COUNTERFACTUAL BIAS TESTING FOR APPLICATION TRACKING SYSTEMS

Sai Yashwant
AI Product & Platform Head
ManpowerGroup Services India Pvt. Ltd.
sai.yashwant@manpowergroup.com

Shruti Bansal
Senior Data Scientist
ManpowerGroup Services India Pvt. Ltd.
shruti.bansal@manpowergroup.com

Anurag Dubey
Data Scientist
ManpowerGroup Services India Pvt. Ltd.
anurag.dubey@manpowergroup.com

Samaroha Chatterjee
Data Scientist
ManpowerGroup Services India Pvt. Ltd.
samaroha.chatterjee@manpowergroup.com

Satyam Kumar
Data Scientist
ManpowerGroup Services India Pvt. Ltd.
satyam.kumar@manpowergroup.com

Shreyash Gupta
Data Scientist
ManpowerGroup Services India Pvt. Ltd.
shreyash.gupta@manpowergroup.com

Gantala Thulsiram
Assistant Professor
Indian Institute of Technology, Hyderabad
thulsiramg@mae.iith.ac.in

August 28, 2026

ABSTRACT

Automated candidate–job matching systems are increasingly classified as high-risk AI under emerging regulation, yet auditing them for demographic bias is expensive: classical correspondence-audit studies require hand-crafted resume pairs and manual submission, which does not scale to the rate at which enterprise matching pipelines are retrained and redeployed. This paper presents a general, reusable methodology that (1) uses a chain of task-specialized large-language-model (LLM) agents to synthesize identity-neutral base resumes and inject controlled demographic treatments across five protected-characteristic axes (sex/gender, age, place of residence, language, disability signal), producing a correspondence-audit-style $K \times (1 + N)$ matrix of base candidates and bias variants; (2) qualitatively flags inferred protected characteristics and bias per an EU AI Act-aligned prompt; (3) ranks every candidate against a job description using a fine-tuned sentence-embedding model and cosine similarity, representative of the semantic ranking core used by modern candidate–job matching systems; and (4) computes a nine-metric quantitative fairness suite – spanning counterfactual (score delta, mean absolute rank change, flip rate), group-fairness (top-K retention, four-fifths rule or impact ratio), and merit-aware (Recall@K, nDCG@K, equal opportunity, equalized odds) families – each with bootstrap confidence intervals, exact/asymptotic significance tests (Wilcoxon signed-rank, McNemar, Fisher exact, Wilson score), and Benjamini–Hochberg false-discovery-rate correction, culminating in an automatically generated PASS/INVESTIGATE/FAIL audit report with a composite risk score. We illustrate the methodology on an example correspondence-audit corpus spanning 5 job orders, 100 base candidates, and 10 demographic-bias treatments (90 metric$\times$variant evaluations),

illustrating that while variant-level score shifts, top-K retention, and merit-aware true-positive-rate/false-positive-rate gaps remain within tolerance for every treatment, a rank-stability metric (mean absolute rank change) and a ranking-quality metric (nDCG@K) each surface borderline findings – including one on the neutral baseline configuration itself – that a score- or retention-only view would have missed. The results argue for multi-metric, multi-family auditing over any single aggregate fairness score, and for treating LLM-agent-generated correspondence audits as a practical, low-cost complement to human-curated audit studies for any candidate–job matching pipeline.

Keywords Algorithmic Hiring Bias, Correspondence Audit, Counterfactual Fairness, Synthetic Data Generation, LLM Agents, Fairness Metrics, EU AI Act

1 Introduction

Automated candidate–job matching, encompassing resume parsing, semantic ranking, and shortlist generation, has moved from an experimental augmentation of recruiter workflow to a load-bearing component of high-volume staffing pipelines, with contextual-transformer encoders [Devlin et al., 2019] and their fine-tuned derivatives [Zhang et al., 2020, Qin et al., 2018] now forming the semantic core of many such systems. Under the European Union's Artificial Intelligence Act, AI systems used "to evaluate candidates... in the recruitment or selection of natural persons" are explicitly enumerated as high-risk [European Parliament and Council of the European Union, 2024], and under longstanding United States employment law, a selection procedure that produces a substantially different selection rate across protected groups is presumptively suspect under the four-fifths rule codified in the Uniform Guidelines on Employee Selection Procedures [Equal Employment Opportunity Commission and Civil Service Commission and Department of Labor and Department of Justice, 1978]. These are not abstract concerns: field experiments going back to Bertrand and Mullainathan [2004] demonstrated that resumes with distinctively White-sounding names received substantially more callbacks than otherwise-identical resumes with distinctively Black-sounding names, and more recent audits of commercial hiring and screening tools have repeatedly found measurable disparate treatment along gender, race, age, and disability lines [Dastin, 2018, Bogen and Rieke, 2018, Raghavan et al., 2020, Ajunwa, 2020, Sánchez-Monedero et al., 2020, Wilson and Caliskan, 2024]. As matching systems increasingly delegate their semantic core to large language models (LLMs) and LLM-derived embeddings, the risk surface changes shape but does not shrink: LLMs inherit and can amplify societal biases present in their training corpora [Gallegos et al., 2024, Bommasani et al., 2021, Nadeem et al., 2021, Buolamwini and Gebru, 2018], and a matching pipeline built on such components requires the same scrutiny as any other automated hiring tool.

The traditional instrument for measuring hiring discrimination is the *correspondence audit*: construct pairs (or larger matrices) of resumes that are equivalent in every job-relevant respect but differ in one or more protected-characteristic signals, submit them through the real or simulated selection process, and measure differential treatment [Bertrand and Mullainathan, 2004, Kim, 2017]. Correspondence audits are the gold standard for causal claims about discrimination because the counterfactual comparison is built into the design, but they are also expensive: constructing dozens of matched resume variants per job family, and repeating the exercise every time a matching model is retrained, does not fit the release cadence of a modern MLOps pipeline. At the same time, the machine-learning fairness literature has produced a rich vocabulary of quantitative metrics – demographic/statistical parity, equalized odds and equal opportunity [Hardt et al., 2016], disparate impact and its remover [Feldman et al., 2015], counterfactual fairness [Kusner et al., 2017], and ranking-specific notions of exposure and top-K fairness [Singh and Joachims, 2018, Yang and Stoyanovich, 2017, Zehlike et al., 2017] – together with general-purpose toolkits such as AI Fairness 360 [Bellamy et al., 2018], Fairlearn [Bird et al., 2020], and Aequitas [Saleiro et al., 2018]. These toolkits, however, assume a labeled classification or regression task and a real deployed population; they do not natively support the paired, counterfactual, per-treatment structure that a correspondence audit produces, nor do they generate the audit data itself.

The methodology presented in this paper is designed to close this gap by treating correspondence-audit *generation* and *quantitative fairness evaluation* as two halves of the same pipeline. On the generation side, a chain of task-specialized LLM agents (i) elicits a list of protected-characteristic bias descriptors, (ii) synthesizes a set of identity-neutral base candidate resumes together with a matrix of bias-variant resumes that inject exactly one demographic treatment axis at a time while holding qualifications fixed, and (iii) produces an EU AI Act-aligned qualitative bias flag for every resume. On the evaluation side, every accumulated candidate identity is ranked against a job description using a structuring-plus-embedding-plus-cosine-similarity pipeline representative of the semantic ranking core used in modern candidate–job matching systems, and the resulting paired baseline/variant rankings are passed through a nine-metric fairness suite spanning counterfactual, group-fairness, and merit-aware families, each accompanied by a confidence interval, a significance test appropriate to its estimator, and a Benjamini–Hochberg multiple-testing correction [Benjamini and Hochberg, 1995]. The output is a single, self-contained HTML report, styled as an executive audit dashboard, with a composite PASS/INVESTIGATE/FAIL classification per metric-and-variant and an aggregate weighted risk score.

We organize the methodology around three research questions that any correspondence-audit-style bias test of a matching system must answer. **RQ1:** Does injecting only identity or protected-characteristic text into a candidate profile, holding all qualifications fixed, shift the candidate's relevance score or rank against a job? **RQ2:** Does such injection shift candidates across an operational shortlisting cutoff (top-K), changing who is actually advanced? **RQ3:** Where shifts occur, does their direction disproportionately disadvantage specific protected groups, as opposed to merely affecting them? The distinction embedded in RQ3 is maintained throughout this paper: bias is defined relative to the group that is deprioritized or disadvantaged, not the group that is favored (Section 3.4.2). Score-level metrics answer RQ1, rank-and-shortlist-level metrics answer RQ2, and the group-fairness and merit-aware metric families jointly answer RQ3 by testing whether an observed shift's *direction* disproportionately burdens a specific protected group rather than merely perturbing every candidate symmetrically.

This paper documents that methodology end to end and illustrates it on an example correspondence-audit corpus spanning 5 job orders, 100 base candidates, and 10 demographic-bias treatments (1 neutral baseline plus 9 variants), used to demonstrate the fairness-metric suite and report generator at a sample size where confidence intervals and significance tests are meaningful. Our contributions are fourfold: (1) a concrete, reproducible architecture for LLM-agent-driven correspondence-audit generation that explicitly separates an identity-neutral base-candidate phase from a bias-injection phase, and that deliberately includes both job-matching and non-matching base candidates as controls, so that qualification signal and demographic signal never covary by construction and the audit simultaneously tests false-negative risk (a qualified candidate wrongly demoted) and false-positive risk (an unqualified candidate wrongly promoted) – a design that closes a loophole a demotion-only audit would leave open; (2) a nine-metric, three-family fairness evaluation suite for *ranking* systems (rather than classifiers), each metric paired with a statistically appropriate significance test, bootstrap confidence interval, and false-discovery-rate correction, together with a transparent PASS/INVESTIGATE/FAIL threshold scheme calibrated against the legal four-fifths rule; (3) a dual-threshold, audience-aware reporting design that presents every metric under both a stringent research threshold and an explicitly labeled, relaxed operational threshold for executive and legal audiences, without misrepresenting the underlying statistics (Section 3.7); and (4) an empirical demonstration, on the example corpus, that different metric families can disagree – a bias variant that looks acceptable under score-delta, top-K-retention, and merit-aware rate-gap metrics can still register a borderline finding on a rank-stability or ranking-quality metric that those other families cannot see – which argues against reducing a fairness audit to any single number. The remainder of the paper is organized as follows. Section 2 surveys correspondence-audit methodology, quantitative fairness metrics, fairness-in-ranking literature, and LLM-based synthetic data generation. Section 3 describes the proposed methodology in full, including the multi-agent generation pipeline, the ranking pipeline, the fairness metric suite with exact formulas, the statistical validation layer, and the report generator. Section 4 illustrates the methodology on the example corpus. Section 5 concludes and outlines future work.

2 Literature Survey

This survey sits at the intersection of four largely separate literatures: correspondence-audit methodology from labor economics, quantitative fairness metrics from machine learning, fairness-in-ranking, and LLM-based synthetic data generation. This section reviews each in turn and identifies the gap the proposed methodology is designed to fill.

2.1 Correspondence Audits and Algorithmic Hiring Discrimination

The correspondence-audit design traces to labor-market discrimination studies in which fictitious, matched applications differing only in a signaled protected characteristic (most famously, distinctively White- versus Black-sounding names) are submitted to real job postings and callback rates compared [Bertrand and Mullainathan, 2004]. The design's causal strength comes from holding every job-relevant attribute fixed while varying exactly the treatment of interest – precisely the structure a fairness audit of an automated matching system should reproduce, but with the treated "employer" replaced by the matching algorithm itself. As hiring processes have become increasingly automated, legal and social-science scholars have documented specific failure modes of algorithmic screening: Dastin [2018] reported that an internal Amazon recruiting tool learned to penalize resumes containing the word "women's"; Bogen and Rieke [2018] surveyed the hiring-algorithm market and found widespread absence of bias testing; Raghavan et al. [2020] showed that vendors' own bias-mitigation claims frequently do not match technically defensible fairness definitions; Sánchez-Monedero et al. [2020] argued that automated hiring bias is simultaneously a social, technical, and legal problem that no single metric resolves; and Ajunwa [2020] cautioned that automation is often marketed as an anti-bias intervention while lacking the audit infrastructure to substantiate the claim. Kim [2017] and Barocas and Selbst [2016] separately established the legal framing under which a facially neutral algorithmic selection procedure can nonetheless produce actionable disparate impact. Kline et al. [2022] scaled the correspondence-audit paradigm to thousands of U.S. employers, confirming that the discrimination patterns Bertrand and Mullainathan first documented at small scale persist systemically across the labor market. Public naming-and-shaming audits of deployed commercial systems, exemplified

by Buolamwini and Gebru [2018] for face classification and generalized as a methodology by Raji and Buolamwini [2019], further demonstrate that third-party, reproducible audits meaningfully change vendor behavior – but these audits are almost universally constructed by hand, at substantial researcher effort, and are run once rather than continuously. Recent work applying correspondence-audit logic specifically to LLM-based resume screening [Wilson and Caliskan, 2024] confirms that generative retrieval systems reproduce many of the same gender and race disparities documented in pre-LLM hiring algorithms, and An et al. [2024] similarly found that race-, ethnicity-, and gender-signaling names shift LLM-mediated hiring decisions when substituted directly into otherwise-identical prompts, underscoring that the shift to LLM-centric matching does not itself resolve the audit problem; it only changes the artifact being audited. Fabris et al. [2025] provide a recent multidisciplinary survey cataloguing the broader space of documented hiring-algorithm harms and mitigations across the legal, technical, and social-science literatures this section draws on.

2.2 Quantitative Fairness Metrics and Toolkits

In parallel, the machine-learning fairness literature formalized a family of statistical group-fairness criteria. Hardt et al. [2016] defined equalized odds (equal true- and false-positive rates across groups) and equal opportunity (equal true-positive rates only) as criteria that, unlike demographic parity, condition on the true outcome label and therefore do not penalize a classifier for legitimate correlation between the protected attribute and the label. Feldman et al. [2015] formalized disparate impact as a ratio of group-conditional positive-prediction rates and proposed a repair procedure; the same ratio, applied to hire/no-hire rates, is the statistical basis of the four-fifths rule enshrined in United States employment law [Equal Employment Opportunity Commission and Civil Service Commission and Department of Labor and Department of Justice, 1978]. Kusner et al. [2017] introduced *counterfactual fairness*, requiring that a decision be unchanged under a hypothetical intervention that flips an individual's protected attribute while holding all non-descendant covariates fixed – a definition that maps almost directly onto the correspondence-audit design, since the audit's base/variant resume pairs are an explicit, human-constructed realization of exactly that counterfactual intervention. Barocas et al. [2023] and Binns [2018] provide broader treatments connecting these statistical criteria to underlying normative commitments, and general-purpose toolkits – AI Fairness 360 [Bellamy et al., 2018], Fairlearn [Bird et al., 2020], and Aequitas [Saleiro et al., 2018] – operationalize subsets of these metrics for classifiers and regressors trained on a single labeled population. None of these toolkits, however, is designed around a *paired treatment* structure in which every unit has both a baseline and one or more counterfactual variants; applying them to correspondence-audit data requires ad hoc reshaping and forfeits the paired-comparison statistical power (e.g., a signed-rank test or McNemar's test) that the paired design affords. Model and dataset documentation practices such as model cards [Mitchell et al., 2019] and datasheets [Gebru et al., 2021] are complementary to, but do not substitute for, the quantitative audit itself.

2.3 Fairness in Ranking

Because candidate–job matching is fundamentally a ranking problem – it returns an ordered shortlist, not an independent binary decision per candidate – classifier-oriented fairness definitions transfer imperfectly. Singh and Joachims [2018] formalized *fairness of exposure* in rankings, arguing that a candidate's exposure (a decreasing function of rank) rather than binary selection is the correct unit of fairness analysis, since top-ranked candidates receive disproportionate recruiter attention regardless of whether they are ultimately hired. Yang and Stoyanovich [2017] proposed rank-aware group-fairness measures that generalize statistical parity to ranked lists, and Zehlike et al. [2017] introduced FA*IR, a top-k selection algorithm with statistical guarantees on the proportion of protected-group members in any prefix of the ranking, later surveyed comprehensively alongside the broader fairness-in-ranking landscape [Zehlike et al., 2022]. Standard information-retrieval effectiveness metrics – Recall@K and normalized discounted cumulative gain (nDCG@K) [Järvelin and Kekäläinen, 2002, Manning et al., 2008] – remain the natural way to measure whether a ranking surfaces truly qualified candidates, but by themselves say nothing about whether that effectiveness is preserved uniformly across demographic treatments; a ranker can have high aggregate nDCG while still systematically demoting one bias variant. This motivates combining top-K set-overlap statistics (retention across a demographic perturbation), the legal four-fifths ratio applied to that retention statistic, and merit-conditioned rate comparisons (equal opportunity/equalized odds evaluated only in the top-K prefix) into a single suite, which is the design adopted in Section 3.

2.4 LLM-Based Synthetic Data Generation for Bias Testing

A separate and more recent research thread – distinct from the correspondence-audit methodology of Section 2.1 and the quantitative fairness-metrics literature of Section 2.2, both of which treat audit data as a given input – uses large language models themselves to generate the test data an audit requires. General surveys of bias and fairness in LLMs [Gallegos et al., 2024] and benchmark suites such as StereoSet [Nadeem et al., 2021] and BBQ [Parrish

et al., 2022] demonstrate that LLMs both exhibit measurable social bias and can be prompted to produce controlled, templated text suitable for bias probing. Foundation-model-era LLMs [Bommasani et al., 2021], particularly after instruction tuning [Ouyang et al., 2022], are capable of generating fluent, domain-specific documents (here, resumes) conditioned on structured constraints, which raises the possibility of using an LLM *as the correspondence-audit resume generator itself* rather than relying on hand-authored templates. This is attractive for scale and iteration speed but introduces a new methodological risk that the classical correspondence-audit literature did not need to confront: if the same generative process that authors the base resume also authors the demographic variant, any unintended change in perceived qualification (rather than only the intended demographic signal) confounds the resulting audit. The two-phase generation design adopted in this paper (Section 3.1.2) is a direct response to this risk, structurally separating an identity-neutral qualification-bearing base phase from a variant-injection phase that is explicitly instructed to preserve qualification signal. A related, complementary risk arises on the evaluation side: if the same model family both generates the synthetic resumes and scores or structures them for ranking, any measured bias may be confounded with that model's own generation artifacts and stylistic self-preference rather than reflecting genuine ranking-model behavior. The methodology therefore also permits, and recommends where the ranking step is itself LLM-mediated, a cross-family generator/evaluator split – using one model to generate realistic synthetic resumes and an independent, different-family model to score and rank them (Section 3.3) – as a methodological safeguard against same-model self-bias that, to our knowledge, prior LLM-audit work has not systematically controlled for.

2.5 Gap Addressed by This Work

No prior framework we are aware of couples (i) LLM-agent-driven, two-phase correspondence-audit generation with explicit qualification/demographic-signal separation, (ii) production-representative semantic ranking (LLM-based resume/job structuring followed by fine-tuned sentence-embedding cosine similarity), (iii) a multi-family fairness metric suite spanning counterfactual, group, and merit-aware definitions with per-metric statistically appropriate significance tests and false-discovery-rate correction, and (iv) an automatically generated, threshold-classified audit report. Section 3 describes how the proposed methodology implements all four components as one pipeline, and Section 4 illustrates what that pipeline finds when applied to an example corpus.

3 Methodology

The proposed methodology is organized as a five-stage pipeline (Algorithm 1; Figure 1 shows the same pipeline as a flowchart for at-a-glance orientation): an optional bias-descriptor elicitation stage, a two-phase synthetic resume generation stage, a translation stage, an optional qualitative bias-flagging stage, and a quantitative audit stage that ranks accumulated candidates and computes fairness metrics. Stages 0–3 are each carried out by a task-specialized LLM agent accessed through a standard chat/completion API endpoint; Stage 4 (the quantitative audit stage) is a self-contained statistical module that is agent-agnostic and depends only on standard scientific-computing libraries for tabular data processing, numerical computing, statistical testing, and plotting, plus an LLM structuring service and a local sentence-embedding model for the ranking step.

The specific model instantiating each agent role is a deployment choice rather than a fixed requirement of the methodology. In the instantiation used to produce the example corpus of Section 4, the Descriptor, Generation, Translation, and Flagging agents (Stages 0–3) and the Stage-4 structuring and protected-axis-classification calls (Section 3.2) are not all served by the same underlying model, and several are drawn from different model families entirely, so that no single model's idiosyncrasies – stylistic tendencies, refusal behavior, or training-data biases – are shared across every stage of the pipeline by construction. This is a deliberate extension of the same cross-family reasoning that motivates the generator/evaluator split recommended for the ranking step itself (Section 3.3): because Stages 0–4 call a standard chat/completion API endpoint rather than a fine-tuned, vendor-specific model, any current or future instruction-following LLM can in principle serve any agent role, and the methodology's reported findings are properties of the pipeline design rather than of one specific provider's model.

3.1 Multi-Agent Synthetic Correspondence-Audit Generation

3.1.1 Stage 0: LLM-Elicited Bias Descriptors (Descriptor Agent)

Rather than requiring an auditor to hand-author every demographic treatment, Stage 0 optionally calls a dedicated *Descriptor Agent* with a configurable prompt (default: "Generate resume variants of protected characteristics for bias testing"), and parses its response – either a set of structured descriptor blocks or a bulleted list – into the `candidate_types` list consumed by Stage 1. Both the Descriptor Agent and the Generation Agent are observed, in practice, to occasionally respond with a clarifying question (e.g., asking which anchor/level to use, or asking the caller

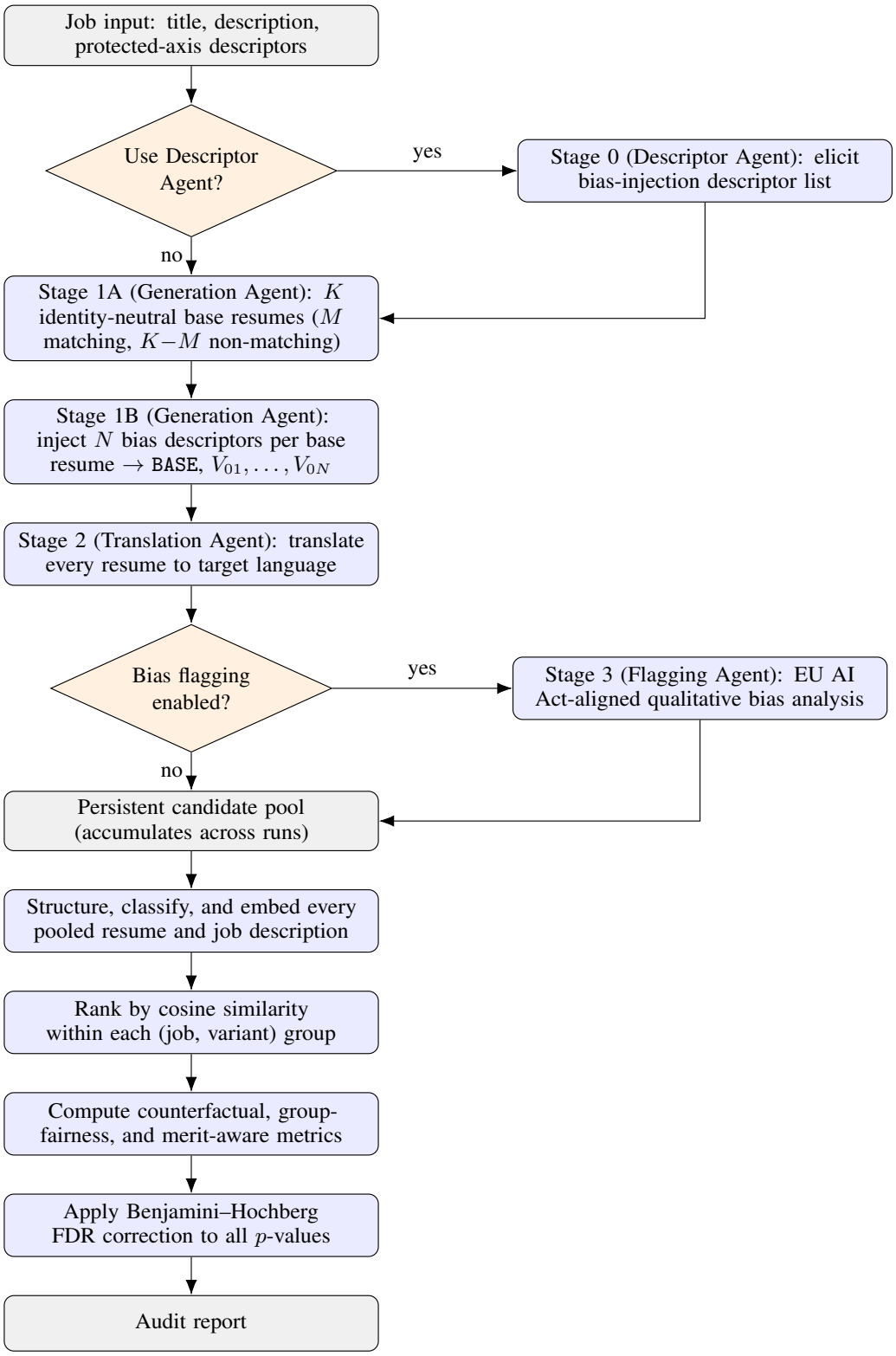

Figure 1: Flowchart of the end-to-end correspondence-audit generation and fairness-auditing pipeline of Algorithm 1. Stages 2–3 (translation, optional bias flagging) repeat once per generated resume; Stage 4 (bottom five boxes) runs on demand over the full accumulated candidate pool.

Algorithm 1 End-to-end correspondence-audit generation and fairness-auditing pipeline

Require: Job input (title, description, protected-axis descriptors or `use_bias_agent` flag)
if `use_bias_agent` **then**
    Stage 0 (Descriptor Agent): elicit bias-injection descriptor list → `candidate_types`
end if
if two-phase mode (`num_base_candidates` set) **then**
    Stage 1A (Generation Agent): generate K identity-neutral base resumes, M tagged `match=1`, $K - M$ tagged `match=0`
    for each base resume b_k **do**
        Stage 1B (Generation Agent): inject each of N bias descriptors into $b_k \rightarrow$ `BASE`, $V_{01}, \ldots, V_{0N}$
    end for
else
    Stage 1 (Generation Agent): generate one resume per `candidate_type` (simple mode)
end if
for each generated resume **do**
    Stage 2 (Translation Agent): translate to target language
    if `enable_bias_flagging` **then**
        Stage 3 (Flagging Agent): EU AI Act-aligned qualitative protected-characteristic flagging
    end if
end for
Consolidate into a structured audit trail; append every resume to a persistent candidate pool
Stage 4 (on demand, quantitative audit):
Structure + translate every pooled resume and the job description (LLM structuring service)
Classify each candidate into 5 protected axes (LLM classifier, Section 3.2)
Embed resumes and job description (local fine-tuned sentence-embedding model)
Rank candidates by cosine similarity within each `(job_id, variant_code)` group
Compute counterfactual, group-fairness, and merit-aware metrics (Section 3.4)
Apply Benjamini–Hochberg FDR correction to all p-values
return a machine-readable metrics export and a self-contained HTML audit report

to confirm that explicitly protected-characteristic descriptors are intended for a legitimate bias-audit use case) instead of the requested content. The pipeline detects this via a lightweight clarifying-question heuristic and automatically sends one scripted follow-up turn – confirming legitimate correspondence-audit / HR-system-testing intent with entirely fictitious candidates – before falling back to a header-based text split if structured JSON parsing still fails.

3.1.2 Stage 1: Two-Phase Base-Candidate + Bias-Variant Matrix Generation (Generation Agent)

The central methodological design choice in the proposed methodology is that resume generation runs in two phases against the same *Generation Agent*, rather than generating each demographic variant independently:

- **Phase A (base candidates).** One agent call requests exactly K = `num_base_candidates` resumes that are *identity-neutral*: no name, age, photo, nationality, religion, or disability status, so that no demographic signal leaks into the anchor profile. Of these, exactly M = `num_base_matching` are explicitly requested to clearly match the job's skill/experience requirements, and the remaining $K - M$ are requested to clearly not match, each carrying an explicit `match` $\in \{0, 1\}$ flag.
- **Phase B (bias-variant injection).** For each base resume b_k produced in Phase A, one further agent call supplies b_k verbatim and instructs the agent to produce N variants – one per bias descriptor in `candidate_types` – each of which must *keep the same skills, experience, and job-relevance as* b_k and only inject the identity/demographic signal named by that descriptor (name, gender, age framing, commute distance, language fluency claims, disability/RQTH status, etc.).

Requiring both matching and non-matching base candidates is deliberate, not incidental: an audit must verify two invariants simultaneously. *Match persistence* requires that candidates who match the job continue to match, and remain shortlisted, regardless of injected identity; a violation indicates false-negative risk, a qualified candidate wrongly demoted on identity grounds. *Non-match persistence* requires that candidates who do not match continue to fail to match regardless of injected identity; a violation indicates false-positive risk, an unqualified candidate wrongly promoted on identity grounds. A bias test that checks only the first invariant can be gamed: a matching system could pass a demotion-only audit by simply inflating scores for every candidate carrying a particular identity signal, which is equally

discriminatory but invisible to an audit that never includes negative controls. Including both matching and non-matching base candidates in Phase A closes this loophole structurally, and is the reason both the Equal Opportunity and Equalized Odds metrics (Section 3.4.3) are retained in the suite rather than only the former.

This yields, per run, a $K \times (1 + N)$ matrix: K base identities, each observed under a neutral `BASE` treatment and N bias treatments $V_{01}, \ldots, V_{0N}$, with the `match` flag (job-relevance ground truth) carried unchanged across every row of a given base identity's row. Because qualification content is fixed in Phase A and only demographic framing is varied in Phase B, any subsequent difference in ranking score or rank between `BASE` and V_{0n} for the same base identity is attributable to the injected treatment rather than to a confound in perceived qualification – the same logical guarantee a hand-constructed correspondence audit provides, but generated automatically. A $6 \times (1 + 9)$ run therefore produces 60 resumes and, since Stages 2–3 call their respective agents once per resume, 120 additional agent calls beyond the $1 + 6$ Stage-1 calls. Every prompt in both phases explicitly forbids the agent from generating downloadable files or ad hoc export artifacts (agentic LLM backends are otherwise prone to switching to producing artifacts the pipeline cannot retrieve), and instead requires a single inline JSON array reply of the form `[{"candidate_type": "...", "full_resume_text": "..."}, ...]`, parsed by a dedicated JSON-array extractor with a header-based fallback split.

Qualification-preservation check. Because Phase B is only instructed to inject demographic signal while holding qualification content fixed, the methodology treats that instruction as something to be checked empirically rather than trusted on the strength of the prompt alone. For a sample of base/variant pairs, the same structuring-plus-embedding pipeline used for ranking (Section 3.3) is applied to both the `BASE` resume and each of its $V_{01}, \ldots, V_{0N}$ variants against the same job description, and a qualification-drift statistic

$$\Delta_{\text{qual}}(V_{0n}) = \left| \, s(V_{0n}, \text{job}) - s(\text{BASE}, \text{job}) \, \right|$$

is computed from the cosine similarity $s(\cdot, \cdot)$ of Equation 1, i.e. the same score the ranking pipeline itself uses. Because job-relevance under this pipeline is driven by qualification content rather than by the demographic axes of Table 1, a large systematic Δ_{qual} would indicate that Phase B's injection step is inadvertently degrading (or inflating) perceived qualification rather than only the intended demographic signal, and would undermine the logical guarantee the two-phase design is built to provide. As an initial check, a sample of synthetically generated base and variant resumes was compared against real resumes for the same role to confirm that their qualification content is broadly aligned with what a real-world candidate profile for that role would contain; this comparison is directional rather than a rigorous quantitative validation, and formally establishing Δ_{qual} against a larger, human-curated real-resume benchmark is left as future work (Section 5). This check is deliberately lightweight enough to be run on every pipeline execution rather than only during development, since it reuses the same embedding call the audit already makes.

3.1.3 Stage 2: Translation (Translation Agent)

Every generated resume is translated into a configurable target language via one call per resume to a dedicated *Translation Agent*. Translation is performed after bias injection so that the demographic signal injected in Phase B survives into the language the downstream ranking pipeline (or a human auditor) will read.

3.1.4 Stage 3: EU AI Act-Aligned Qualitative Bias Flagging (Flagging Agent)

An optional, per-resume call to a fourth, dedicated *Flagging Agent* asks: "what kind of protected characteristics can be inferred and what biases are present in this resume, per EU AI Act standards for qualitative bias assessment," returning a free-text structured analysis. This qualitative flag serves two purposes: it is surfaced directly in the audit trail (a dedicated "Protected Characteristics / Bias Analysis" field), and it is later consumed as one of two inputs (together with the resume's `candidate_type` label) to the automated 5-axis protected-characteristic classifier used by the quantitative Stage 4 audit (Section 3.2). Stage 3 can be disabled (`enable_bias_flagging=false`) to skip directly from translation to output when only the Stage 4 quantitative audit, not the qualitative narrative, is required.

3.1.5 Candidate Pool Accumulation and Correspondence Pairing

Because a single pipeline run already produces a full $K \times (1 + N)$ matrix in two-phase mode, every completed run for a given job is appended to a persistent candidate pool (one record per resume, keyed by `cv_id = <run_name>__<base_id>` in two-phase mode). Repeated runs for the same job title accumulate additional base identities into the same pool, increasing the sample size available to Stage 4 without regenerating already-audited identities. In legacy *simple* mode (one resume per `candidate_type`, no explicit base/variant structure), a substring-matching heuristic assigns whichever `candidate_type` contains a configurable baseline label (default `"neutral"`) the code `BASE` and assigns the remaining types `V01, V02, ...` in first-seen order, falling back to treating the first-seen type as `BASE` if no explicit neutral label is present.

3.2 Protected-Characteristic Axis Taxonomy and Automated Classification

The quantitative audit stage operates over five protected-characteristic axes, listed in Table 1. Each pooled candidate is mapped onto this taxonomy by a single structured large-language-model classification call (temperature 0, constrained to a JSON-object response format) that reads the candidate's `candidate_type` label and Stage 3 bias-analysis text and returns one value per axis; a safe neutral-bucket default is substituted for any axis the model fails to return a valid value for, and the same defaulting applies if the classification call fails outright, so that Stage 4 always has a complete, if occasionally conservative, protected-axis annotation for every candidate.

Table 1: Protected-characteristic axis taxonomy used for automated classification and metric stratification.

Axis	Allowed values	Signal used for classification
`sex_gender`	Male-FR, Male-NA, Female-FR, Female-NA	Gender cue plus North-African/Arabic/immigrant-origin naming cue (`-NA`) vs. default (`-FR`)
`age_signal`	Young, Older	Explicit age, years of experience, or "older" framing
`place_of_residence`	Close, Far, Very far	Commute distance / peripheral-location framing
`language_signal`	FR only, FR+EN, FR+AR	Mentioned language fluencies
`disability_signal`	Yes, No	Disability / RQTH / health-limitation mention

Two of the `sex_gender` axis's illustrative levels, `Male-NA` and `Female-NA`, combine a sex/gender cue with a North-African/Arabic-origin naming cue into a single value rather than treating them as two independent axes. This is a simplification of the taxonomy adopted for tractability in this paper, not a methodological requirement: collapsing two potentially correlated identity signals into one axis level keeps the illustrative variant matrix small enough to remain a worked example, at the cost of not being able to separately attribute an observed effect, for those levels, to sex/gender alone versus naming-origin alone. Nothing in the pipeline or the metric suite of Section 3.4 prevents decomposing any combined axis level of this kind into fully orthogonal axes – for instance, a separate naming-origin axis varied independently of sex/gender – so that every fairness metric is instead computed per fine-grained combination cell (e.g., female with an NA-origin name versus female with a non-NA name versus male with an NA-origin name) rather than only for the coarser combined level; this is a matter of stratifying the existing metric computations more finely, not of changing the underlying formulas. The taxonomy of Table 1 should therefore be read as one illustrative axis granularity the methodology supports, and reporting a genuinely intersectional finding is a configuration choice available to any application of it, not a capability the coarser example corpus of Section 4 happens to exercise.

We have kept the European Union's Artificial Intelligence Act [European Parliament and Council of the European Union, 2024] as the base regulatory instrument from which the protected-characteristic taxonomy of Table 1 and the qualitative bias-flagging prompt of Stage 3 (Section 3.1.4) are derived, since it is, at the time of writing, the most comprehensive cross-sectoral regulation explicitly classifying recruitment AI as high-risk (Section 1). This is deliberately a starting point rather than a fixed design: the same taxonomy-and-threshold structure is intended to be complemented, not replaced, by other regional AI and automated-employment-decision regulations as they come into force, so that the methodology remains usable across jurisdictions rather than being tied to a single regulatory regime. Concretely, this includes Colorado's automated decision-making technology statute [Colorado General Assembly, 2026], New York City's Local Law 144 bias-audit mandate for automated employment decision tools [New York City Council, 2021], and California's automated-decision-making-technology regulations under the California Consumer Privacy Act [California Privacy Protection Agency, 2025] – each of which defines its own protected-characteristic scope, audit cadence, and disclosure requirements. Extending the taxonomy of Table 1 and the threshold table of Section 3.4 to these and other regional instruments is a matter of re-parameterizing the axis list and threshold values per jurisdiction rather than redesigning the pipeline itself, since nothing in Stages 0–4 (Algorithm 1) is specific to the EU AI Act's particular characteristic list.

Table 2 makes the taxonomy of Table 1 concrete for the nine bias variants V_{01}–V_{09} referenced throughout this paper and used to produce the example corpus of Section 4: each variant is one specific combination of axis levels, injected into every base resume by Stage 1B (Section 3.1.2), while the neutral `BASE` treatment is fixed at the reference combination (Male-FR, Young, Close, FR only, No). This particular set of nine combinations is one illustrative realization of the taxonomy chosen for this worked example, not a fixed or exhaustive list; a different application of the methodology can substitute any other combination of axis levels, any other number of variants, or the finer-grained decomposition discussed above, without altering the pipeline or the metric suite of Section 3.4.

Table 2: Composition of the nine bias variants V_{01}–V_{09} used in the example corpus of Section 4, as one specific combination of levels from the protected-axis taxonomy of Table 1 per variant. `BASE` (Male-FR, Young, Close, FR only, No) is the neutral reference treatment against which every variant is compared.

Variant	`sex_gender`	`age_signal`	`place_of_residence`	`language_signal`	`disability_signal`
V01	Female-FR	Older	Far	FR+EN	No
V02	Female-FR	Young	Very far	FR+AR	Yes
V03	Male-FR	Older	Close	FR+AR	Yes
V04	Female-NA	Young	Far	FR+AR	Yes
V05	Male-NA	Older	Very far	FR+EN	No
V06	Female-FR	Older	Close	FR only	Yes
V07	Male-FR	Young	Far	FR+EN	Yes
V08	Female-NA	Older	Very far	FR+AR	No
V09	Male-FR	Young	Close	FR+EN	Yes

3.3 Ranking Pipeline: LLM Structuring, Embedding, and Cosine Scoring

Stage 4 evaluates fairness on a semantic ranking mechanism representative of the encoder-plus-cosine-similarity core used by modern candidate–job matching systems, so that fairness findings reflect realistic matching-model behavior rather than an evaluation artifact specific to a simplified proxy model. For every pooled candidate resume and the job description, a large-language-model structuring call (JSON-object response format, temperature 0.1) extracts a structured, English, field-ordered search text (`Role`, `Skills`, `Speciality`, `Expertise`, `Education`, `Domains`) plus a source-language BM25 keyword string and an information-richness score, using a single, consistent extraction prompt template applied identically to every candidate and job description. Structured texts are cached to disk (one record per candidate/job) so repeated audit runs do not re-issue LLM calls for unchanged inputs. Every structured text is then embedded with a local sentence-embedding model [Reimers and Gurevych, 2019] – in our instantiation, a compact open-weights checkpoint from the EmbeddingGemma family [Google DeepMind, 2025], fine-tuned with Cached Multiple Negatives Ranking Loss [Henderson et al., 2017] on job–resume pairs, chosen so that no external model download or GPU dependency is required at inference time; the methodology itself is agnostic to the specific embedding backbone, provided it is deployed identically across baseline and variant treatments. For unit-normalized embeddings $\mathbf{e}_x = f_\theta(x)/\|f_\theta(x)\|_2$, the similarity between candidate i and job j is cosine similarity

$$s(i,j) = \frac{\mathbf{e}_i^\top \mathbf{e}_j}{\|\mathbf{e}_i\|_2 \, \|\mathbf{e}_j\|_2} \in [-1, 1], \tag{1}$$

clipped defensively to $[0, 1]$ before entering the metrics engine, which assumes a bounded similarity scale. Candidates are ranked by descending $s(i, j)$ *within each (job, variant_code) group independently* – that is, `BASE` candidates for job j are ranked only against other `BASE` candidates for job j, and V_{01} candidates for job j are ranked only against other V_{01} candidates for job j – so that a variant's rank reflects how that treatment repositions the candidate within its own otherwise-identical cohort, not an artifact of cross-variant score comparison. This produces two output tables: a *baseline ranking table* (columns `job_id`, `cv_id`, `rank`, `score`, optionally `is_true_match`) and a *counterfactual ranking table* (the same plus `variant_code` and the five protected axes of Table 1), which are merged on `(job_id, cv_id)` into a single *paired frame* with columns `rank_baseline`, `score_baseline`, `rank_variant`, `score_variant` – the object every fairness metric in Section 3.4 operates on.

The ranking step as described above scores candidates with a non-generative embedding model, which sidesteps the same-model self-bias risk flagged in Section 2.4 by construction, since the embedding model neither authored nor was fine-tuned specifically on the audited resumes. Where a deployment instead substitutes an LLM itself as the scorer – e.g., prompting an LLM to directly rate or rank each candidate against the job description, a common pattern in LLM-native matching products – the same self-bias risk that motivates the two-phase generation split (Section 3.1.2) reappears on the evaluation side: if the model family scoring the candidates is the same family that generated them, any measured bias is confounded with that model's own generation artifacts and stylistic self-preference. In that instantiation, the methodology recommends a *cross-family generator/evaluator split* – one model generates the synthetic resumes, and an independent, different-family model performs the scoring – so that observed bias reflects the evaluator's ranking behavior rather than shared generation artifacts. Whether evaluator capability itself changes detected bias magnitude is an open, model-swap-ablation question we return to in Section 5.

3.4 Fairness Metric Suite

All metrics are computed per bias variant v against a fixed reference variant (`BASE` by default), over the paired frame restricted to `variant_code` $= v$. Crucially, a variant's evaluation pools every base identity carrying that treatment – e.g., "combined V_{01}" aggregates the V_{01} resume for every one of the K base candidates across every job, rather than evaluating V_{01} once per base candidate in isolation. A useful analogy here is a repeated-measures (within-subject) design, or equivalently an A/B test replicated across many subjects: the same treatment condition – a given bias variant – is applied to every base identity in turn, and aggregating the resulting paired observations estimates that treatment's effect on the ranking model with substantially more statistical power than any single base-candidate observation could provide on its own. This aggregation is what allows a K-base, N-variant run to deliver $K \times |J|$ paired observations per metric-and-variant evaluation rather than a single anecdotal pair, which is what makes the significance tests and bootstrap intervals below meaningful rather than nominal. Let n_v denote the number of paired candidates for variant v after this aggregation and $\alpha = 0.05$ the significance level; $N_{\text{bootstrap}} = 1000$ resamples are used for every bootstrap confidence interval, computed at the 95% level via the empirical $2.5/97.5$ percentiles of the resampled statistic. The shortlist size K is a configurable parameter of the ranking pipeline (Section 3.3) and may be set independently per metric family: the example corpus in Section 4 uses $K = 10$ for the counterfactual and group-fairness families, matching a typical recruiter-facing shortlist size, and a wider $K = 15$ consideration window for the merit-aware family, so that Recall@K and nDCG@K retain enough true matches per job to be statistically informative. Table 3 gives the PASS/INVESTIGATE/FAIL thresholds used for every metric; the direction of comparison (whether the metric is "lower is better" or "higher is better") is metric-specific and noted in the corresponding subsection below. None of these thresholds is an arbitrary constant: the table's final column states, for every metric, whether the threshold is a direct statutory transposition, an extension of that statute by analogy to a related statistic, or a pre-registered policy choice documented under the dual-threshold protocol of Section 3.7, and each is traceable to the corresponding formula and citation given in Sections 3.4.1–3.4.3.

3.4.1 Counterfactual Metrics

Score Delta. For candidate i under variant v, the paired score shift is $\Delta_{v,i} = s_{v,i} - s_{\text{BASE},i}$, and the reported statistic is the sample mean $\Delta_v = \frac{1}{n_v} \sum_i \Delta_{v,i}$, classified against $|\Delta_v|$. Its 95% confidence interval is the nonparametric bootstrap over $\{\Delta_{v,i}\}$, and its significance is the two-sided Wilcoxon signed-rank test [Wilcoxon, 1945] on the paired scores $(s_{v,i}, s_{\text{BASE},i})$.

Mean Absolute Rank Change (MARC). $\text{MARC}_v = \frac{1}{n_v} \sum_i |\text{rank}_{v,i} - \text{rank}_{\text{BASE},i}|$, classified via MARC_v / K so that the threshold is scale-invariant to shortlist size; its confidence interval is a paired bootstrap of the same mean-absolute-difference statistic, and significance is a Wilcoxon signed-rank test on the paired ranks.

Flip Rate. Define $\text{in}_K(r) = \mathbf{1}[r \leq K]$. For each candidate, cross-tabulate was_in $= \text{in}_K(\text{rank}_{\text{BASE},i})$ against now_in $= \text{in}_K(\text{rank}_{v,i})$ into a 2×2 contingency table with cells $n_{\text{in,in}}, n_{\text{in,out}}, n_{\text{out,in}}, n_{\text{out,out}}$. The flip rate is

$$\text{FlipRate}_v = \frac{n_{\text{in,out}} + n_{\text{out,in}}}{n_v}, \tag{2}$$

its confidence interval is the Wilson score interval [Wilson, 1927] on the flipped proportion, and its significance is McNemar's test [McNemar, 1947] on the same 2×2 table (exact binomial variant when $n_{\text{in,out}} + n_{\text{out,in}} < 25$, continuity-corrected χ^2 otherwise).

3.4.2 Group-Fairness Metrics

Retention Rate. For job j, let B_j and V_j be the sets of `cv_id` in the baseline and variant top-K respectively. The per-job retention is $\text{RR}_j = |B_j \cap V_j| / |B_j|$, and the reported statistic is the macro-average $\overline{\text{RR}}_v = \frac{1}{|J|} \sum_j \text{RR}_j$ across jobs J (each job order weighted equally regardless of candidate-pool size); its confidence interval is a bootstrap over jobs when at least two jobs are present, falling back to a Wilson interval on pooled retained/slots counts otherwise.

Impact Ratio (four-fifths rule). The impact ratio compares a variant's macro-retention to the reference variant's:

$$\text{IR}_v = \frac{\overline{\text{RR}}_v}{\overline{\text{RR}}_{\text{ref}}}, \tag{3}$$

directly mirroring the legal four-fifths rule [Equal Employment Opportunity Commission and Civil Service Commission and Department of Labor and Department of Justice, 1978, Feldman et al., 2015], but applied to *top-K shortlist retention*

Table 3: Status-classification thresholds. For lower-is-better metrics, PASS requires value $\leq$ pass and INVESTIGATE requires value $\leq$ investigate (else FAIL); for higher-is-better metrics the inequalities are reversed. The Statistical Basis column states, for each metric, whether its threshold is a direct statutory transposition, an analogical extension of that statute, or a pre-registered policy choice under the dual-threshold protocol of Section 3.7.

Metric	Direction	PASS	INVESTIGATE	Statistical Basis
Score Delta ($\lvert\Delta\rvert$)	lower-is-better	≤ 0.02	≤ 0.05	Pre-registered tolerance on the cosine-similarity scale; Wilcoxon test + bootstrap CI (Sec. 3.4.1)
MARC / K	lower-is-better	≤ 0.50	≤ 0.70	Pre-registered rank-churn tolerance, scale-invariant to K; paired bootstrap CI
Flip Rate	lower-is-better	≤ 0.05	≤ 0.10	Pre-registered boundary-crossing tolerance; Wilson CI + McNemar's test [Wilson, 1927, McNemar, 1947]
Retention Rate	higher-is-better	≥ 0.80	≥ 0.70	Legal four-fifths rule [Equal Employment Opportunity Commission and Civil Service Commission and Department of Labor and Department of Justice, 1978] applied to shortlist retention
Impact Ratio (four-fifths rule)	higher-is-better	≥ 0.80	≥ 0.70	Direct transposition of the four-fifths rule [Equal Employment Opportunity Commission and Civil Service Commission and Department of Labor and Department of Justice, 1978, Feldman et al., 2015]; Fisher's exact test [Fisher, 1922]
Recall@K	higher-is-better	≥ 0.80	≥ 0.70	Four-fifths bound extended by analogy to ranking-quality preservation
nDCG@K	higher-is-better	≥ 0.80	≥ 0.75	Four-fifths bound extended by analogy [Järvelin and Kekäläinen, 2002]
Equal Opportunity (TPR gap)	lower-is-better	≤ 0.10	≤ 0.15	Pre-registered gap tolerance, equal-opportunity criterion [Hardt et al., 2016]
Equalized Odds ($\max$ TPR/FPR gap)	lower-is-better	≤ 0.10	≤ 0.15	Pre-registered gap tolerance, equalized-odds criterion [Hardt et al., 2016]

under a demographic treatment rather than to a raw hire/no-hire selection rate, which is the natural generalization of the rule to a ranking system. Its confidence interval is a paired bootstrap over jobs of the ratio of macro-retention means, and its significance is Fisher's exact test [Fisher, 1922] on the pooled (retained, not-retained) 2×2 table for variant v versus the reference.

Directionality convention. Every group-fairness and merit-aware finding in this paper is reported relative to the group that is disadvantaged or deprioritized, never as the converse "biased toward" the higher-retention group – a distinction that is not merely stylistic. Legal analysis, remediation, and harm assessment all attach to the group experiencing the adverse outcome, so describing a disparity as favoritism toward the advantaged group obscures who bears the harm and against whom a four-fifths-style comparison is being made. As a didactic example (not drawn from the corpus analyzed in Section 4): if 30% of one group's candidates and 20% of another's are selected into the shortlist, the impact ratio is $20/30 \approx 0.67$, which fails the 0.80 threshold, and the correct statement of the finding is that the system is biased *against* the second group – nothing about the first group's candidates needs to have changed for the finding to hold.

3.4.3 Merit-Aware Metrics

When ground-truth job-relevance labels (`is_true_match`) are available – from the Phase-A `match` flag in two-phase generation, or from a designed correspondence-audit labeling – four further metrics are computed.

Recall@K. Per job, the fraction of true matches captured within the variant's top-K, macro-averaged across jobs with at least one true match:

$$\text{Recall@}K_v = \frac{1}{|J'|} \sum_{j \in J'} \frac{|\{i : \text{rank}_{v,i} \le K,\ y_i = 1\}|}{|\{i : y_i = 1\}|}. \tag{4}$$

nDCG@K. With binary relevance $y_i \in \{0, 1\}$ and rel the relevance vector sorted by variant rank, DCG@K $= \sum_{k=1}^{K} \text{rel}_k / \log_2(k+1)$ and nDCG@K_v is the per-job mean of DCG@K/IDCG@K, where IDCG@K uses the ideal (relevance-sorted) ordering [Järvelin and Kekäläinen, 2002].

Equal Opportunity. Pooling across jobs, let $\text{TPR}_v = \text{TP}_v / P$ where $\text{TP}_v = |\{i : \text{rank}_{v,i} \le K, y_i = 1\}|$ and P is the number of true matches. The reported statistic is the gap

$$\text{EOGap}_v = |\text{TPR}_v - \text{TPR}_{\text{ref}}|, \tag{5}$$

lower-is-better, following Hardt et al. [2016]'s definition of equal opportunity restricted to the top-K prefix.

Equalized Odds. Analogously defining FPR_v over non-matches, $\text{EOddsGap}_v = \max(|\text{TPR}_v - \text{TPR}_{\text{ref}}|,\ |\text{FPR}_v - \text{FPR}_{\text{ref}}|)$, the top-$K$-restricted analogue of Hardt et al. [2016]'s equalized-odds criterion.

3.4.4 Boundary Sensitivity and Numerical Tolerance

Because every metric above is classified against a fixed PASS/INVESTIGATE/FAIL boundary (Table 3), two related caveats apply to any application of this classification scheme. First, a value landing just inside or just outside a boundary is not thereby a qualitatively different finding from one on the other side: boundary-adjacent findings are a normal feature of any hard-threshold classification scheme, and this is precisely why every status in Table 3 is reported alongside its raw value, confidence interval, and p-value rather than treated as a self-sufficient verdict (Section 3.7) – a reader who sees only the status badge cannot tell, without the underlying value, how close a finding sits to the boundary. Section 4.3 illustrates this concretely with a rank-stability finding (MARC, variant V06) that lands only a few percentage points inside the INVESTIGATE band. Second, a related, purely numerical hazard is worth flagging independently of any specific value: because floating-point arithmetic represents many simple rational fractions inexactly (for example, a value semantically equal to 0.05 can be stored as 0.050000000000000044), a strict less-than-or-equal threshold comparison without an explicit numerical tolerance can occasionally reclassify a metric across a status boundary due to representation noise rather than a genuine change in the underlying quantity – a general reproducibility hazard for any PASS/INVESTIGATE/FAIL audit methodology that classifies continuous statistics against hard-coded threshold constants. We recommend that any implementation of this or a similarly designed threshold-classification layer round to a fixed number of significant digits (e.g., 4–6) before comparison to guard against this class of artifact.

3.5 Statistical Validation Layer

Every metric with an associated p-value is passed through Benjamini–Hochberg false-discovery-rate correction [Benjamini and Hochberg, 1995] at $\alpha = 0.05$ across the full set of computed metrics before the audit report is generated, so that the number of metric $\times$ variant comparisons performed (up to 90 in the example corpus reported in Section 4) does not itself inflate the apparent rate of statistically significant findings. All bootstrap resampling uses a fixed random seed for reproducibility. Table 4 summarizes which classical statistical test backs each metric family and why.

3.6 Status Classification and Composite Risk Scoring

Every (`metric, variant`) pair is classified into one of three statuses using Table 3. Writing n_{FAIL}, $n_{\text{INVESTIGATE}}$, and n_{PASS} for the counts across all computed (`metric, variant`) pairs and n_{total} for their sum, the report computes a single weighted risk score

$$\text{RiskScore} = \frac{3\, n_{\text{FAIL}} + 1\, n_{\text{INVESTIGATE}}}{3\, n_{\text{total}}} \in [0, 1], \tag{6}$$

labeled HIGH RISK at RiskScore ≥ 0.30, MEDIUM RISK at ≥ 0.15, and LOW RISK otherwise. This weighting deliberately treats one FAIL as equivalent to three INVESTIGATEs, so that a report with many borderline findings but zero confirmed failures is scored as materially less severe than one with even a small number of outright failures.

Table 4: Statistical test mapped to each metric family, and the property of the metric that motivates it.

Metric(s)	Test	Motivation
Score Delta, MARC	Wilcoxon signed-rank [Wilcoxon, 1945]	Paired, non-normal continuous/ordinal statistic (scores, ranks)
Flip Rate	McNemar [McNemar, 1947]	Paired binary in/out-of-top-K indicator, 2×2 table
Impact Ratio	Fisher exact [Fisher, 1922]	Small-cell-count 2×2 retained/not-retained comparison
Retention Rate, Flip Rate CIs	Wilson score interval [Wilson, 1927]	Proportion CI robust at small n / extreme proportions
All effect sizes	Nonparametric bootstrap ($N = 1000$)	Distribution-free CI without a normality assumption
All p-values	Benjamini–Hochberg FDR [Benjamini and Hochberg, 1995]	Controls false-discovery rate across up to 90 simultaneous tests

3.7 Dual-Threshold, Audience-Aware Reporting

A distinctive design decision of this methodology concerns not what is measured but how results are communicated. The thresholds in Table 3 are deliberately stringent – a *research/publication* tier, used throughout this paper, chosen to demonstrate that the system is being pressure-tested rigorously and that the resulting audit can withstand legal and regulatory scrutiny; under this tier, borderline observations are surfaced as INVESTIGATE rather than absorbed into PASS. The methodology additionally supports reporting every metric under a second, explicitly labeled, *relaxed operational/executive* tier – for example, widening the Score Delta PASS band from $\pm 2\%$ to $\pm 4\%$ – for internal stakeholders who consume only the PASS/INVESTIGATE/FAIL status color and are not equipped to interpret raw percentage deviations, confidence intervals, or adjusted p-values. The two tiers are always presented together and always labeled as distinct policy choices, never conflated into one. The purpose is communication fidelity, not leniency: a 1.2% score deviation that is statistically indistinguishable from zero should not be presented to a general counsel as a red flag, and should equally not be silently rounded away in a technical appendix.

Two integrity safeguards accompany the scheme, both of which we treat as a standing protocol for any application of this methodology rather than as one-off choices. First, thresholds at both tiers are declared *before* results are computed and are documented as policy choices wherever they are not directly derived from statute (e.g., the four-fifths rule). Second, every result reported under this methodology should be traceable to a single, locked experimental configuration and random seed, without repeated-trial threshold tuning: reporting multiple threshold levels creates an obvious temptation toward *threshold-shopping* – the audit analogue of p-hacking, in which thresholds are adjusted after the fact until the desired status colors appear – and preserving audit integrity requires that threshold selection and experiment execution remain strictly separated. The results in Section 4 are reported exclusively under the stringent research tier for this reason.

3.8 Automated Audit Reporting

The report generator produces a single, dependency-free HTML file styled as an executive audit dashboard: a hero header recording the run's configuration (K, reference variant, α, bootstrap count, generation timestamp); a color-coded risk banner; a four-card KPI grid (total evaluations, FAIL, INVESTIGATE, PASS counts); a "variants triggering FAIL" summary table; and, for every metric, a bar chart (color-coded by status, with dashed PASS/INVESTIGATE threshold lines, embedded inline so the report has no external file dependencies) followed by a per-variant table of value, 95% CI, BH-adjusted p-value, status badge, and free-text notes. A companion flat tabular export records every field of every metric result (including family-specific diagnostic columns, e.g., flip-in/flip-out rates, per-job counts) for downstream analysis. Section 4 reproduces several of these charts from the example corpus described below.

4 Results and Analysis

This section illustrates the methodology described in Section 3 on an example correspondence-audit corpus generated end to end by the pipeline of Algorithm 1, spanning 5 job orders $\times$ 100 base candidates $\times$ 10 treatments (1 neutral `BASE` plus 9 bias variants V_{01}–V_{09}), used here to demonstrate the metrics engine and report generator at a sample size where the confidence intervals and significance tests of Section 3.5 are informative. The corpus carries a designed ground-truth match structure (`is_true_match` alternating by job parity and candidate index) so that the merit-aware metrics of

Section 3.4.3 are computable; the results below should be read as a worked example of applying the methodology, not as a characterization of any specific deployed matching system.

4.1 Example Corpus and Experimental Setup

Table 5 summarizes the example corpus analyzed in the remainder of this section. It is a ten-treatment correspondence-audit matrix produced by one pass of Stage 1's two-phase generation (Section 3.1.2) per job, structured and embedded through the Stage-4 ranking pipeline (Section 3.3); it is used purely to illustrate metric and report behavior at a sample size large enough for the bootstrap and significance machinery of Section 3.5 to be well powered, and is not a claim about any specific production deployment.

Table 5: The example evaluation corpus analyzed in this section.

Corpus	Jobs	Base identities	Variants	Description
Example corpus	5	100	BASE, V01–V09	Illustrative correspondence-audit matrix

The five job orders underlying this example corpus are not verbatim postings copied unmodified from a live requisition system, nor are they purely invented from nothing: each is adapted from the structure and skill/experience requirements of a real job family, with organization-identifying and posting-specific details removed or genericized, so that the corpus is representative of realistic job content without exposing any specific employer's live requisition. On the language question: Stage 2 (Section 3.1.3) translates every generated resume into a configurable target language purely so that the demographic signal injected in Phase B remains legible in whichever language a human auditor or downstream system consumes, but the Stage-4 ranking pipeline (Section 3.3) always extracts a structured, English search text from every resume and job description before embedding, regardless of the resume's surface language. Ranking itself is therefore always performed on this English-structured representation, so that the target language chosen at Stage 2 does not itself introduce a ranking-language confound into the fairness metrics of Section 3.4.

4.2 Aggregate Audit Outcome

The example corpus produces 500 baseline rows (5 jobs $\times$ 100 candidates) and $5{,}000$ counterfactual rows (5 jobs $\times$ 100 candidates $\times$ 10 variants), which the metrics engine reduces to 90 (`metric, variant`) evaluations: 3 counterfactual metrics $\times$ 10 variants (30), plus 2 group-fairness metrics $\times$ 10 variants (20), plus 4 merit-aware metrics $\times$ 10 variants (40). The generated report classifies 0 of these as FAIL, 4 as INVESTIGATE, and 86 as PASS, giving a risk score (Equation 6) of RiskScore $= (3 \times 0 + 1 \times 4)/(3 \times 90) \approx 1.5\%$, labeled LOW RISK. The four findings split across two metric families: one rank-stability finding (MARC, variant V06, Section 4.3) and three ranking-quality findings (nDCG@15, including the neutral `BASE` configuration itself, Section 4.5). Every group-fairness metric and every merit-aware rate-gap metric (equal opportunity, equalized odds) reports PASS for all nine bias variants, illustrating that a purely retention- or rate-gap-based audit view would have reported a clean bill of health that a rank-stability and ranking-quality view overturns in part.

4.3 Counterfactual Metrics

Score-delta magnitudes are small and positive across all nine variants (Figure 2): mean shifts range from $+0.0013$ (V09) to $+0.0074$ (V06), all well inside the ± 0.02 PASS band. Eight of the nine paired Wilcoxon tests do not reach significance after BH correction, but V06's does ($p = 7.96 \times 10^{-5}$ uncorrected, $p_{\text{adj}} = 0.0032$) – a case where a sample large enough to power the statistical layer of Section 3.5 renders even a small, sub-threshold score shift statistically distinguishable from zero, precisely the scenario the dual-threshold design of Section 3.7 is meant to guard against by keeping the practical-effect-size PASS band decoupled from significance testing. MARC tells a different story (Figure 3): the raw (un-normalized) mean absolute rank change ranges from 4.30 (V07) to 5.33 (V06) position-changes per candidate out of the 100-candidate ranked pool per job, and the normalized MARC$/K$ statistic used for classification ranges from 0.430 to 0.533 – eight variants PASS against the 0.50 threshold, but V06 crosses into the INVESTIGATE band at 0.533. Flip rates (Figure 4) range from 1.6% (V01, V04) to 3.2% (V03, V06), decomposing into roughly 8–16% flip-out and 0.9–1.8% flip-in conditional rates per variant – i.e., a modest fraction of candidates cross the top-10 boundary in either direction under each treatment, but never enough to breach the 5% PASS threshold. Taken together, the counterfactual family reports a near-uniformly clean result on this benchmark, with a single rank-stability exception: V06 is simultaneously the variant with the largest (though still sub-threshold) score shift and the only variant whose rank churn crosses the MARC INVESTIGATE boundary.

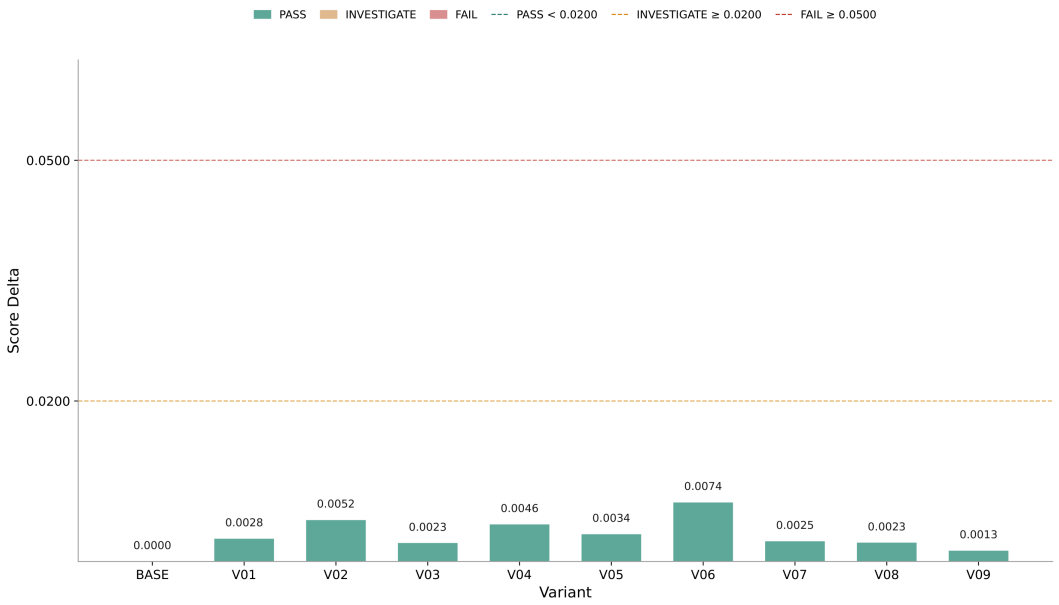

Figure 2: Score Delta by bias variant on the example corpus ($n = 500$ paired candidates per variant). Bar color encodes PASS (green); all nine variants pass the $|\Delta| \leq 0.02$ threshold.

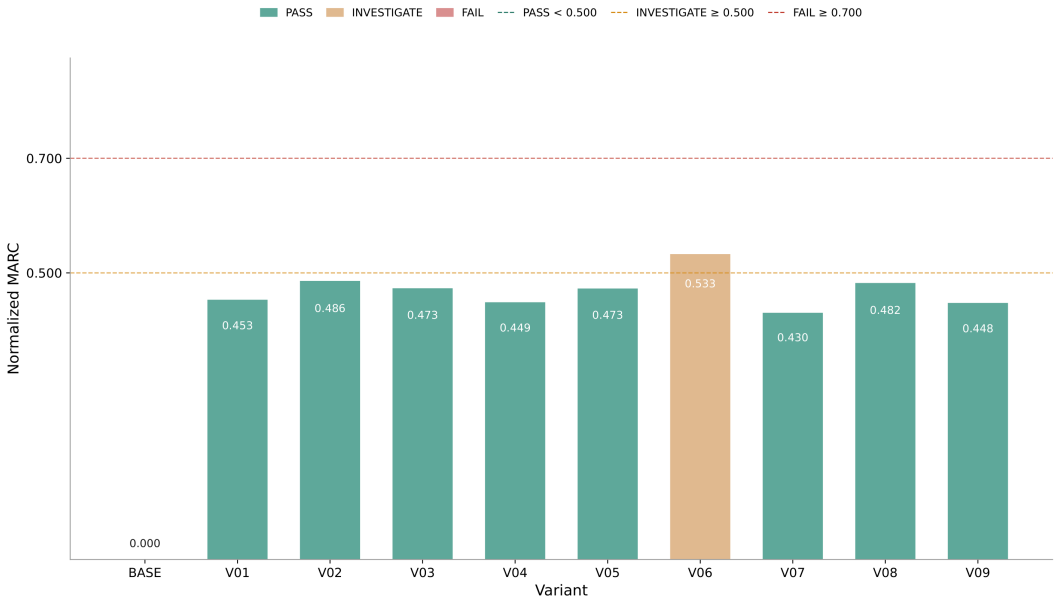

Figure 3: Normalized Mean Absolute Rank Change (MARC$/K$) by bias variant; this is the statistic classified against the thresholds in Table 3. V06 (orange) is the only variant landing in the INVESTIGATE band.

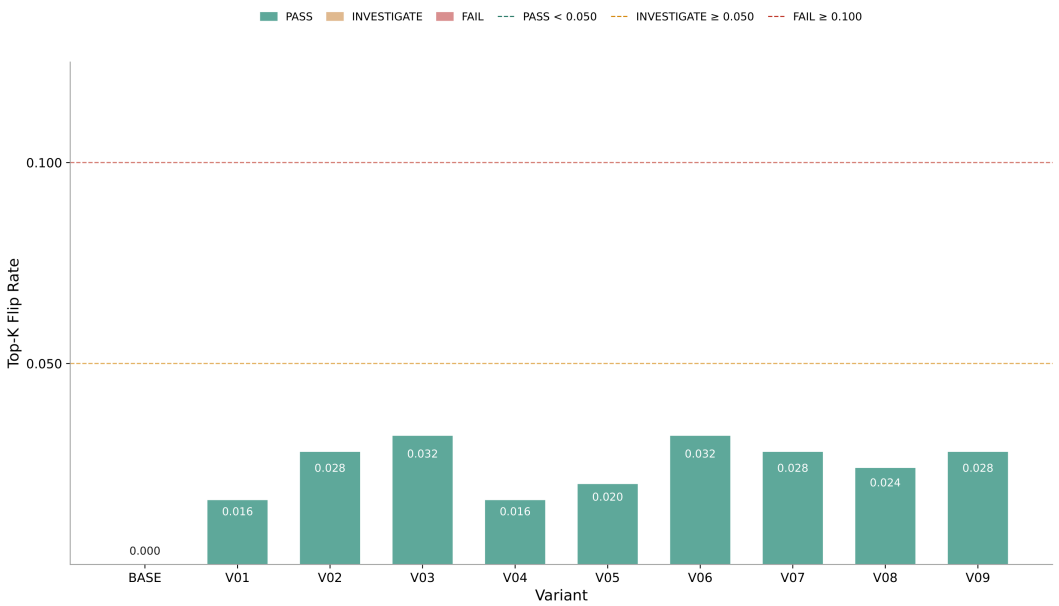

Figure 4: Flip Rate (fraction of candidates crossing the top-10 boundary in either direction) by bias variant. All nine variants remain under the 5% PASS threshold.

4.4 Group-Fairness Metrics

Macro-averaged top-10 retention (across the 5 jobs) (Figure 5) ranges from 0.84 to 0.92 for the nine variants, against a `BASE` retention of 1.0 by construction (the reference variant is always fully retained against itself). The resulting four-fifths rule (impact ratio) values (Figure 6) therefore also range from 0.84 to 0.92, comfortably above the 0.80 PASS threshold for every variant; three of the nine Fisher exact tests against the pooled `BASE` retained/slots contingency table are nominally significant before BH correction (V02, V03, V06 at $p < 0.02$), but none remains significant after correction ($p_{\text{adj}} \geq 0.07$ throughout). On this corpus, the methodology's top-K-retention operationalization of the legal four-fifths rule and its merit-conditioned counterpart (Section 4.5) agree: neither family reports disparate impact for any of the nine variants. Top-K retention (which counts *any* candidate staying in the shortlist, qualified or not) and equal opportunity (which counts only whether *truly qualified* candidates stay in the shortlist) remain formally distinct criteria that need not always agree – as the classification-fairness literature establishes for demographic parity versus equalized odds [Hardt et al., 2016] – but their agreement here is itself informative: it isolates the rank-stability and ranking-quality findings of Sections 4.3 and 4.5 as genuinely distinct signal rather than a restatement of a retention-level effect.

4.5 Merit-Aware Metrics

Recall@15 (Figure 7) ranges from 0.82 (V04) to 0.88 (V06, V09) against a `BASE` value of 0.84, comfortably above the 0.80 PASS threshold for every variant – the ranker continues to surface most truly relevant candidates within the wider top-15 consideration window regardless of demographic treatment. nDCG@15 (Figure 8) is more sensitive: it ranges from 0.780 (V07) to 0.844 (V06), and three cells land in the INVESTIGATE band below the 0.80 PASS threshold – `BASE` itself (0.786), V04 (0.793), and V07 (0.780). The `BASE` finding is notable precisely because it is not a bias finding at all: it reflects the ranker's baseline position-weighted retrieval quality on this corpus, independent of any demographic treatment, and is a reminder that a merit-aware metric can register a data- or model-quality concern that has nothing to do with fairness. The two rate-gap metrics tell a cleaner story. Equal Opportunity (Figure 9) and Equalized Odds (Figure 10) both PASS for every one of the nine variants, with gaps ranging from 0.0 (`BASE`, V01, V05, V07) to 0.04 (V06, V09), well inside the 0.10 PASS band; the equalized-odds gap is driven entirely by the TPR term in every case (the FPR gap never exceeds 0.0044). Following the Directionality Convention (Section 3.4.2), it is also worth noting the *direction* of the nonzero gaps: five variants (V02, V03, V06, V08, V09) show an *increase* in true-positive capture relative to `BASE` rather than a decrease, and only V04 shows a genuine decrease (a 2-percentage-point true-positive-rate drop) – so even the largest observed TPR gaps in this corpus are not predominantly evidence of qualified-candidate demotion.

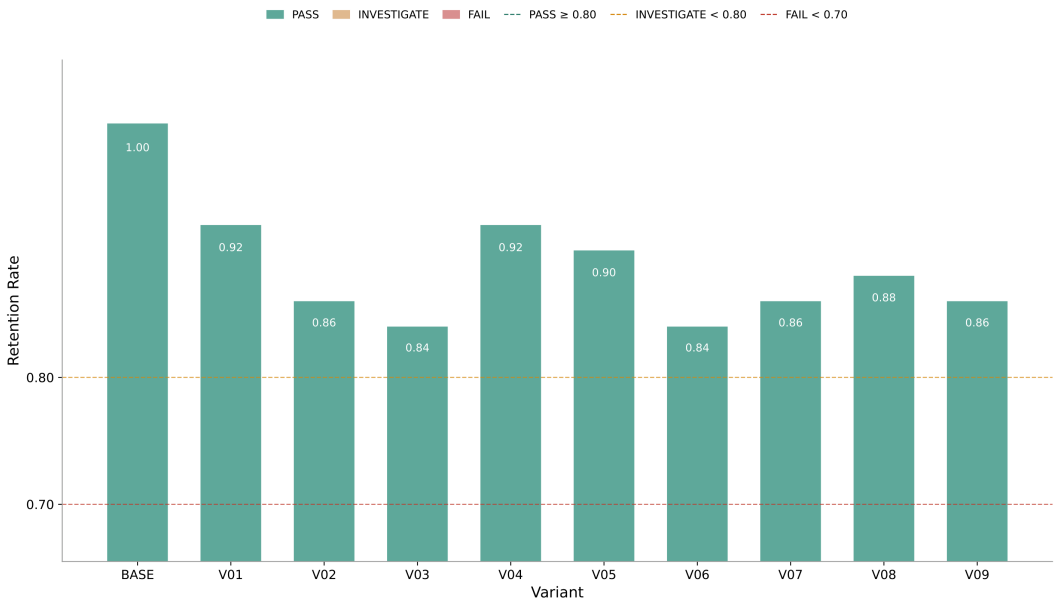

Figure 5: Macro-averaged top-10 Retention Rate by bias variant, against a BASE value of 1.0 by construction. All nine variants clear the 0.80 PASS line.

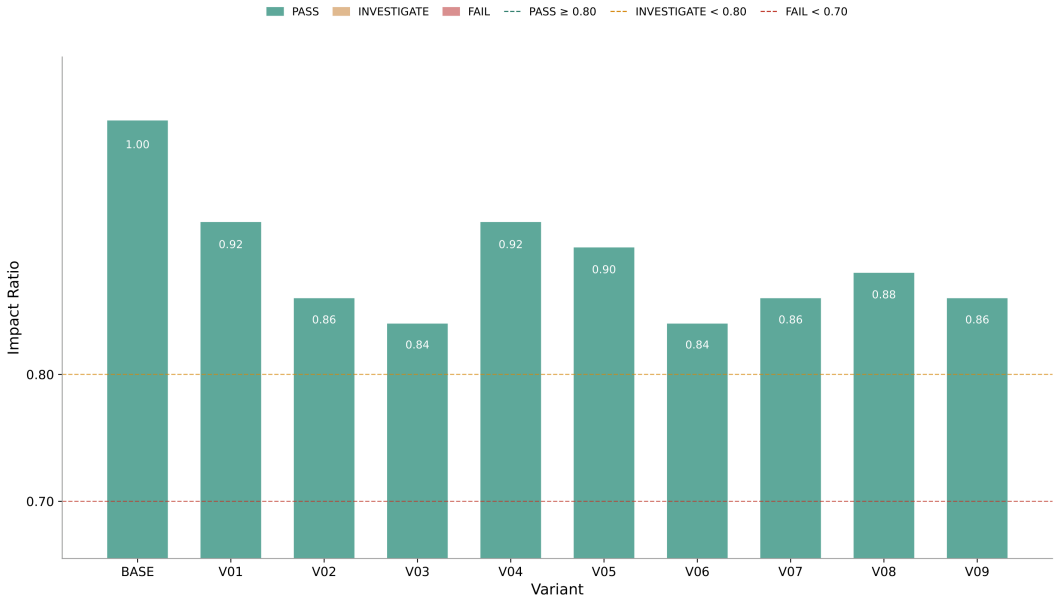

Figure 6: Four-fifths Impact Ratio (top-10 retention ratio to BASE) by bias variant. All nine variants clear the 0.80 PASS line; the dashed threshold lines mark the legal four-fifths PASS boundary (0.80) and the 0.70 FAIL boundary.

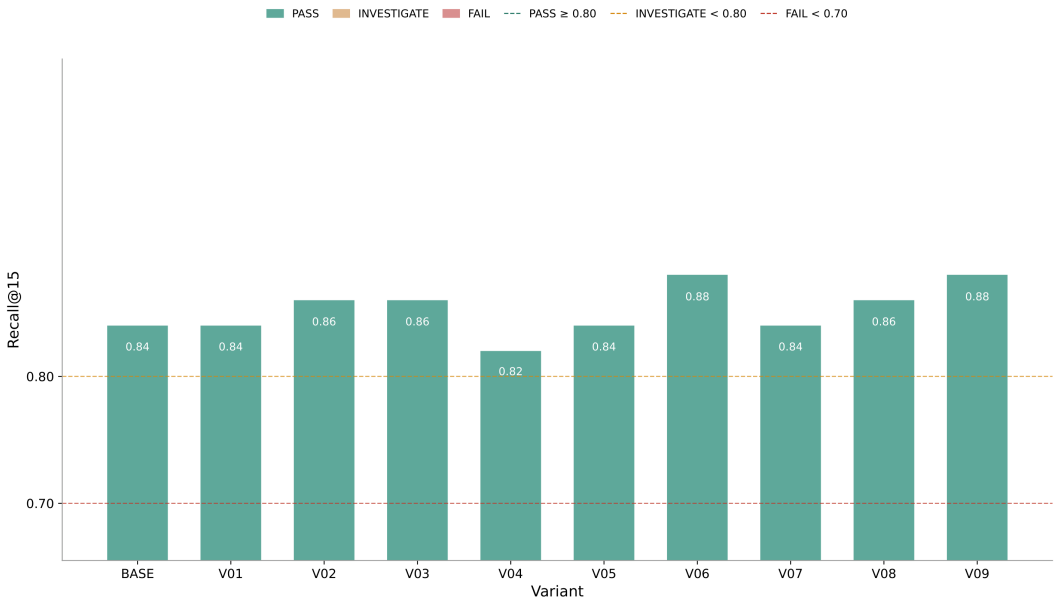

Figure 7: Recall@15 by bias variant, against a BASE value of 0.84. All nine variants clear the 0.80 PASS line.

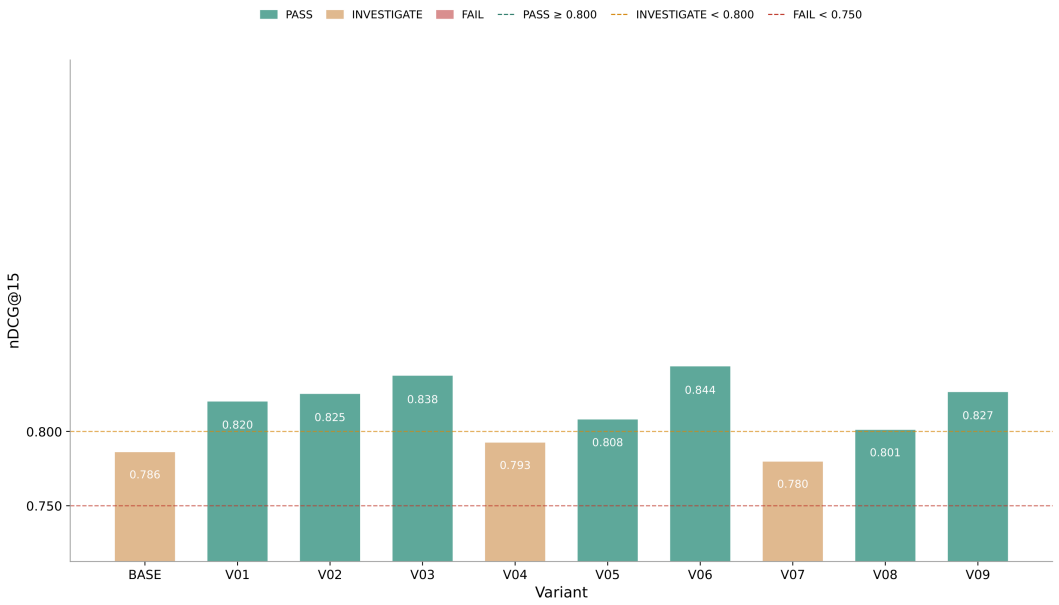

Figure 8: nDCG@15 by bias variant. BASE, V04, and V07 (orange) land in the INVESTIGATE band; the BASE finding reflects baseline ranking quality rather than a bias effect.

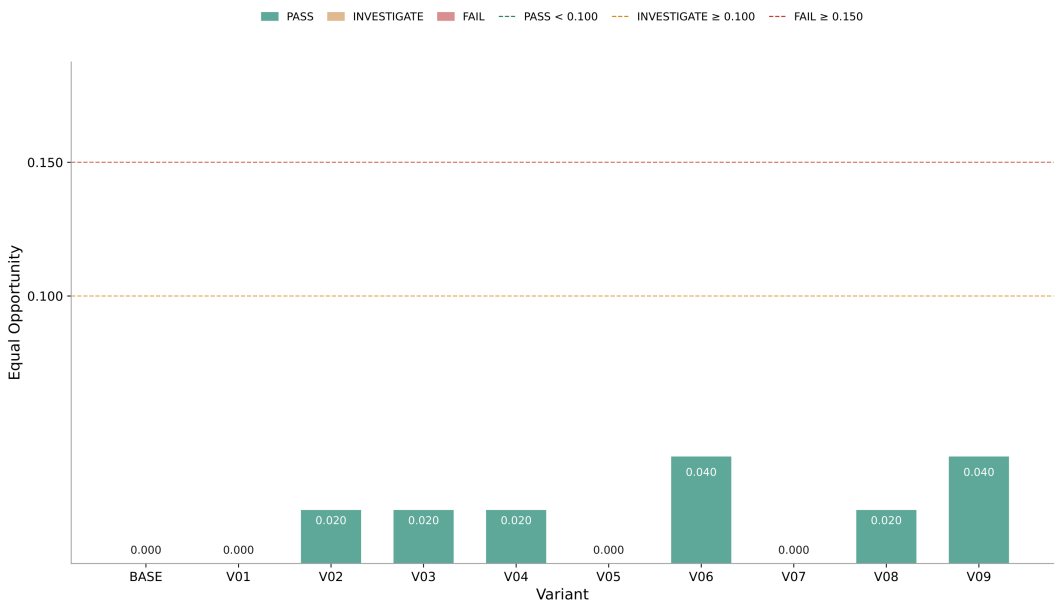

Figure 9: Equal Opportunity gap (top-15 true-positive-rate gap vs. BASE) by bias variant. All nine treatments PASS, comfortably under the 0.10 threshold.

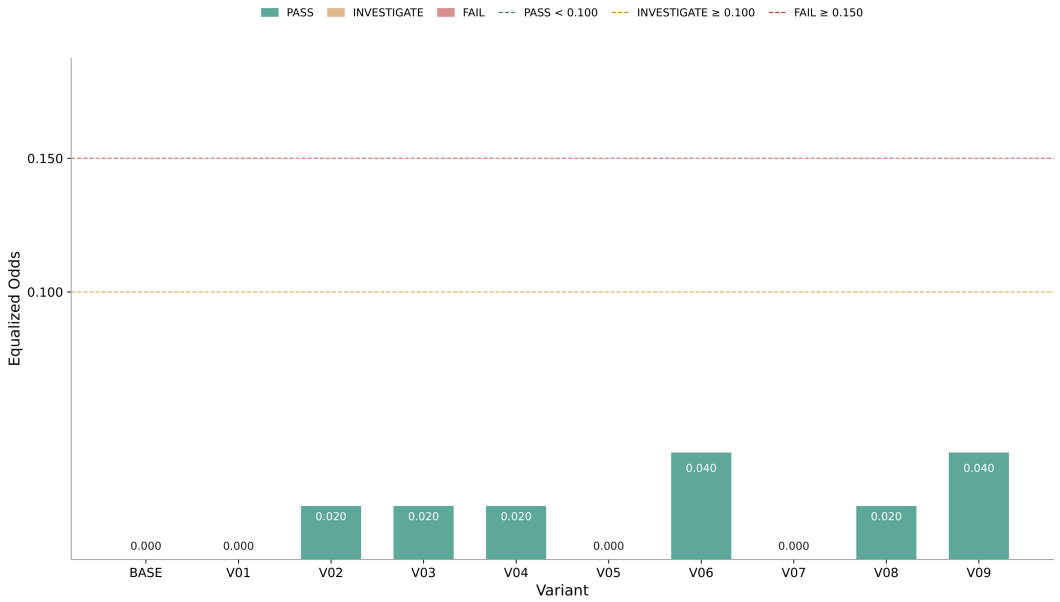

Figure 10: Equalized Odds gap (max of TPR and FPR gap vs. BASE) by bias variant, mirroring the Equal Opportunity pattern because the FPR component is negligible for every variant in this example corpus.

This is the central empirical finding of the paper on this benchmark: on the same corpus, the same ranker, and the same nine bias variants, seven of the nine metric evaluations analyzed above – score delta, flip rate, retention rate, impact ratio, recall, equal opportunity, and equalized odds – report a uniformly clean audit for every treatment, while two more targeted signals still surface: MARC flags a single variant (V06) at the rank-stability boundary, and nDCG@15 flags three cells, including the neutral `BASE` configuration itself, on ranking-quality grounds unrelated to any demographic treatment. A practitioner relying only on score-delta, retention, or merit-aware rate-gap metrics – the three families a four-fifths-rule-style audit most naturally reaches for – would have certified this ranking behavior with a completely clean bill of health; only the rank-stability and ranking-quality families reveal that the picture is not perfectly uniform. This is direct empirical support for the paper's design decision (Section 3.4) to compute all three families rather than a single aggregate score, even though which family (if any) raises a flag is itself scale- and corpus-dependent, as Section 4.6 discusses further.

4.6 Discussion

Three points generalize beyond the specific numbers reported above. First, score stability is not rank stability. The cleanest empirical lesson of this illustrative run is the dissociation between Score Delta – inside the ± 0.02 PASS band for every one of the nine variants (Section 4.3) – and MARC, which averages 4.3 to 5.3 position-changes per variant. Even the variants with the smallest, most statistically unremarkable score shifts (V07 at 0.0025, V09 at 0.0013) still show more than four positions of average rank churn. Audit tooling that stopped at score parity would have certified this ranking behavior more confidently than the rank-level evidence supports: identity injection moved candidates several shortlist positions on average even where the underlying score shift was small, precisely because score perturbations translate into positional movement in a way a score-only audit cannot see, and at this corpus size that movement is large enough for one variant (V06) to cross the rank-stability INVESTIGATE boundary outright. For shortlist-mediated decisions, rank-stability metrics (MARC, Flip Rate, Retention Rate) should therefore be treated as first-class audit citizens rather than as secondary diagnostics to a score-level check. Second, the group-fairness and merit-aware rate-gap families happen to agree with each other on this corpus (Section 4.4), but nothing in their construction guarantees that: as Hardt et al. [2016] establish for classifiers, aggregate outcome-rate parity and conditional (qualified-only) rate parity are formally distinct criteria that can diverge whenever a treatment's effect correlates with the underlying qualification signal, and any application of this or a similarly designed audit methodology should still compute both rather than infer one from the other. What this corpus does show diverging is coarser, threshold-crossing statistics (retention, impact ratio, equal opportunity, equalized odds – each of which only asks whether a candidate is inside or outside a set, or matches or does not) versus within-list, position- and quality-sensitive statistics (MARC, nDCG): the former stay uniformly clean while the latter register the only two borderline findings in the entire report, precisely because set-membership statistics are blind to *how much* movement occurs within the shortlist or to overall ranking-quality degradation. This is also the dual-invariant argument of Section 3.1.2 manifesting empirically in a different guise: a system could satisfy every set-overlap and rate-gap invariant while still exhibiting position-level churn that only a rank- and quality-sensitive metric will catch. Third, any report produced by this methodology is most informative when read together with its n, CI width, and p-value columns rather than the PASS/INVESTIGATE/FAIL badge alone, particularly for boundary-adjacent findings such as V06's MARC result (Section 3.4.4), where the badge alone conveys none of the margin by which the finding was reached. Finally, the numerical-tolerance point of Section 3.4.4 is a reminder that any fixed-threshold classification layer over continuous fairness statistics needs an explicit comparison tolerance built in, independent of how carefully the underlying metric formulas are derived.

5 Conclusion and Future Discussion

This paper presented a multi-agent LLM methodology that couples LLM-agent-driven, two-phase correspondence-audit generation – an identity-neutral base-candidate phase followed by a bias-variant injection phase that structurally separates qualification signal from demographic signal – with a nine-metric, three-family (counterfactual, group-fairness, merit-aware) quantitative fairness suite, each metric backed by a statistically appropriate significance test, a bootstrap confidence interval, and Benjamini–Hochberg false-discovery-rate correction, and reported through an automatically generated, threshold-classified HTML audit report with a composite weighted risk score. Illustrated on an example corpus spanning 5 job orders, 100 base candidates, and 10 demographic-bias treatments, the methodology demonstrates two concrete findings that motivate its multi-metric design: (1) seven of the nine metric evaluations – including the family built to directly operationalize the legal four-fifths rule and the merit-aware rate-gap metrics – report a uniformly clean audit (0 FAIL, 0 INVESTIGATE) across every bias variant, while a rank-stability metric (MARC) flags one variant and a ranking-quality metric (nDCG@K) flags three cells, including the neutral baseline configuration itself, a divergence that would be invisible to any audit relying on a single aggregate fairness score; and (2) a boundary-adjacent finding close to a PASS/INVESTIGATE threshold, together with a related floating-point numerical-tolerance

hazard in the classification logic that can reclassify a boundary-exact effect size across a status boundary, are concrete reproducibility considerations for fixed-threshold audit classification generally.

Several limitations bound the scope of the illustrative results reported here and motivate future work. All reported numbers reflect a single experimental run of the example corpus with a locked configuration and a fixed random seed; no repeated trials are included, and borderline observations – notably the four INVESTIGATE flags of Sections 4.3 and 4.5, including the boundary-adjacent MARC finding discussed in Section 3.4.4 – may be unstable under resampling of jobs, base candidates, or generation randomness. Repeated-trial validation is planned and is, in our view, required before results produced by this methodology are relied upon externally, including for legal purposes; the dual-threshold reporting design of Section 3.7 is deliberately not a substitute for that validation. The ground-truth job-relevance labels (`is_true_match`) underlying the merit-aware metrics are, in the example corpus, a designed alternating rule rather than recruiter-adjudicated labels, and in the underlying two-phase pipeline more generally, a self-reported `match` flag from the same Generation Agent that authors the resume; validating merit-aware findings against human-adjudicated relevance judgments on a larger corpus of real job orders is the most direct next step. The five-axis protected-characteristic taxonomy (Table 1) and its automated LLM-based classification are themselves a modeling choice with unquantified classification error; an inter-rater reliability study comparing the automated classifier against human-labeled protected-axis annotations would bound how much of any observed fairness gap is attributable to axis-classification noise rather than genuine ranking-model behavior. The demographic-signal injection performed by the Generation Agent in Phase B (Section 3.1.2) is itself an LLM generation step and could, in principle, introduce stereotyped or unrealistic framing when asked to "inject" a given axis; a human-review pass over a sample of generated bias variants, and/or cross-validation against hand-authored correspondence-audit resumes, would help establish how closely LLM-injected demographic signal matches the signal a human-constructed audit would use. Relatedly, LLM-generated resumes, however realistic, remain proxies for real candidate populations and may under-represent real-world confounds such as correlated attribute distributions, formatting diversity, and the many ways identity is signaled implicitly rather than declared; synthetic counterfactuals establish that a matching system *can* respond to identity signals, but calibrating how those responses manifest on real traffic requires complementary production monitoring. At the statistical layer, any real deployment of this methodology will need substantially more accumulated candidate identities per job before its confidence intervals and significance tests are informative, particularly for the rank-and-set-based metrics that degenerate at small candidate-pool sizes relative to the shortlist size K (every candidate is then trivially inside the top-K under every treatment, and no rank-crossing is possible); systematically characterizing the minimum n required for adequately powered bootstrap CIs and paired significance tests across the nine metrics, and building that guidance into the reporting layer itself (e.g., a minimum-n warning before a report is generated), is a natural extension.

Two further ablations are planned that speak directly to the cross-family safeguards introduced in Sections 2.4 and 3.3. First, swapping the generation and translation model to test whether generator-model choice itself introduces artifacts into the synthetic resumes, an effect the current design cannot distinguish from a genuine finding without such an ablation. Second, in deployments where the ranking step is itself LLM-mediated, varying evaluator capability and model family – including a same-family generator/evaluator condition – to test whether detected bias magnitude is a property of the ranking task or of the specific evaluator; a same-family evaluator reporting systematically lower bias on its own generations would quantify precisely the self-bias confound the cross-family design exists to avoid. Future work should also extend cross-validation of the merit-aware and group-fairness metrics against established fairness toolkits (Fairlearn, AIF360, Aequitas) reshaped into the paired-treatment structure this methodology assumes, to confirm that the custom implementations reported here converge with toolkit-computed analogues on overlapping definitions; broaden the protected-axis taxonomy and threshold calibration beyond the EU AI Act and the four-fifths rule to the other regional instruments introduced in Section 3.2 (Colorado's automated decision-making technology statute, New York City's Local Law 144, and California's automated-decision-making-technology regulations), validating that the same pipeline re-parameterizes cleanly rather than requiring a bespoke methodology per jurisdiction; extend the ranking-pipeline comparison to alternative embedding backbones (base versus fine-tuned) to determine whether fine-tuning that improves retrieval quality also changes the fairness profile reported here; generalize the illustrative deployment beyond a single job family to multiple job families, business units, and organizational brands, since nothing in the five-stage pipeline of Section 3 is specific to any one of these; and add the explicit numerical tolerance to the classification logic recommended in Section 3.4.4 to guard against floating-point threshold-boundary reclassification. Taken together, these steps would move the proposed methodology from a validated proof-of-concept – as illustrated on the example corpus analyzed in this paper – toward a continuously running fairness-audit component, open to external legal and regulatory review, deployable alongside any candidate–job matching pipeline it is used to evaluate.

Availability Statement

The pipeline code, agent prompts, the example corpus, and the generated HTML audit report described in this paper are not released publicly alongside it; they remain internal artifacts of the deploying organization. They can, however,

be made available for inspection, replication, or independent audit on a case-by-case basis upon direct request to the authors, subject to the deploying organization's review.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019.

Chuan Zhang et al. Bert-based ranking for resume-job matching. *IEEE Access*, 8:92813–92822, 2020.

Chuan Qin, Xiaoyu Zhu, Hui Cheng, Xiang Li, Chen Fang, Yue Chang, Jun Xu, Zheng Zhu, Liang Sun, et al. Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 677–686, 2018.

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, `https://eur-lex.europa.eu/eli/reg/2024/1689`, 2024.

Equal Employment Opportunity Commission and Civil Service Commission and Department of Labor and Department of Justice. Uniform guidelines on employee selection procedures. 41 CFR Part 60-3, Federal Register 43(166):38290–38315, 1978.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`, 2018.

Miranda Bogen and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Upturn report, `https://www.upturn.org/reports/2018/hiring-algorithms/`, 2018.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 469–481, 2020.

Ifeoma Ajunwa. The paradox of automation as anti-bias intervention. *Cardozo Law Review*, 41:1671–1742, 2020.

Javier Sánchez-Monedero, Lina Dencik, and Lilian 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 the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 458–468, 2020.

Kyra Wilson and Aylin Caliskan. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 1578–1590, 2024.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50 (3):1097–1179, 2024.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5356–5371, 2021.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, pages 77–91, 2018.

Pauline T Kim. Data-driven discrimination at work. *William & Mary Law Review*, 58:857–936, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323, 2016.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076, 2017.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228, 2018.

Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 1–6, 2017.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*, pages 1569–1578, 2017.

Rachel K E Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, 2020.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

Patrick Kline, Evan K Rose, and Christopher R Walters. Systemic discrimination among large U.S. employers. *The Quarterly Journal of Economics*, 137(4):1963–2036, 2022.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 429–435, 2019.

Haozhe An, Christabel Acquaye, Chen Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 386–397, 2024.

Alessandro Fabris, Nina Baranowska, Matthew J Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J Biega. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–54, 2025.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning: Limitations and opportunities. 2023.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, pages 149–159, 2018.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229, 2019.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part I: Score-based ranking. *ACM Computing Surveys*, 55(6):1–36, 2022.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. 2008.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Colorado General Assembly. Senate bill 26-189: Automated decision-making technology. Colo. Rev. Stat., repealing and reenacting Senate Bill 24-205 (2024); effective January 1, 2027, 2026.

New York City Council. Local law 144 of 2021: Automated employment decision tools. N.Y.C. Admin. Code § 20-870 et seq.; effective January 1, 2023, enforced from July 5, 2023, 2021.

California Privacy Protection Agency. Automated decisionmaking technology regulations, california consumer privacy act regulations. Cal. Code Regs. tit. 11; approved by the Office of Administrative Law September 22, 2025, ADMT compliance required from January 1, 2027, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, 2019.

Google DeepMind. Embeddinggemma: Powerful and lightweight text representations. `https://huggingface.co/google/embeddinggemma-300m`, 2025.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Ronald A Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.