

Linguistic Triggers of Gender and Racial Bias in Open-Weight LLMs Applied to Recruitment

Kosuke Kitahara¹, Nobuhiro Yamaguchi¹

¹Recruit Co., Ltd.

1-9-2 Marunouchi

Chiyoda-ku, Tokyo 100-6640 Japan

k.kitahara@r.recruit.co.jp, nobuhiro.yamaguchi@r.recruit.co.jp

Abstract

Open-weight large language models are rapidly entering hiring pipelines, yet their discriminatory failure modes—and the regulatory exposure these create under the EU AI Act high-risk classification (Annex III) and U.S. EEOC adverse-impact analysis—remain poorly understood. We present the first systematic, multi-model audit of open-weight LLMs that treats *job-posting language* as the primary experimental variable, evaluating six models (Llama 3.2, Mistral, Gemma 3, Qwen 3, Phi 3, DeepSeek-R1) across four controlled experiments that jointly probe recruiter-simulation and job-seeker-simulation tasks. We find that (1) agentic posting language depresses recruiter recommendation scores for female candidates ($r_{rb} = 0.309$, $p_{Bonf} = 7 \times 10^{-5}$; model-fixed-effects $r_{rb} = 0.448$), while communal language partially reverses the penalty; and (2) coded-exclusion language suppresses non-White recruiter scores at large effect sizes ($r_{rb} = 0.646$ – 0.758) and, on the *job-seeker* side, selectively deters non-White personas from expressing interest—operationalizing a chilling-effect mechanism at scale. A label-ablation experiment isolates the explicit demographic persona label as the primary causal driver, and Word Embedding Association Tests corroborate these findings at the representational level ($d = 1.01$ – 1.45 under Caliskan et al.’s multi-word gender attribute lists). We translate these results into a concrete pre-deployment audit protocol—posting-vocabulary scoring, persona-conditioned LLM probing, and adverse-impact flagging against the four-fifths threshold—that operationalizes the documentation and risk-management obligations Annex III imposes on high-risk AI in recruitment.

1 Introduction

In recent years, the integration of Large Language Models (LLMs) into corporate recruitment processes—ranging from automated job description generation and candidate screening to job-matching recommendation systems and chatbot-based applicant interaction—has become a fundamental component of the hiring workflow. While LLM adoption promises improved consistency and operational efficiency, it also harbors significant legal and ethical risks. The European Union’s AI Act (EU AI Act) classifies AI use in recruitment and employment as high-risk, emphasizing strict governance and human oversight (European Parliament and

Council of the European Union 2024). Furthermore, the U.S. Equal Employment Opportunity Commission (EEOC) has warned of Adverse Impact risks, where AI selection may disadvantage specific groups (Commission et al. 2023). We deliberately ground the present audit in both the U.S. EEOC adverse-impact tradition and the EU AI Act high-risk classification because, despite differing legal definitions of protected characteristics and liability allocation, these Western jurisdictions converge on a shared concern: language-mediated discrimination in recruitment is a regulatable harm whose detection requires pre-deployment auditing of how AI systems respond to the linguistic surface of job postings. Cross-jurisdictional engagement is therefore not a stylistic choice but a substantive one—the auditing methodology must be portable across compliance regimes even though the downstream categories (race/ethnicity in the U.S., a broader set of protected characteristics in the EU) are not directly interchangeable. Model outputs can reproduce or amplify stereotypes rooted in training data or prompt design; thus, elucidating these behaviors is a critical challenge for ensuring AI governance and accountability across these frameworks.

We design the present audit so that its outputs map directly onto two operational obligations these frameworks already impose. First, Annex III of the EU AI Act lists “recruitment or selection of natural persons” as a high-risk use case, triggering provider obligations on documented risk management, data governance, transparency, and human oversight throughout the system life-cycle (European Parliament and Council of the European Union 2024); providers must therefore be able to evidence, before deployment, that their AI does not respond to lawful posting variations in ways that disadvantage protected groups. Second, U.S. EEOC adverse-impact analysis operationalizes disparate-impact concerns through the “four-fifths rule” (Commission et al. 2023): a selection rate for a protected group that falls below 80% of the majority group’s rate is treated as *prima facie* evidence of adverse impact, requiring a job-relatedness defense. Linguistic auditing of job postings sits at exactly the leverage point both frameworks demand: the same job ad can be pre-processed for AI-suppressive vocabulary and the same LLM can be probed with controlled persona variations to estimate group-conditional output distributions *before* a posting is published, yielding a documented, replica-

ble audit trail consistent with Annex III and a quantifiable adverse-impact indicator consistent with the four-fifths rule. This is the deployment-relevant target our experiments and proposed protocol (Section 5.6) are designed to support.

Gender bias in job advertisements is well-documented. Gaucher et al. demonstrated that job titles and vocabulary in job advertisements are clearly divided into masculine traits (such as competitiveness and dominance) and feminine traits (such as cooperation and supportiveness) (Gaucher, Friesen, and Kay 2011). Male-dominated professions tend to use “agentic” (power-oriented) vocabulary such as “competitive” or “dominant,” which can inhibit the “sense of belonging” for women and decrease their intent to apply. As a result, job seekers may unconsciously judge whether a position is suitable for them, leading to the maintenance of “occupational segregation”—the exclusion of specific genders from certain professions. According to Eagly and Karau’s Role Congruity Theory, the traditional “leader image” often aligns more closely with masculine traits and is perceived as inconsistent with feminine traits (Eagly and Karau 2002). Thus, even highly skilled women may be unfairly underrated at hiring and evaluation stages, or labeled with negative attributes such as “emotional,” due to a perceived mismatch between job titles and stereotypical images. Damelang et al. analyzed large-scale job advertisement data (on the order of one million listings) and confirmed that, even in today’s digital labor market, linguistic gender segregation persists according to job titles and job descriptions (Damelang, Rückel, and Stops 2024). This finding suggests a risk that AI-driven automated screening or algorithms, which learn from data containing historical biases, may further entrench inequalities.

Bertrand and Mullainathan empirically demonstrated how candidate names alone can function as powerful selection filters in the labor market (Bertrand and Mullainathan 2004). Even when candidates had identical educational backgrounds and job experience, those with White-sounding names received 50% more interview callbacks than those with Black-sounding names. Critically, improvements in resume quality benefited White candidates substantially but yielded no comparable gains for Black candidates, suggesting that evaluators use name as a proxy for ability—a form of statistical discrimination rooted in unconscious bias rather than information about actual qualifications.

Rivera traced a related mechanism to the interview stage, showing that recruiters in elite professional firms systematically favor candidates who share their leisure activities, cultural references, and self-presentation styles—a process he termed “cultural matching” (Rivera 2012). Although framed as assessing “organizational fit,” this criterion in practice advantages candidates from racially and socioeconomically privileged backgrounds, placing equally qualified minority applicants at risk of exclusion for reasons unrelated to job performance.

Building on the preceding discussion of persistent gender and race biases in traditional and AI-assisted hiring processes, this study aims to systematically investigate the emergence and mechanisms of such disparities within large language models (LLMs) applied to recruitment scenarios.

We experimentally vary job posting language—agentic versus communal vocabulary for gender bias, and inclusive versus culturally-coded language for racial bias—with job titles co-varied to reflect the target gender-stereotype theme. Synthetic candidate profiles are constructed with explicit gender, race, and ethnicity attributes, allowing targeted analysis of model responses from both the recruitment agent and applicant perspectives. This approach is designed to shed empirical light on the specific conditions under which LLM biases manifest during simulated hiring, an area where robust evidence remains limited. Accordingly, this study poses the following two Research Questions, examined from both the recruiter and job-seeker perspectives:

- **RQ1:** Does an LLM produce systematic differences in evaluation outcomes for candidates of different genders, given identical qualifications and varying job description linguistic styles (agentic vs. communal)?
- **RQ2:** Does an LLM favor “cultural fit” over diversity, or vice versa, when evaluating candidates from different racial or ethnic backgrounds under controlled and otherwise equivalent job description language?

To address these questions, we conduct four controlled experiments spanning both the recruiter and job-seeker perspectives, supplemented by two follow-up analyses (H1–H2). Results are organized around RQ1–RQ2 above.

2 Related Work

2.1 Gender and Racial Biases in AI-Assisted Hiring

In the context of algorithmic systems, Köchling and Wehner conducted a systematic review of algorithmic hiring tools and found that discrimination and fairness failures are pervasive across the HR lifecycle, from initial screening through development decisions (Köchling and Wehner 2020). Raghavan et al. further showed that vendor claims about bias mitigation in commercial hiring tools are often unsubstantiated, and that adverse impact can arise even when protected attributes are excluded from model inputs (Raghavan et al. 2020). De-Arteaga et al. demonstrated that semantic representations of occupational biographies encode gender in ways that propagate to downstream inference, a pattern directly relevant to LLM-based resume evaluation (De-Arteaga et al. 2019).

Evidence that gendered vocabulary in job advertisements shapes applicant pools predates the AI era. Gaucher et al. demonstrated empirically that agentic wording (e.g., *competitive*, *dominant*) in postings correlates with male-dominated occupations and suppresses women’s sense of belonging (Gaucher, Friesen, and Kay 2011). Damelang et al. extended this finding to nearly one million contemporary job listings, confirming that linguistic gender segregation persists in the digital labor market (Damelang, Rückel, and Stops 2024). Role Congruity Theory provides a theoretical explanation: when leadership or job titles carry masculine connotations, even highly qualified women face a perceived mismatch that translates into lower evaluations (Eagly and Karau 2002).

These linguistic patterns are structurally encoded in word representations as well. Caliskan et al. showed via the Word Embedding Association Test (WEAT) that static embeddings associate male names with career concepts and female names with family concepts (Caliskan, Bryson, and Narayanan 2017), while Bolukbasi et al. demonstrated that the direction “man is to computer programmer as woman is to homemaker” is geometrically encoded in word2vec (Bolukbasi et al. 2016).

More recently, studies have shown that LLMs trained on such biased corpora reproduce and sometimes amplify these patterns. Kotek et al. found systematic gender stereotyping in GPT-4 outputs across occupational roles (Kotek, Dockum, and Sun 2023), and Wan et al. documented that LLM-generated reference letters use warmer, less achievement-oriented language for women than for men (Wan et al. 2023). A large-scale audit using over 360,000 randomized resumes found that leading LLMs (e.g., GPT-4o, Gemini 1.5 Flash) systematically assigned lower evaluation scores to Black male candidates while over-correcting in favor of Black female candidates (An et al. 2025); a complementary audit of embedding-based resume retrieval found that White-associated names were favored in 85% of comparisons, with Black men disadvantaged most (Wilson and Caliskan 2024); intersectional biases of this kind resist current safety-training methods, consistent with the “gendered race prototype” hypothesis.

Bertrand and Mullainathan provided foundational evidence that race-associated names function as selection filters: resumes with White-sounding names received 50% more callbacks than otherwise identical resumes with Black-sounding names, an advantage that persisted even with resume improvements (Bertrand and Mullainathan 2004). Rivera 2012 traced a related mechanism to the interview stage, where interviewers in elite firms systematically favored candidates who shared their leisure activities and self-presentation styles—a process she termed “cultural matching.” Cultural matching in Rivera’s sense is interpersonal (an evaluator–candidate alignment in class-coded markers) and is conceptually distinct from the broader human-resources discourse of “cultural fit” or “person–organization fit,” which casts the same criterion in organizational-similarity terms; Rivera’s contribution was precisely to show that what is reported as organizational “fit” is, mechanically, interpersonal cultural matching that disadvantages candidates from racially and socioeconomically marginalized backgrounds.

Human-in-the-loop (HITL) setups do not reliably mitigate these biases. Wilson et al. conducted a resume-screening experiment ($N = 528$) in which participants collaborated with simulated LLMs exhibiting race-based preferences; even with moderately biased AI, participants followed the AI’s recommendations up to 90% of the time (Wilson et al. 2025). These findings echo broader concerns about bias in machine learning systems (Mehrabi et al. 2021) and the risks of delegating high-stakes decisions to language models trained on unrepresentative corpora (Navigli, Conia, and Ross 2023).

2.2 Novelty of the Present Study

Existing LLM hiring audits predominantly treat job postings as fixed stimuli and manipulate *candidate attributes*—names, demographic markers, or resume content—to measure differential treatment (An et al. 2025; Kotek, Dockum, and Sun 2023; Wan et al. 2023). This design answers whether LLMs treat demographically distinct candidates differently, but leaves the linguistic environment of the posting itself unexamined as a potential bias driver.

The present study departs from this paradigm in two ways. First, and most centrally, we treat *job posting language* as the primary experimental variable—systematically varying both the agentic–communal dimension of vocabulary (for gender bias) and the inclusive–coded-exclusion dimension (for racial bias)—and demonstrate that LLM hiring bias is *triggered* by the linguistic style of the posting, not only by the demographic label attached to the candidate; to our knowledge no prior multi-model audit has operationalized this posting-language manipulation across both gender and race bias in a unified framework. Second, by jointly simulating both the recruiter and job-seeker perspective within the same experimental framework, we show that the same linguistic trigger affects evaluator and applicant behavior *asymmetrically*.

3 Methods

3.1 Models

We selected six open-weight models deployable via Ollama on consumer hardware: **Llama 3.2** 3B (Meta), **Mistral** 7B v0.3 (Mistral AI), **Gemma 3** 4B (Google), **Qwen 3** 8B (Alibaba), **Phi 3** 3.8B Mini (Microsoft), and **DeepSeek-R1** 7B distill (DeepSeek); see Table 1. All six models were used in all experiments. All models were queried with temperature 0 to ensure deterministic outputs,¹ and responses were constrained to a JSON schema via Ollama’s structured-output API, minimizing free-form variation.

Model	Developer	Params
Llama 3.2	Meta	3B
Mistral	Mistral AI	7B (v0.3)
Gemma 3	Google	4B
Qwen 3	Alibaba	8B
Phi 3	Microsoft	3.8B (Mini)
DeepSeek-R1	DeepSeek	7B (distill)

Table 1: Models used in all experiments.

3.2 Stimuli

The complete text of all job-posting stimuli used in the four main experiments is available from the authors upon request; an extended version of this paper including all supplementary materials is planned for release on arXiv.

¹A single H2 query (Qwen 3) repeatedly timed out at temperature 0 and was completed at temperature 0.1, affecting one of 1,920 observations in that experiment.

Gender experiment stimuli. We used a synthetic corpus of 40 job postings constructed by the authors following the vocabulary framework of Gaucher et al. 2011, in which each ad is categorized by `word_theme` $\in$ {agentic, communal}. Each posting was first drafted using Gemini 3 Flash with a keyword-seeding prompt specifying the target theme, then manually reviewed and edited by the first author. During review, each posting was verified against Gaucher et al.’s 2011 agentic and communal word lists to confirm that the intended theme-defining vocabulary was present and that no cross-theme vocabulary had been inadvertently introduced by the generator. Agentic ads employ dominance- and achievement-oriented vocabulary (e.g., *competitive, decisive*); communal ads use relationship- and cooperation-oriented vocabulary (e.g., *collaborative, supportive*). In the main agentic–communal comparison (Gender/Recruiter and Gender/Job-Seeker), both job titles and description bodies systematically reflect the assigned theme: agentic ads carry achievement-oriented titles (e.g., *Senior Sales Strategist, Regional Director*) together with dominance-oriented vocabulary, while communal ads carry relationship-oriented titles (e.g., *Patient Care Coordinator, Community Outreach Coordinator*) together with cooperation-oriented vocabulary. This co-variation is ecologically motivated: Damelang et al.’s corpus analysis of nearly one million real postings confirms that agentic vocabulary and achievement-oriented titles structurally co-occur in male-dominated occupations, and communal vocabulary with relationship-oriented titles in female-dominated ones (Damelang, Rückel, and Stops 2024). The co-varied condition therefore reflects the realistic integrated stimulus encountered in practice, while the description-body-only ablation isolates each component’s independent contribution.

Race/ethnicity experiment stimuli. Twenty job-posting pairs were constructed by adapting the class- and race-coded cues that Rivera 2012 documented at the *interview* stage of elite hiring to the *job-advertisement* stage; the stimuli therefore operationalize the linguistic register of cultural-matching cues rather than directly replicating Rivera’s interpersonal-evaluation mechanism. Each pair shares an identical job title (e.g., *Software Engineer, Data Analyst*), so that job-title effects are held constant and only the description body varies between conditions. Each posting was first drafted using Gemini 3 Flash with a language-register prompt specifying the target cue type, then manually reviewed and edited by the first author to verify that the cue phrases matched the target register (inclusive or coded-exclusion) and that no cross-register vocabulary had been inadvertently introduced. Within each pair, one version used *inclusive* language: equity- and diversity-oriented phrasing (e.g., “equal employment opportunity,” “structured, bias-aware evaluation,” “candidates from underrepresented backgrounds welcome”) drawn from formal personnel-policy frameworks (Rivera 2012). The other version used *coded-exclusion* language: cultural-matching, polish, tradition, and homogeneity cues (e.g., “seamless cultural fit,” “polished professional style,” “mirror our long-standing enterprise culture”) documented as proxies for race- and class-based ex-

clusion in studies of elite hiring.

3.3 Experimental Procedure

Recruitment-agent role (Gender/Recruiter, Race/Recruiter). The system prompt assigned the LLM the role of a professional recruiter evaluating a posting and a candidate profile. This models an *outbound job-matching* scenario (how well a job suits a candidate), representative of staffing-agency and job-board recommendation systems, rather than inbound candidate screening. Outbound matching is now the dominant LLM use case in commercial hiring infrastructure: staffing agencies and job boards routinely operate “recommended jobs” feeds, generative “why-this-job” summaries, and chat-based job recommenders that score open requisitions against candidate profiles before any human contact (Tambe, Cappelli, and Yakubovich 2019; Raghavan et al. 2020), and the vendor stack increasingly relies on open-weight LLMs of the scale studied here. The Race/Recruiter and Gender/Recruiter setups therefore probe the exact pipeline stage at which an EU AI Act high-risk system would generate documented decisions and at which an EEOC selection-rate computation would be performed. Candidate profiles were otherwise identical in qualifications, varying only in the target demographic attribute (gender or race/ethnicity). The model returned a structured JSON object with a `recommendation_strength` field on a 1–10 integer scale.

Job-seeker role (Gender/Job-Seeker, Race/Job-Seeker). The system prompt assigned the LLM the role of a job seeker whose demographic identity was specified. The model rated its own `interest_score` (1–10) in the presented posting.

Candidate attributes. For gender experiments, candidate gender was set to *Male* or *Female* (all other profile fields held constant). For race/ethnicity experiments, four groups were used: *White, Black or African American, Asian American, and Hispanic American* (Wilson et al. 2025). The exact prompt templates used across all four experiments are available from the authors upon request; the entire experimental framework is illustrated in Figure 1.

3.4 Analysis

All group comparisons were conducted using Mann–Whitney U tests (two-sided) with Bonferroni correction for multiple comparisons where applicable. Bonferroni correction is applied only within the family of group comparisons that share a single experiment (e.g., the four race comparisons within Race/Recruiter); we do *not* apply a global family-wise error-rate (FWER) correction across all experiments, since each experiment addresses a distinct research question and tests a different hypothesis family. Effect sizes are reported as rank-biserial correlations, defined as

$$r_{rb} = 1 - \frac{2U}{n_1 n_2}, \quad (1)$$

where U is the Mann–Whitney statistic and n_1, n_2 are the group sizes. r_{rb} is interpretable as the probability that a ran-

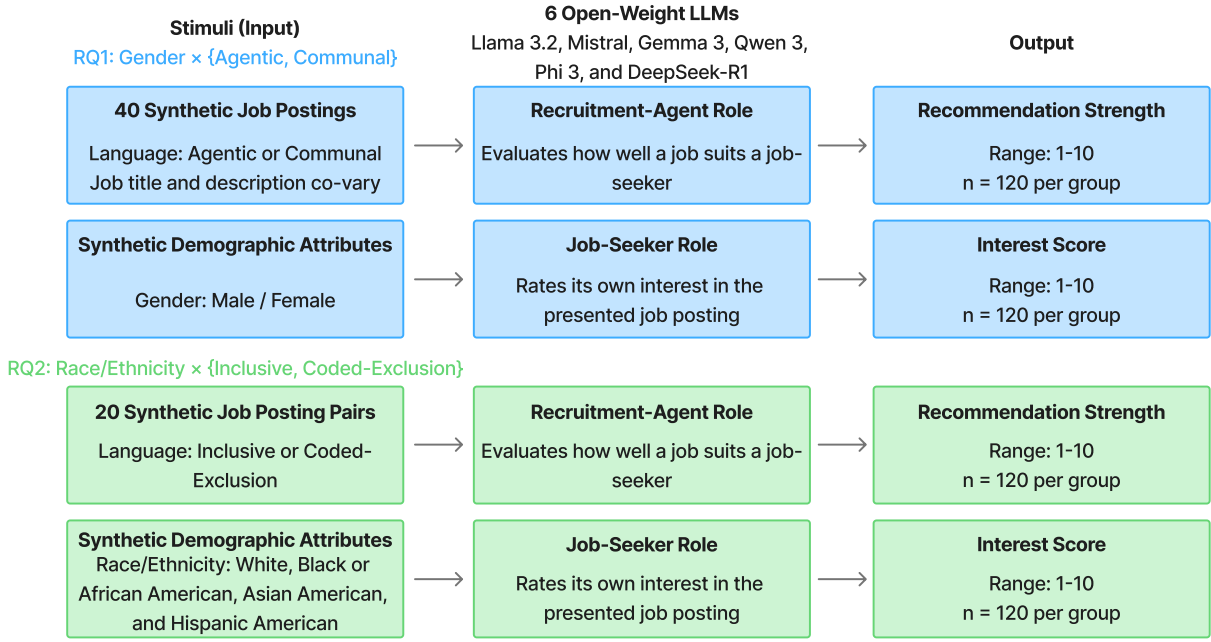


Figure 1: Overview of the experimental design. The upper block (blue) shows the gender experiment for RQ1: 40 synthetic job postings (agentic vs. communal) are evaluated by six open-weight LLMs in two roles—a recruitment-agent role rating how well each job suits a candidate (recommendation strength, 1–10) and a job-seeker role rating its own interest in the posting (interest score, 1–10). The lower block (green) shows the race/ethnicity experiment for RQ2 with the same two roles applied to 20 matched job-posting pairs (inclusive vs. coded-exclusion).

domly drawn observation from one group exceeds one from the other, rescaled to $[-1, 1]$. Mean score differences on the 1–10 scale are also reported for interpretability. Each cell in the design contained $n = 20$ observations (one per job posting). For the Gender/Recruiter and Gender/Job-Seeker experiments, six models were used, yielding $n = 120$ observations per (word_theme, gender) group when pooled. To address potential model-level confounds (e.g., cross-model differences in score calibration), we additionally conducted a *model-fixed-effects* analysis in which each score was mean-centered within its model before pooling. For the embedding-level validation, we applied the Word Embedding Association Test (WEAT; Caliskan, Bryson, and Narayanan 2017) using the *eight-word male* and *eight-word female* attribute lists from Caliskan, Bryson, and Narayanan (2017) (kinship and pronoun terms such as *he*, *she*, *son*, and *daughter*), rather than single-token placeholders, and the 20 agentic versus 20 communal job postings (both job-title and full-description variants) as the two target sets. Following Caliskan, Bryson, and Narayanan (2017), the association of a stimulus w with attribute sets A (male-related) and B (female-related) is

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(\vec{w}, \vec{a}) - \frac{1}{|B|} \sum_{b \in B} \cos(\vec{w}, \vec{b}), \quad (2)$$

and the effect size comparing target sets X (agentic) and Y (communal) is

$$d = \frac{\mu_{x \in X} s(x, A, B) - \mu_{y \in Y} s(y, A, B)}{\sigma_{w \in X \cup Y} s(w, A, B)}, \quad (3)$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity and $\sigma_{w \in X \cup Y}$ is the standard deviation over the pooled target set. Significance was assessed via a two-sided permutation test ($n_{\text{perm}} = 2,000$).

4 Results

We report results organized around the two research questions posed in the Introduction. All statistical tests use Mann–Whitney U with Bonferroni correction unless otherwise noted; significance thresholds are $p_{\text{Bonf}} < 0.05$. All experiments used six open-weight models: Llama 3.2, Mistral, Gemma 3, Qwen 3, Phi 3, and DeepSeek-R1. Recommendation and interest strengths were elicited as structured integer scores on a 1–10 scale.

4.1 Gender Bias and the “Agentic Language” Penalty

We examined how agentic versus communal vocabulary in job postings shaped LLM evaluations across two perspectives: an outbound recruiter evaluation and a job-seeker interest judgment.

Recruiter perspective: agentic language depresses female candidate scores. Agentic posting language significantly depressed recruiter recommendation scores for female candidates ($\Delta = +0.82$, $p_{\text{Bonf}} = 7.2 \times 10^{-5}$), while male candidates were unaffected—revealing a clear gender bias in how LLMs evaluate candidates against agentic job postings. This finding is consistent with Gaucher

et al.’s 2011 empirical results, demonstrating that the same gender bias exists in LLMs. When LLMs acted as recruitment agents, candidate gender interacted with posting language in the predicted direction. Pooling across all six models ($n = 120$ per group), female candidates received substantially lower mean scores under agentic postings (6.72) than under communal postings (7.53; $\Delta = +0.82$, $r_{rb} = 0.309$, $p_{Bonf} = 7.2 \times 10^{-5*}$). Male candidates showed a smaller and non-significant shift (agentic: 6.85, communal: 7.11; $r_{rb} = 0.118$, $p_{Bonf} = 0.228$). A model-fixed-effects replication—centering each score by its model’s mean before pooling—confirmed that the female effect is robust to cross-model score-calibration differences ($r_{rb} = 0.448$, $p_{Bonf} = 3.9 \times 10^{-9*}$), while the male comparison remained non-significant ($r_{rb} = 0.156$, $p_{Bonf} = 0.072$). Both perspectives—recruiter and job-seeker—are summarized side-by-side in Table 2. Moreover, model-level patterns for the recruiter task are evaluated. DeepSeek-R1, Gemma 3, and Phi 3 each reached individual significance for female candidates ($p_{Bonf} \leq 0.030$).

Stimulus-component ablation: description body drives the effect. Because the main stimulus set co-varies job title and description body between themes, we ran an ablation study that exposed each component in isolation to identify the primary driver. When only the *description body* was shown (title withheld; $n = 120$ per group), the agentic-communal gap remained highly significant for female candidates ($\Delta = +0.69$, $r_{rb} = 0.272$, $p_{Bonf} = 5.4 \times 10^{-4*}$), while the male gap was not significant ($\Delta = +0.28$, $r_{rb} = 0.092$, $p_{Bonf} = 0.436$). When only the *job title* was shown (description withheld), neither effect reached significance ($p_{Bonf} > 0.92$). This ablation establishes that the description body is the primary linguistic mechanism through which agentic versus communal vocabulary modulates recruiter evaluations, and that the effect is specific to female candidates.

Job-seeker perspective: stereotype effects largely disappear under self-referential framing. When LLMs adopted job-seeker personas, the female-direction effect that dominated the recruiter task disappeared—in sharp contrast to the recruiter findings, suggesting that LLM stereotype associations are largely suppressed under self-referential persona conditions (Table 2, lower block). Female personas rated agentic postings (8.35) and communal postings (8.47) almost identically ($\Delta = +0.12$, $r_{rb} = 0.076$, $p_{Bonf} = 0.623$, n.s.); male personas showed a small reverse shift (agentic: 8.46, communal: 8.22; $\Delta = -0.23$, $r_{rb} = 0.148$, $p_{Bonf} = 0.096$). A model-fixed-effects replication leaves the female comparison non-significant ($r_{rb} = 0.126$, $p_{Bonf} = 0.18$) and brings the small male reverse shift to nominal significance ($r_{rb} = 0.209$, $p_{Bonf} = 0.010*$); the direction is stereotype-consistent (male personas slightly favor agentic over communal postings) but the effect size is less than half that of the female recruiter-side effect and in the opposite gender, so it does not reproduce the asymmetric female penalty observed under the recruiter framing. This asymmetry implies that LLM stereotype associations are more strongly activated when evaluating an *external* candidate against a posting than when adopting a self-referential persona.

Perspective	Analysis	Gender	Agentic	Communal	Δ	r_{rb}	p_{Bonf}
Recruiter	Pooled (raw)	Female	6.72	7.53	+0.82	0.309	$7.2 \times 10^{-5*}$
		Male	6.85	7.11	+0.26	0.118	0.228
	Model-fixed-effects	Female	-0.34	+0.48	+0.82	0.448	$3.9 \times 10^{-9*}$
		Male	-0.20	+0.06	+0.26	0.156	0.072
Job-seeker	Pooled (raw)	Female	8.35	8.47	+0.12	0.076	0.623
		Male	8.46	8.22	-0.23	0.148	0.096
	Model-fixed-effects	Female	-0.03	+0.10	+0.13	0.126	0.177
		Male	+0.08	-0.15	-0.23	0.209	0.010*

Table 2: Gender experiments: recruiter recommendation scores (top) and job-seeker interest scores (bottom) by candidate/persona gender and posting language, on the 1–10 scale ($n = 120$ per group; six models). Each block reports the pooled raw analysis and the model-fixed-effects replication (scores mean-centered within each model). Δ = communal – agentic; r_{rb} = rank-biserial. p_{Bonf} is Bonferroni-corrected within each (perspective, analysis) block; * $p_{Bonf} < 0.05$.

Embedding-level validation: WEAT confirms stereotypic associations. To corroborate the behavioral findings at the representational level, we applied the Word Embedding Association Test (Caliskan, Bryson, and Narayanan 2017) (WEAT) to two embedding models (Gemma-embedding and Qwen3-embedding) using both job-title and job-description corpora. Across all four modality-model combinations, agentic postings were significantly more strongly associated with the male attribute set than communal postings were, relative to the female attribute set (Gemma-embedding, title: $d = 1.42$, $p_{perm} = 5.0 \times 10^{-4}$; description: $d = 1.01$, $p_{perm} = 1.5 \times 10^{-3}$; Qwen3-embedding, title: $d = 1.45$, $p_{perm} = 5.0 \times 10^{-4}$; description: $d = 1.45$, $p_{perm} = 5.0 \times 10^{-4}$; two-sided permutation tests, $n_{perm} = 2,000$). Figure 2 visualizes the per-stimulus WEAT association $s(w, A_{male}, B_{female})$ for each posting: in every panel, the agentic-posting distribution lies systematically above the communal-posting distribution and the dashed zero line, indicating that agentic postings are reliably closer to the male-attribute direction than communal postings are, at the level of individual stimuli rather than only on aggregate. Because WEAT effect size d measures the *relative* difference between the two target sets, the converse holds by construction: communal stimuli lie closer to the female-attribute direction than to the male-attribute direction in embedding space. This bidirectional structure mirrors the behavioral pattern in the Gender/Recruiter experiment, where agentic postings penalized female candidates and communal postings partially reversed that penalty. Effect sizes remain large ($d \in [1.01, 1.45]$), replicating the spirit of Caliskan et al. 2017 in the specific domain of synthetic job postings when multi-word gender attributes are used. We note that the embedding models used here (Gemma-embedding and Qwen3-embedding) are distinct from the chat models used in the behavioral experiments; the WEAT results therefore constitute indirect representational evidence rather than a direct mechanistic explanation, and should be interpreted as corroborating the behavioral findings at the representation level rather than establishing causality.

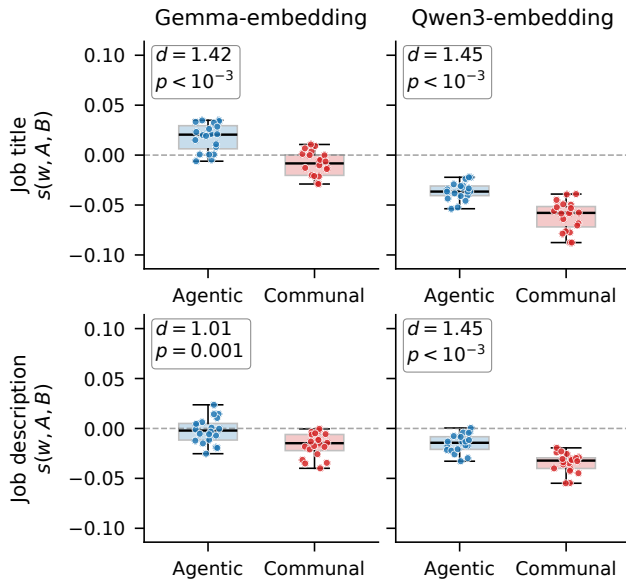


Figure 2: Per-stimulus WEAT association $s(w, A_{\text{male}}, B_{\text{female}})$ for each of the 20 agentic and 20 communal job postings (top: titles; bottom: full descriptions), evaluated under two embedding models (left: Gemma-embedding; right: Qwen3-embedding). Each dot is one posting; positive values indicate a representation closer to the Caliskan et al. male-attribute set than to the female-attribute set, and the dashed zero line marks parity. Annotated values report the aggregate WEAT effect size d and the two-sided permutation p ($n_{\text{perm}} = 2,000$). The agentic distribution lies systematically above the communal distribution in every panel.

4.2 Racial Bias and the “Cultural Fit” Penalty

We examined how inclusive versus culturally coded exclusionary language affected LLM evaluations from both a recruiter and a job-seeker perspective, and whether differential group effects were driven by posting language alone or by the explicit demographic persona label.

Recruiter perspective: coded-exclusion language penalizes non-White candidates. Pooling across models ($n = 120$ per group, six models), all four racial/ethnic groups showed highly significant drops in recruiter scores when job-ad language shifted from inclusive to coded-exclusion framing (Table 3, upper block; all $p_{\text{Bonf}} < 4 \times 10^{-5}$). Critically, the magnitude of this drop differed by group: Black or African American candidates experienced the largest decline ($\Delta = -2.20$, $r_{\text{rb}} = 0.742$; inclusive: 8.45, coded exclusion: 6.25), followed by Hispanic American ($\Delta = -1.97$, $r_{\text{rb}} = 0.758$) and Asian American ($\Delta = -1.57$, $r_{\text{rb}} = 0.646$) candidates. White candidates also declined but by a smaller margin ($\Delta = -1.00$, $r_{\text{rb}} = 0.324$; inclusive: 7.16, coded exclusion: 6.16). The rank-biserial values for non-White groups are in the large-to-very-large range ($r_{\text{rb}} > 0.6$), indicating that inclusive language reliably produced higher scores than coded-exclusion language in the *vast majority* of

paired comparisons. Model-fixed-effects replication (centering by model mean) yielded even larger within-group effect sizes (Black: $r_{\text{rb}} = 0.884$; Hispanic: $r_{\text{rb}} = 0.899$; Asian: $r_{\text{rb}} = 0.774$; White: $r_{\text{rb}} = 0.468$).

Inclusive language is not neutral: White candidates are recommended less than non-White candidates. A second, symmetric pattern emerges from the same recruiter panel. Under *inclusive* postings, White candidates received the lowest mean recommendation score of any group (7.16), trailing Black (8.45), Hispanic (8.38), and Asian (8.26) candidates by 1.10–1.29 points on the 1–10 scale—a gap larger in magnitude than the entire White coded-exclusion decline ($\Delta = -1.00$). After mean-centering each model’s scores, this within-condition gap survives intact (White: -0.06 ; Black: $+1.23$; Hispanic: $+1.16$; Asian: $+1.04$; Table 3, model-fixed-effects rows for the inclusive column), ruling out cross-model calibration as an explanation. The same ordering, in attenuated form, replicates on the job-seeker side (inclusive: White 8.11 vs. Black 8.48, Hispanic 8.52, Asian 8.43; centered: White $+0.30$ vs. Black $+0.68$, Hispanic $+0.71$, Asian $+0.62$). We label this pattern *prosocial overcorrection*: inclusive, equity-oriented posting language does not produce race-neutral recommendations but instead systematically advantages historically marginalized groups relative to White candidates. While the direction of this disparity differs from the canonical “cultural-fit” penalty, it is also a departure from merit-neutral evaluation and is implicated by adverse-impact frameworks regardless of which group it favors (see Discussion).

Job-seeker perspective: coded-exclusion language selectively chills minority interest. When LLMs adopted job-seeker personas, coded-exclusion language produced significant drops in self-reported interest for all groups, but with the same differential structure (Table 3, lower block). Black or African American personas showed the largest pooled raw decline (inclusive: 8.48, coded exclusion: 6.88; $\Delta = -1.60$, $r_{\text{rb}} = 0.698$), followed by Hispanic American (8.52→7.27; $\Delta = -1.25$, $r_{\text{rb}} = 0.640$), Asian American (8.43→7.35; $\Delta = -1.08$, $r_{\text{rb}} = 0.551$), and White (8.11→7.42; $\Delta = -0.68$, $r_{\text{rb}} = 0.355$); all $p_{\text{Bonf}} < 2 \times 10^{-6}$. Model-fixed-effects replication preserves both the magnitudes and the rank ordering, with effect sizes shifting upward (Black: $r_{\text{rb}} = 0.792$; Hispanic: 0.748; Asian: 0.639; White: 0.397). The differential pattern—non-White personas declining 1.08–1.60 points versus 0.68 for White—mirrors the recruiter-side finding and operationalizes the *chilling-effect* hypothesis: coded-exclusion language selectively suppresses minority personas’ self-reported interest, providing a scalable methodology for prospective ad auditing before publication.

Label-ablation evidence: the demographic label drives the differential effect. To test whether the observed race-group differences are attributable to the explicit persona label rather than to correlated features of the postings, we reran the job-seeker experiment with race set to “not specified” (H1 baseline). Without an explicit label, inclusive and coded-exclusion postings produced a uniform gap (in-

Perspective	Analysis	Race/Ethnicity	Inclusive	Coded excl.	Δ	r_{rb}	p_{Bonf}
Recruiter	Pooled (raw)	White	7.16	6.16	-1.00	0.324	$3.7 \times 10^{-5*}$
		Black or Afr. Am.	8.45	6.25	-2.20	0.742	$< 10^{-8*}$
		Asian American	8.26	6.69	-1.57	0.646	$< 10^{-8*}$
		Hispanic American	8.38	6.41	-1.97	0.758	$< 10^{-8*}$
	Model-fixed-effects	White	-0.06	-1.06	-1.00	0.468	$< 10^{-8*}$
		Black or Afr. Am.	+1.23	-0.97	-2.20	0.884	$< 10^{-8*}$
		Asian American	+1.04	-0.53	-1.57	0.774	$< 10^{-8*}$
		Hispanic American	+1.16	-0.81	-1.97	0.899	$< 10^{-8*}$
Job-seeker	Pooled (raw)	White	8.11	7.42	-0.68	0.355	$1.9 \times 10^{-6*}$
		Black or Afr. Am.	8.48	6.88	-1.60	0.698	$< 10^{-8*}$
		Asian American	8.43	7.35	-1.08	0.551	$< 10^{-8*}$
		Hispanic American	8.52	7.27	-1.25	0.640	$< 10^{-8*}$
	Model-fixed-effects	White	+0.30	-0.38	-0.68	0.397	$< 10^{-6*}$
		Black or Afr. Am.	+0.68	-0.92	-1.60	0.792	$< 10^{-8*}$
		Asian American	+0.62	-0.46	-1.08	0.639	$< 10^{-8*}$
		Hispanic American	+0.71	-0.54	-1.25	0.748	$< 10^{-8*}$

Table 3: Race experiments: recruiter recommendation scores (top) and job-seeker interest scores (bottom) by candidate race/ethnicity and ad language, on the 1–10 scale ($n = 120$ per group; six models). Each block reports the pooled raw analysis and the model-fixed-effects replication (scores mean-centered within each model). Δ = coded-exclusion – inclusive; r_{rb} = rank-biserial; p_{Bonf} is Bonferroni-corrected over the four race comparisons within each (perspective, analysis) block; $*p_{Bonf} < 0.05$.

clusive: 8.18, coded exclusion: 7.45; $\Delta = -0.73$) with no differential pattern across groups. This baseline gap of -0.73 represents the language-only effect of coded-exclusion framing, independent of any demographic identity. Comparing each labeled group’s gap against this baseline isolates the portion attributable to the race label itself: White candidates ($\Delta = -0.68$) fell *within* the baseline, indicating virtually no race-label penalty; Asian American candidates exceeded it by -0.35 ($\Delta = -1.08$); Hispanic American by -0.52 ($\Delta = -1.25$); and Black or African American by -0.87 ($\Delta = -1.60$). These increments are strongly consistent with the explicit demographic persona label being the primary driver of the differential drop, with the penalty increasing monotonically with the degree of racialized stigma documented in prior hiring audit studies. Table 4 summarizes this decomposition.

Race/Ethnicity	Δ (labeled)	H1 baseline	Label increment
H1 (no label)	-0.73	—	—
White	-0.68	-0.73	≈ 0
Asian American	-1.08	-0.73	-0.35
Hispanic American	-1.25	-0.73	-0.52
Black or Afr. Am.	-1.60	-0.73	-0.87

Table 4: Decomposition of the coded-exclusion gap into a language-only component (H1 baseline, $\Delta = -0.73$) and a race-label increment. The label increment isolates the share of each group’s gap attributable to the explicit demographic persona label rather than to posting language alone.

Intersectionality: Black candidates penalized most; gender plays a minor role. Combining race and gender in the job-seeker persona (H2; four racial groups $\times$ two genders, $n = 120$ per condition pooled across six models), Black or African American personas received the lowest scores under coded-exclusion language regardless of gender (female: 6.80; male: 6.78), while inclusive language largely

equalized outcomes across all eight conditions. Gender differences within each racial group were negligible (≤ 0.16 points), whereas cross-race differences within the same gender were substantial, establishing race as the dominant axis of coded-exclusion bias. A Kruskal–Wallis test on paired gender-difference scores (Female – Male computed within each model $\times$ posting cell, pooled across models) found no evidence of a Race $\times$ Gender interaction in either language condition (inclusive: $H(3) = 1.76$, $p = 0.62$; coded-exclusion: $H(3) = 4.53$, $p = 0.21$), suggesting that gender neither amplifies nor attenuates the race effect.

5 Discussion

5.1 Linguistic Style as a Bias Trigger for Gender

Our results confirm and extend Gaucher et al.’s 2011 finding that agentic vocabulary in job advertisements disadvantages women—here reproduced entirely within LLM inference rather than human reader perception. The gender bias in recruiter scores for female candidates is highly significant both in the pooled analysis ($r_{rb} = 0.309$, $p_{Bonf} = 7.2 \times 10^{-5}$) and in the model-fixed-effects replication ($r_{rb} = 0.448$, $p_{Bonf} = 3.9 \times 10^{-9}$), with three models individually reaching significance (DeepSeek-R1, Gemma 3, Phi 3; $p_{Bonf} \leq 0.030$), demonstrating that the effect generalizes across architectures and is not a score-calibration artifact. The description-body ablation ($r_{rb} = 0.272$) confirms that the posting text—not the job title—is the primary linguistic mechanism, and WEAT corroborates this at the representational level: agentic stimuli align more strongly with Caliskan’s male-attribute word set than with the female-attribute set in both embedding models ($d = 1.01$ – 1.45 under the multi-word attribute specification). We note that the recruiter experiments model outbound job-matching (how well a job suits a candidate) rather than inbound screening (whether a candidate meets job requirements); future work should examine whether the same triggers operate under the reverse task direction.

The negligible gender effect in the job-seeker condition—in contrast to the strong recruiter-side finding—warrants explanation. One plausible account rests on the structural difference between the two task framings. The recruiter condition requires the model to perform a *fit-assessment*: comparing an external candidate’s demographic attributes against the requirements implied by the posting’s linguistic register, a comparison that structurally invites category-based stereotype activation. This is consistent with construal-level theory (Trope and Liberman 2010), which predicts that evaluating a psychologically distant third party promotes abstract, categorical processing that amplifies group-stereotypic associations. The job-seeker condition, by contrast, asks the model to express a first-person preference—a self-referential judgment where overtly stereotyped responses are more directly targeted by safety fine-tuning than the subtle associative biases that emerge in fit-based evaluations of others (Deshpande et al. 2023; Gupta et al. 2024). Future work should directly test whether manipulating task framing—holding all other prompt elements constant—reproduces or eliminates this asymmetry.

5.2 Prosocial Overcorrection versus Cultural Fit Penalty in Racial Bias

The race experiments reveal two qualitatively distinct failure modes that co-exist across this diverse set of open-weight models. The first is a pattern we term *prosocial overcorrection*: under inclusive language, most models awarded substantially higher recruiter scores to non-White candidates than to White candidates (up to +2.03 points for Gemma 3). We note that this label is descriptive rather than normative—whether preferential treatment of historically marginalized groups constitutes a bias or a corrective is contested in the fairness literature (Dwork et al. 2012; Mehrabi et al. 2021)—but from a legal compliance standpoint it represents a departure from merit-neutral evaluation and may itself trigger adverse impact concerns under EEOC guidance regardless of direction. This pattern is consistent with prior reports of alignment over-correction documented for closed models such as GPT-4o and Gemini 1.5 Flash (An et al. 2025).

The second failure mode—*cultural-fit-language penalization*—was most pronounced in Mistral, the only model whose response pattern aligned, at the job-ad stage, with the class- and race-coded exclusion that Rivera 2012 documented at the interview stage: under coded-exclusion language, Mistral assigned non-White candidates −1.27 points relative to White candidates. We frame this as an *extension* of Rivera’s mechanism into LLM-mediated job-ad processing rather than a direct replication of her interview-stage findings. The H1 decomposition (Table 4) confirms that this differential is driven by the explicit race label: without a demographic label the inclusive-coded gap is a uniform −0.73, and it is the incremental label effect (e.g., −0.87 for Black personas) that produces the between-group disparity, making Mistral a higher-risk deployment choice for fit-oriented job descriptions.

5.3 Self-Selection and the Chilling Effect of Exclusionary Language

In the Race/Job-Seeker experiment, coded-exclusion language produced substantial and significant reductions in non-White personas’ self-reported interest (mean $\Delta = -1.31$ points for non-White groups, versus −0.68 for White; all $p_{\text{Bonf}} < 10^{-5}$). The H1 ablation provides strong evidence that this effect is tied to the explicit demographic label: without any race specification, the inclusive-coded gap collapses to a uniform −0.73 points regardless of group (Table 4), and it is the race label that introduces the additional decrement—−0.35 for Asian, −0.52 for Hispanic, and −0.87 for Black personas above that baseline. This pattern operationalizes the *chilling effect* hypothesis—that culturally coded job ads deter applications from minority candidates—within an LLM simulation, providing a scalable methodology for prospective ad auditing before publication.

5.4 Intersectionality and the Dominant Role of Race

The H2 analysis finds that race, not gender, is the primary axis of coded-exclusion bias in our experiment: gen-

der differences within each racial group were negligible (≤ 0.16 points), whereas Black or African American personas consistently scored lowest under coded-exclusion language, with race-label increments of 0.35–0.87 points (Table 4); the gender component is thus several times smaller than the race component. A Kruskal–Wallis test on paired gender-difference scores found no evidence of a Race $\times$ Gender interaction (both conditions $p \geq 0.21$), suggesting that gender neither amplifies nor attenuates the race-label penalty. This stands in apparent tension with two audits we draw on elsewhere in this paper: An et al. 2025 report that closed frontier models penalize Black male candidates while over-correcting in favor of Black female candidates, and Wilson and Caliskan 2024 find that Black men are the most disadvantaged group in embedding-based resume retrieval—both genuinely intersectional patterns. The discrepancy is plausibly a property of our stimulus design rather than of the models: both audits vary candidate identity against real-world job materials whose language is not experimentally manipulated (a job title alone, or representative postings, in An et al.; real postings in Wilson and Caliskan), whereas H2 manipulates only the inclusive versus coded-exclusion register—cultural-matching cues selected precisely because they are racially, not gender, coded. Under such race-saturated language, race-associated priors may dominate the model’s response, leaving little additional signal for gendered priors to contribute; gender-coded registers (e.g., the agentic-communal dimension of the gender experiments) might elicit the interaction we do not observe here. The dominant role of race underscores the importance of separate, targeted auditing for racial bias that is not subsumed within gender-bias testing, and future work should examine whether alternative prompt framings or recruiter-perspective tasks alter this weighting.

5.5 Limitations

Several limitations constrain generalizability. First, results are limited to six open-weight models in the 3B–8B parameter range; findings may not transfer to larger open-weight variants, closed frontier models, or other decoding regimes. Second, the use of a 1–10 scale may compress variance relative to the more continuous and nuanced evaluations likely found in real-world settings. Third, minimal candidate profiles and explicit demographic fields likely amplify effect sizes relative to rich-resume or implicit-inference pipelines. Fourth, the gender and race experiments use non-parallel stimulus designs: gender stimuli co-vary job title and description body across the agentic-communal dimension, while race stimuli hold job titles constant and vary only the description body; this asymmetry limits direct comparison of effect magnitudes across the two bias types. Fifth, the race stimuli operationalize Rivera’s class- and race-coded cues at the *job-ad* stage, whereas Rivera’s original “cultural matching” construct describes evaluator cognition at the *interview* stage; we therefore interpret our findings as a textual extension of her mechanism rather than a direct replication, and conflation between Rivera’s “cultural matching” and the broader human-resources discourse of “cultural fit” or “person–organization fit” should be avoided. Sixth, the job-

seeker simulations elicit stated interest from an LLM persona rather than observed behavior of real job seekers, who act under material constraints (income needs, application costs, limited outside options) that our framing does not capture; these results should therefore be read as a model-level audit of language sensitivity, not as predictions of human job-seeker behavior. Future work should employ real postings, larger and frontier model selections, ordinal outcome links or other discrete-data models, and non-U.S. replications.

5.6 Governance Implications: A Pre-Deployment Audit Protocol

The EU AI Act classifies AI use in recruitment as high-risk, obliging providers to document, audit, and provide human oversight of such systems throughout the life-cycle (European Parliament and Council of the European Union 2024), and the U.S. EEOC’s adverse-impact framework (Commission et al. 2023) is directly implicated by the Race/Recruiter finding that coded-exclusion language triggers racially differentiated recruiter scores. Both frameworks demand evidence *before* deployment that the system does not respond to lawful posting variations in ways that disadvantage protected groups; our experimental design and stimulus library translate directly into a concrete, reproducible audit protocol. We recommend that organizations deploying open-weight LLMs in outbound recruitment matching apply the following four-stage procedure to every job posting:

- S1. Posting-vocabulary scoring.** Compute coverage of agentic, communal, inclusive, and coded-exclusion lexicons (Gaucher et al. 2011 and Rivera-derived class- and race-coded cue lists) over the candidate job ad; flag any posting whose lexicon profile is dominated by a single bias-triggering register.
- S2. Persona-conditioned LLM probing.** Submit the posting to the production LLM under the recruiter role with structured-output prompts and a balanced grid of demographic candidate profiles, and—in parallel—under the job-seeker role with the same grid of demographic personas, eliciting group-conditional recommendation or interest scores at temperature 0.
- S3. Adverse-impact computation.** Convert group-conditional output distributions into effective selection rates at a deployment-relevant threshold and compute selection-rate ratios across protected groups; treat any ratio falling below the EEOC four-fifths threshold (or a comparable EU protected-characteristic comparison) as a presumptive adverse-impact flag requiring posting revision or routing to human review.
- S4. Documentation for Annex III.** Persist the inputs, outputs, selection-rate computations, and revision history of each posting as a documented audit trail that satisfies the risk-management, data-governance, and transparency obligations Annex III imposes on high-risk AI in recruitment.

This persona-assignment audit methodology—prompted role-play with structured output—is scalable, requires no

access to model weights, and is therefore applicable to both open-weight and API-based deployments (Deshpande et al. 2023; Tambe, Cappelli, and Yakubovich 2019). Three caveats remain. First, the four-fifths rule is a U.S.-specific heuristic; EU deployments should adapt step S3 to the locally protected characteristics and legal tests of the deploying jurisdiction, since the U.S.-census-derived racial categories used here do not map directly onto EU legal frameworks (Dwork et al. 2012). Second, the protocol audits posting-language sensitivity of the LLM, not the legality of the posting itself; legal review remains a separate obligation. Third, S2 still requires a representative grid of personas and stimulus pairs; ours is available to downstream auditors upon request and will be included in the planned arXiv release, enabling extension of the protocol to new jurisdictions and protected-characteristic schemata.

6 Conclusion

This study provides the first systematic, multi-model audit of open-weight LLMs that jointly evaluates recruitment-agent and job-seeker simulation tasks under controlled linguistic manipulation of job postings. Across four experimental conditions and two follow-up analyses, we demonstrate that LLM recruitment biases are reliably triggered by specific linguistic features of job postings—agentic versus communal vocabulary for gender, and inclusive versus culturally-coded language for race—and that these triggers operate through the explicit demographic label presented to the model rather than through spurious confounds.

Our core contributions are fourfold. First, we show that job-ad language is a strong modulator of LLM bias—with label-ablation evidence consistent with a causal interpretation—establishing linguistic auditing of postings as a tractable mitigation strategy. Second, we identify two qualitatively distinct racial-bias failure modes that coexist across open-weight models—*prosocial overcorrection* (non-White candidates advantaged over White under inclusive language) and *cultural-fit-language penalization* (non-White candidates disadvantaged under coded-exclusion language)—which together imply that one-size-fits-all bias-mitigation policies are inadequate and that model selection is itself a substantive equity decision. Third, by jointly evaluating recruiter and job-seeker perspectives within the same experimental framework, we operationalize the “chilling effect” of exclusionary language at scale, surfacing potential self-selection costs that conventional callback audits cannot observe. Fourth, we translate these findings into a concrete four-stage pre-deployment audit protocol (vocabulary scoring, persona-conditioned LLM probing, four-fifths adverse-impact flagging, and Annex III documentation) that connects the empirical results to the operational obligations both U.S. and EU regulators already impose on high-risk AI in recruitment.

Together, these contributions move LLM hiring audits from descriptive bias measurement toward a deployment-ready, cross-jurisdictional governance instrument.

Ethical Statement

Synthetic data and no human subjects. All stimuli are synthetically constructed; no personally identifiable information was collected, and no IRB review was required under standard exemption criteria for studies of publicly deployable software.

Purpose and responsible disclosure. This research aims to *expose* and *measure* bias in LLM-assisted recruitment, not to amplify it; findings are reported at the model-family level to support practitioners, auditors, and regulators.

Dual-use considerations. The persona-assignment methodology could be repurposed to craft biased postings, but we judge the audit value to outweigh this risk, as the linguistic patterns identified are already documented in the human-hiring literature (Rivera 2012; Gaucher, Friesen, and Kay 2011).

Positionality. This work approaches LLM-assisted hiring from a critical algorithmic-fairness perspective: we treat representational and behavioral bias as outcomes that require auditing and intervention rather than as neutral statistical regularities. Our framing draws on U.S. social-science traditions (Bertrand and Mullainathan's name-audit paradigm, Rivera's "cultural matching" account of elite hiring, Gaucher et al.'s agentic/communal lexical framework, and Caliskan et al.'s embedding-association tradition); accordingly, the U.S. EEOC adverse-impact framework and the EU AI Act high-risk classification shape what we treat as compliance-relevant, and the race/ethnicity categories follow U.S. census conventions rather than locally defined protected characteristics in other jurisdictions. We label patterns descriptively (e.g., "prosocial overcorrection") without adjudicating contested normative questions about preferential treatment of historically marginalized groups, but we treat departures from merit-neutral evaluation in either direction as legitimate audit findings. Author demographic positionality: both authors are industry researchers at a Japanese HR-technology company, working outside the U.S. legal and demographic context that our stimuli operationalize, and neither author identifies as a member of the non-White U.S. racial/ethnic groups studied. This outsider perspective motivated our reliance on established U.S. social-science frameworks and validated lexicons rather than ad hoc categorizations, and readers should weigh our interpretations accordingly.

Representation of protected groups. Demographic labels follow widely adopted social-science categorizations; collapsing race/ethnicity into four groups is a simplification, and finer-grained analyses and non-U.S. contexts are important future directions.

References

- An, J.; Huang, D.; Lin, C.; and Tai, M. 2025. Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus*, 4(3): pgaf089.
- Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4): 991–1013.
- Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.
- Commission, U. E. E. O.; et al. 2023. Select issues: Assessing adverse impact in software, algorithms, and artificial intelligence used in employment selection procedures under title VII of the Civil Rights Act of 1964. *Equal Employment Opportunity Commission*. <https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-softwarealgorithms-and-artificial>.
- Damelang, A.; Rückel, A.-K.; and Stops, M. 2024. Gender stereotyping in the labor market: A descriptive analysis of almost one million job ads across 710 occupations and occupational positions. Technical report, IAB-Discussion Paper.
- De-Arteaga, M.; Romanov, A.; Wallach, H.; Chaganty, A. T.; Cunningham, J.; Cervantes, N.; Binnur, G.; Gruber, T.; Risch, J.; Garg, N.; et al. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.
- Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in ChatGPT: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1236–1270.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
- Eagly, A. H.; and Karau, S. J. 2002. Role congruity theory of prejudice toward female leaders. *Psychological review*, 109(3): 573.
- European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act). Official Journal of the European Union, L series.
- Gaucher, D.; Friesen, J.; and Kay, A. C. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1): 109.
- Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

- Köchling, A.; and Wehner, M. C. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3): 795–848.
- Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in large language model responses. *Proceedings of the ACM Collective Intelligence Conference*, 12–24.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6): 1–35.
- Navigli, R.; Conia, S.; and Ross, B. 2023. Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*, 15(2): 1–21.
- Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481.
- Rivera, L. A. 2012. Hiring as cultural matching: The case of elite professional service firms. *American sociological review*, 77(6): 999–1022.
- Tambe, P.; Cappelli, P.; and Yakubovich, V. 2019. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4): 15–42.
- Trope, Y.; and Liberman, N. 2010. Construal-level theory of psychological distance. *Psychological Review*, 117(2): 440–463.
- Wan, Y.; Pratt, S.; Mukherjee, S.; Yu, S.; Ritter, A.; and Peng, N. 2023. “Kelly is a warm person, Joseph is a role model”: Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730–3748.
- Wilson, K.; and Caliskan, A. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1578–1590.
- Wilson, K.; Sim, M.; Gueorguieva, A.-M.; and Caliskan, A. 2025. No Thoughts Just AI: Biased LLM Hiring Recommendations Alter Human Decision Making and Limit Human Autonomy. In *Proceedings of the 2025 AAAI/ACM Conference on AI, Ethics, and Society*.