier score from 0.9580 to 0.8252, and ECE from 0.2121 to 0.0401. Sigmoid calibration obtained lower NLL (2.8179) but worse ECE (0.0945), Brier score (0.8308), and macro-F1 (0.2106). Figure 4 confirms that temperature scaling moves confidence toward observed accuracy. Bootstrap ECE had median 0.0467 and a 95% interval of 0.0309–0.0657, so calibration remains imperfect.

Table 6. Calibration ablation for the selected linear SVM.

Model	Accuracy	Macro F1	NLL	Brier	ECE
SVM raw softmax	0.2489	0.2325	3.5529	0.9580	0.2121
SVM temperature (T=0.143)	0.2489	0.2325	2.9329	0.8252	0.0401
SVM sigmoid	0.2478	0.2106	2.8179	0.8308	0.0945

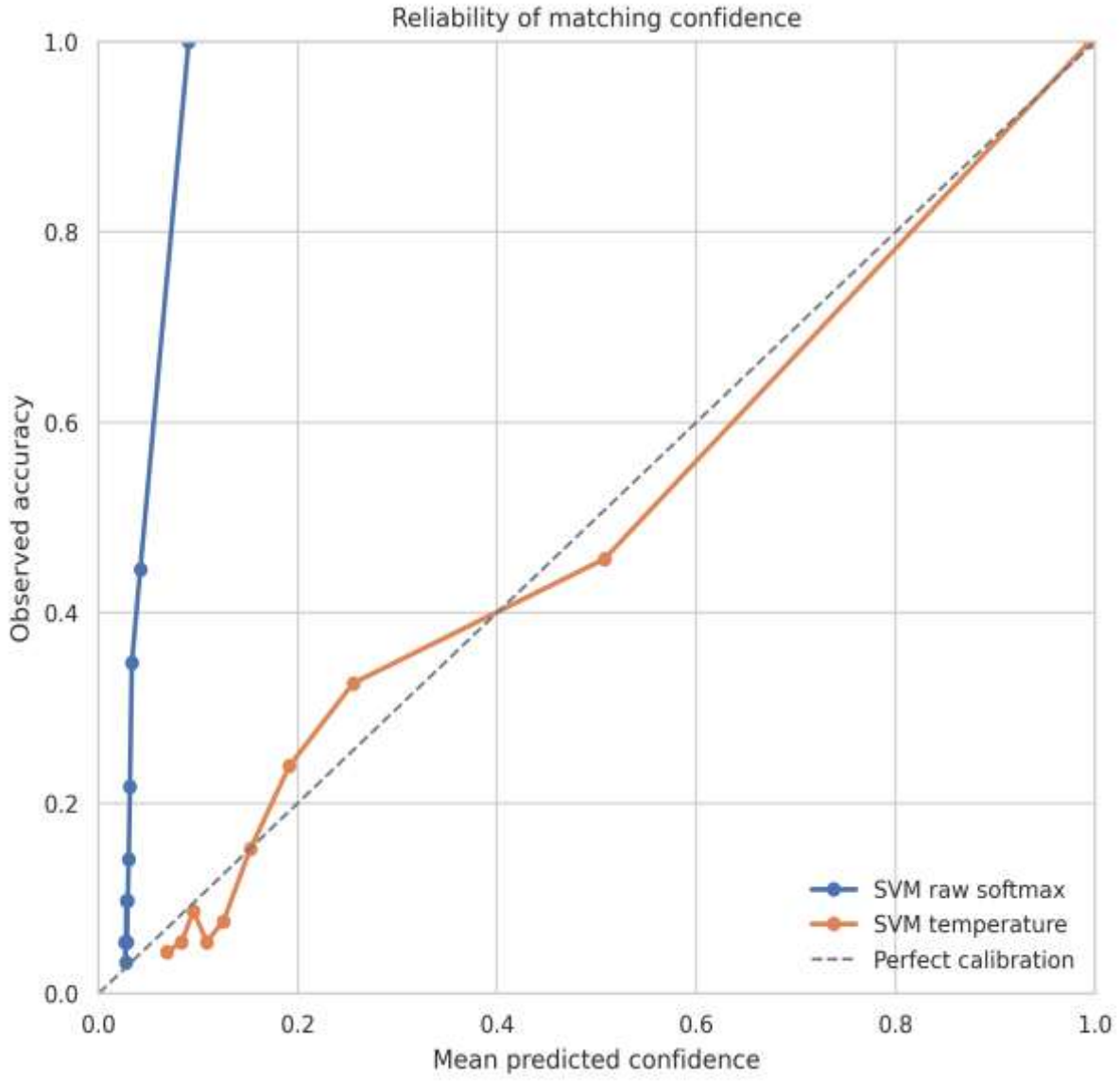

Figure 4. Reliability with equal-frequency bins before and after held-out temperature scaling.

Table 7 quantifies the automation–error trade-off. The threshold selected for 80% calibration coverage produced 0.8130 test coverage, 0.2955 accepted-case accuracy, 0.1870 review rate, and capture of 23.73% of all errors. At the 50% target, observed coverage was 0.5293, accepted accuracy rose to 0.4148, and 58.76% of errors were reviewed. Macro-F1 rose only from 0.2325 at full coverage to 0.2533 at the most selective setting, showing that confidence concentrates on some roles rather than improving all 46 uniformly. The risk–coverage curves in Figure 5 and SVM AURC of 0.4855 confirm useful but limited error ranking. Selective review reduces accepted-case risk; it does not make the underlying model reliable enough for autonomous decisions.

International Journal of Instruction, Technology, and Social Sciences
Vol. 5, 86-105, 2026
ISSN: 2716-6546
www.ijitsc.net

Table 7. Selective-review performance at validation-set coverage targets.

Target coverage	Threshold	Observed coverage	Selective accuracy	Review rate	Errors sent to review
1.0000	0.0000	1.0000	0.2489	0.0000	0.0000
0.9000	0.0774	0.9011	0.2714	0.0989	0.1259
0.8000	0.0871	0.8130	0.2955	0.1870	0.2373
0.7000	0.0998	0.7185	0.3222	0.2815	0.3517
0.6000	0.1136	0.6130	0.3670	0.3870	0.4834
0.5000	0.1302	0.5293	0.4148	0.4707	0.5876

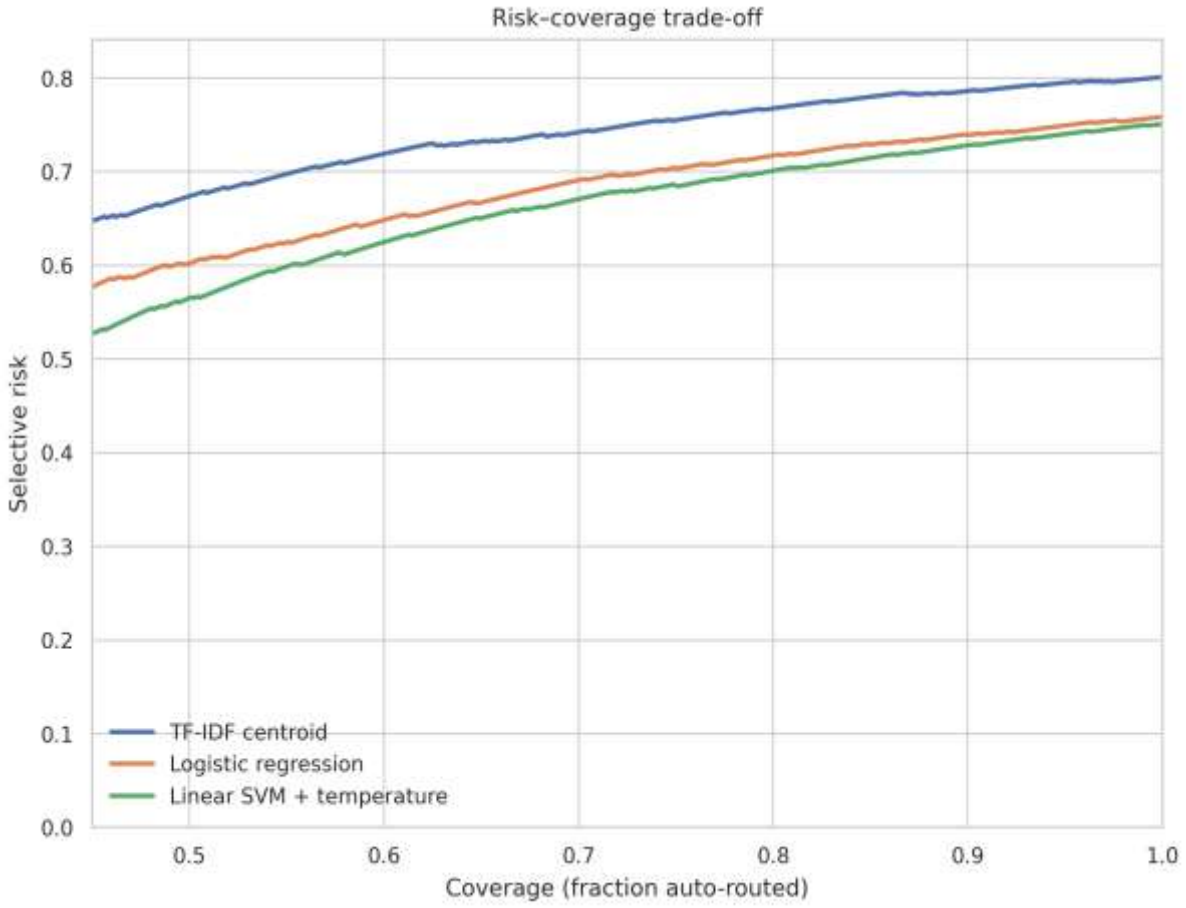

Figure 5. Selective risk as automatic coverage decreases for three strict-view models.

Family, subgroup, and counterfactual results

Table 8 reveals pronounced family variation. Accuracy was 0.4875 for Database and 0.4111 for Software Development, but 0 for Data and AI, 0.0250 for Architecture, 0.0625 for Security, and 0.1071 for Cloud and Platform. QA and Embedded reached 0.3500. This concentration explains the modest macro-F1 and warns against interpreting balanced class counts as balanced difficulty.

Table 8. Held-out performance by role family.

Role family	Test n	Accuracy	Macro F1	Mean confidence	ECE
Architecture	40	0.0250	0.0041	0.1185	0.0935
Cloud and Platform	140	0.1071	0.0263	0.1419	0.0462
Data and AI	160	0.0000	0.0000	0.1223	0.1223
Database	80	0.4875	0.4826	0.5744	0.1482
QA and Embedded	60	0.3500	0.0548	0.1771	0.1729
Security	80	0.0625	0.0142	0.1168	0.0715
Software Development	360	0.4111	0.1822	0.3546	0.0647

The operational slices in Table 9 do not remove that concern. Remote-preference accuracy ranged narrowly from 0.2444 to 0.2526 and automatic coverage from 0.7901 to 0.8386. Seniority accuracy ranged more widely, from 0.2063 for junior to 0.2866 for mid-level profiles; employment-type accuracy ranged from 0.2298 to 0.2651. ECE across all reported slices lay between 0.0357 and 0.0661. These are descriptive corpus slices, not protected-group fairness measures.

Table 9. Operational subgroup performance and automatic-route coverage.

Attribute	Subgroup	Test n	Accuracy	Macro F1	ECE	Automatic-route coverage
Remote Preference	hybrid	324	0.2500	0.2211	0.0392	0.7901
Remote Preference	onsite	311	0.2444	0.2249	0.0420	0.8135
Remote Preference	remote	285	0.2526	0.2392	0.0661	0.8386
Seniority	junior	286	0.2063	0.1883	0.0624	0.8042
Seniority	mid	314	0.2866	0.2435	0.0569	0.8312
Seniority	senior	320	0.2500	0.2405	0.0357	0.8031
Employment Type	contract	298	0.2651	0.2408	0.0627	0.7953
Employment Type	full-time	300	0.2533	0.2312	0.0622	0.8067
Employment Type	internship	322	0.2298	0.2175	0.0439	0.8354

Table 10 and Figure 6 show why feature blinding is operationally important. In the sensitive-fields model, gender-associated swaps changed the reference-role score by 0.0047 on average and flipped the top role in 14.76% of 3,680 comparisons; race-associated swaps yielded 0.0032 mean absolute change and 10.43% top-role flips. The age-cue audit produced 0.0020 mean change and 7.72% flips. Review-route flip rates were 10.54%, 7.66%, and 3.80%, respectively. The blind model was exactly invariant because all inserted cues were excluded by construction. That zero result proves removal of the tested text pathway, not absence of demographic discrimination through unmeasured proxies.

Table 10. Counterfactual score and routing sensitivity.

Model	Audit	Mean absolute shift	95th percentile shift	Top-role flip rate	Review-route flip rate
Sensitive-fields model	Gender-associated name	0.0047	0.0176	0.1476	0.1054
Sensitive-fields model	Race-associated name	0.0032	0.0121	0.1043	0.0766
Sensitive-fields model	Age cue	0.0020	0.0068	0.0772	0.0380
Blind model	Gender-associated name	0.0000	0.0000	0.0000	0.0000
Blind model	Race-associated name	0.0000	0.0000	0.0000	0.0000
Blind model	Age cue	0.0000	0.0000	0.0000	0.0000

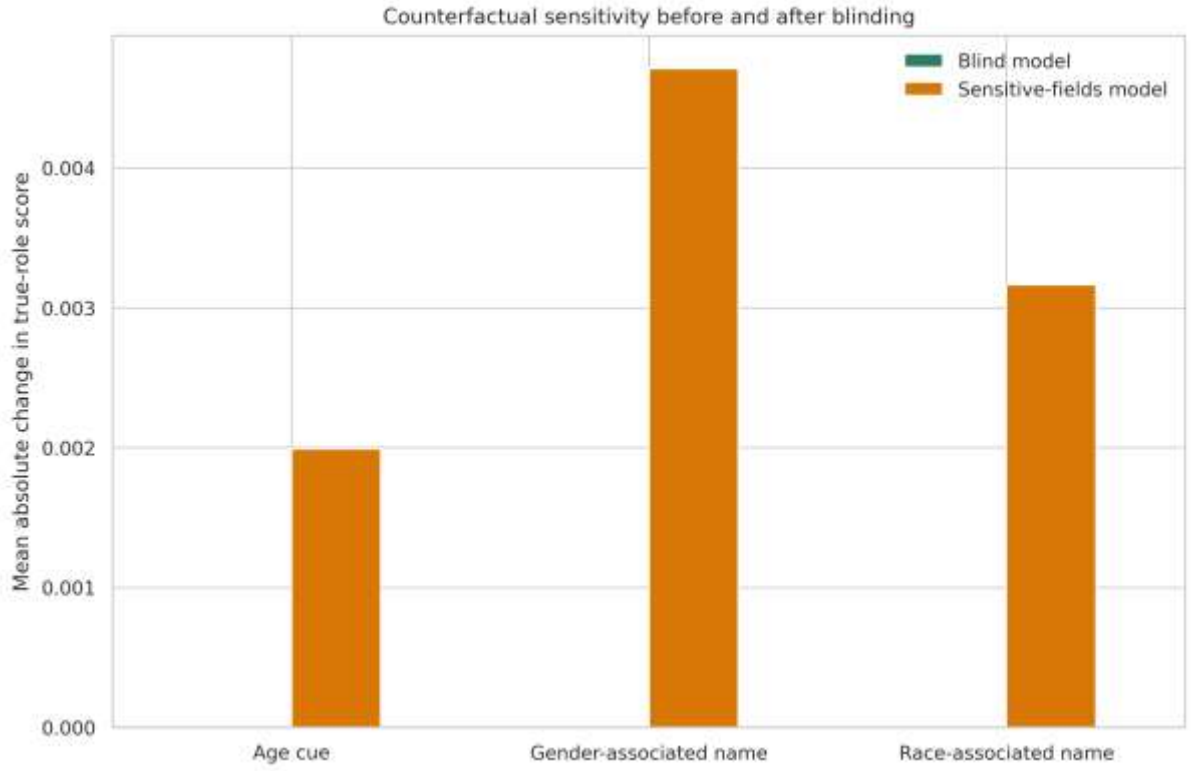

Figure 6. Mean absolute reference-role score change under controlled name and age-cue swaps.

Transfer, support gating, and explanation fidelity

Transfer results in Table 11 were weaker than core results. Only 104 heterogeneous records mapped to the taxonomy after within-pool deduplication; overall accuracy was 0.1346, top-3 accuracy 0.3173, MRR 0.2733, and mean confidence 0.4390. Java Developer, which supplied 53 profiles, reached 0.2453 accuracy, while four of the six represented roles had zero top-1 accuracy. The mismatch between low transfer accuracy and substantial confidence shows that calibration inside the standardized schema does not survive distribution shift.

Table 11. Frozen-model transfer to mapped heterogeneous profiles.

Role	N	Accuracy	Top-3 accuracy	MRR	Mean confidence
Data Scientist	3	0.0000	0.0000	0.0690	0.1113
Database Administrator	1	1.0000	1.0000	1.0000	0.4389
DevOps Engineer	24	0.0000	0.0000	0.0654	0.5276
Java Developer	53	0.2453	0.5283	0.4303	0.5384
Python Developer	22	0.0000	0.1818	0.1271	0.1654
Software Engineer	1	0.0000	0.0000	0.0455	0.0478
Overall	104	0.1346	0.3173	0.2733	0.4390

The support gate in Table 12 addresses that failure. Confidence alone reviewed only 11.52% of all 217 heterogeneous records and had OOD AUROC 0.3366. Maximum role-centroid support achieved OOD AUROC 0.9968; its support-only rule retained 94.35% of core test profiles while reviewing 99.08% of heterogeneous profiles. Combining support and confidence retained 77.28% of core profiles, achieved 0.2954 accepted accuracy, and still reviewed 99.08% of heterogeneous profiles. Support detects schema mismatch; confidence chiefly ranks ambiguity within the training schema. Both checks are therefore necessary for a bounded triage service.

Table 12. Confidence and lexical-support gates under distribution shift.

Gate	Core coverage	Core accepted accuracy	Core review rate	Heterogeneous review rate	Support OOD AUROC	Confidence OOD AUROC
Confidence only	0.8217	0.2897	0.1783	0.1152	0.9968	0.3366
Support only	0.9435	0.2512	0.0565	0.9908	0.9968	0.3366
Confidence + support	0.7728	0.2954	0.2272	0.9908	0.9968	0.3366

Table 13 reports 920 structured cards, all with complete schemas, valid roles, in-range scores, three evidence spans, and exact source fidelity. Their review rate was 18.70% and mean top-versus-runner-up margin was 0.1773. Table 14 audits the 46 LLM verbalizations. Every output named the predicted role and route, used the “not observed” qualifier, avoided hiring recommendations and absolute qualification claims, and reproduced at least one source span verbatim. No summary reproduced all three spans because the one-to-two-sentence format intentionally selected one or two. Thirty-six of the 46 sampled predictions were wrong, making the grounding result nontrivial: the summaries stayed faithful even when the classifier failed. Fidelity does not establish correctness or human usefulness, and no recruiter study was conducted.

Table 13. Structured explanation-card validation.

Measure	Value
Cards generated	920
Schema completeness	1.0000
Evidence fidelity	1.0000
Valid role rate	1.0000
Score validity	1.0000
Three-span coverage	1.0000
Manual-review cards	0.1870
Mean card margin	0.1773

Table 14. Audit of 46 evidence-bound LLM card realizations.

Measure	Value
LLM cards audited	46
Schema Complete	1.0000
Role Mentioned	1.0000
Decision Mentioned	1.0000
At Least One Verbatim Span	1.0000
All Spans Verbatim	0.0000
Not Observed Qualifier	1.0000
No Hiring Recommendation	1.0000
No Absolute Qualification Claim	1.0000

Conclusions

This study built and executed a leakage-controlled evaluation of calibrated resume role retrieval on 4,817 structured records. Direct role-bearing fields transformed a 0.2348-accuracy strict logistic model into a 1.0000 classifier, confirming that unexamined schema fields can dominate benchmark conclusions. On the valid representation, the final SVM achieved 0.2489 accuracy and 0.0401 ECE. Selective review improved accepted-case accuracy to 0.4148 only by reducing coverage to 0.5293, and heterogeneous accuracy fell to 0.1346. A centroid-support gate detected that shift far better than confidence.

The system's appropriate use is internal, human-supervised role triage. It cannot rank employability, recommend rejection, or certify fairness because the data contain no vacancies, outcomes, or protected-group labels. Counterfactual invariance verifies the removal of tested identity cues, while evidence cards verify source fidelity; neither substitutes for outcome-based validation. Deployment requires a job-

linked dataset, occupationally grounded labels, representative applicant populations, protected-attribute governance, human-factor evaluation, and continuing drift audits. Within the present evidence, calibration, abstention, support gating, and grounded cards make uncertainty visible, but they do not convert a weak proxy task into a hiring decision.

References

[1] Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

[2] Bommasani, R., Bana, S. H., Creel, K. A., Jurafsky, D., & Liang, P. (2026). Algorithmic monocultures in hiring. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3805689.3812400

[3] Castleman, J., Shen, Z., Metevier, B., Springer, M., & Korolova, A. (2026). Measuring validity in LLM-based resume screening [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2602.18550

[4] datasetmaster. (2023). *Advanced resume parser & job matcher resumes* [Data set]. Hugging Face. https://huggingface.co/datasets/datasetmaster/resumes

[5] El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research, 11*(53), 1605–1641. https://www.jmlr.org/papers/v11/el-yaniv10a.html

[6] Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., Zuiderveen Borgesius, F., & Biega, A. J. (2025). Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology, 16*(1), Article 16, 1–54. https://doi.org/10.1145/3696457

[7] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

[8] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems, 30*, 4878–4887. https://papers.neurips.cc/paper/7073-selective-classification-for-deep-neural-networks

[9] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[10] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3315–3323. https://papers.neurips.cc/paper/6374-equality-of-opportunity-in-supervised-learning

[11] Hébert-Johnson, Ú., Kim, M. P., Reingold, O., & Rothblum, G. N. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 1939–1948). PMLR. https://proceedings.mlr.press/v80/hebert-johnson18a.html

[12] Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[13] Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[14] Jones, E., Sagawa, S., Koh, P. W., Kumar, A., & Liang, P. (2020). Selective classification can magnify disparities across groups [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2010.14134

[15] Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. *Advances in Neural Information Processing Systems, 32*. https://proceedings.neurips.cc/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html

[16] Lee, J. K., Bu, Y., Rajan, D., Sattigeri, P., Panda, R., Das, S., & Wornell, G. W. (2021). Fair selective classification via sufficiency. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 6076–6086). PMLR. https://proceedings.mlr.press/v139/lee21b.html

[17] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

[18] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596

[19] Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). ACM. https://doi.org/10.1145/3351095.3372828

[20] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[21] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

[22] Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[23] Tan, B. C. Z., Khoo, S., Doan, B. N., Liu, Z., Chen, N. F., & Lee, R. K.-W. (2026). Small changes, big impact: Demographic bias in LLM-based hiring through subtle sociocultural markers in anonymised resumes [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2603.05189

[24] Vaishampayan, S., Leary, H., Alebachew, Y. B., Hickman, L., Stevenor, B., Beck, W., & Brown, C. (2025). Human and LLM-based resume matching: An observational study. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 4823–4838). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.270

[25] Wang, Z., Wu, Z., Guan, X., Thaler, M., Koshiyama, A., Lu, S., Beepath, S., Ertekin, E., & Perez-Ortiz, M. (2024). JobFair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 3227–3246). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.184

[26] Wilson, K., & Caliskan, A. (2024). Gender, race, and intersectional bias in resume screening via language model retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7*(1), 1578–1590. https://doi.org/10.1609/aies.v7i1.31748

[27] Wilson, K., Sim, M., Gueorguieva, A.-M., & Caliskan, A. (2025). No thoughts just AI: Biased LLM hiring recommendations alter human decision making and limit human autonomy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(3), 2692–2704. https://doi.org/10.1609/aies.v8i3.36749

[28] Xin, Q. (2025). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[29] Zhou, B., Jin, J., & Zhao, D. (2025). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529