CIDE

LATIN AMERICAN ECONOMIC REVIEW

2026 | Volume 37 | Article 6

Gender Bias in Automated CV Evaluation: Evidence from Counterfactual Simulations Using Synthetic Data from Mexico

Edgar Cruz[‡] Alejandro T. Moreno-Okuno[†] Johanan Zamilpa[§]

Submission received: June 17, 2025
Final version received: April 19, 2026
Accepted: April 28, 2026
Published: May 14, 2026

Abstract

This paper investigates the presence of demographic bias in automated CV evaluations conducted by a large language model (LLM). We generate a universe of over 14,000 synthetically constructed CVs representative of the Mexican labor market across six occupational categories and implement a counterfactual design that isolates the effect of perceived gender by switching only the candidate's name and reported gender, holding all remaining CV content fixed. The results suggest systematic and occupation-specific asymmetries: female candidates receive higher scores when presented as male in traditionally masculine roles (e.g., Truck Driver), while male candidates gain when reclassified as female in feminized occupations (e.g., Nursing, Primary School Teaching). Notably, we document a statistically significant and operationally meaningful pro-female bias in the high-status Chief Financial Officer role. In a complementary counterfactual exercise, we mechanically vary reported age while keeping the full CV content unchanged, showing that age sensitivity is likewise occupation-specific and markedly non-linear. Our design is contextually grounded, using names, educational institutions, and employers common in Mexico, and offers a scalable methodology for localized bias auditing in LLM-based screening. The findings highlight the necessity of context-specific fairness assessments and raise concerns about the equity implications of deploying general-purpose AI tools in personnel selection.

Keywords: Gender bias, Automated CV screening, Large language models, Synthetic data experiments.
JEL codes: J71, C93, O33, O54.

[‡]Departamento de Economía y Finanzas, Universidad de Guanajuato. Corresponding author.
Email: be.cruz@ugto.mx.
[†]Departamento de Economía y Finanzas, Universidad de Guanajuato.
Email: atatsuo@ugto.mx.
[§]Departamento de Gestión y Dirección de Empresas, Universidad de Guanajuato.
Email: johanan.zamilpa@ugto.mx.

1. Introduction

There is a rapid integration of artificial intelligence (AI) into personnel selection processes, which has raised substantial concerns about the perpetuation and amplification of discriminatory biases, particularly those based on gender and race.[1] Specifically, a growing body of evidence suggests that large language models (LLMs) are not neutral when deployed in hiring tasks. [2] For instance, Kotek et al. (2023) document that LLMs reproduce occupational stereotypes, disproportionately favoring male candidates in masculinized professions and female candidates in feminized roles when CVs are synthetically constructed to be otherwise comparable. Armstrong et al. (2024) find that GPT-4 systematically favors applicants identified as women, with effects strengthening when multiple identity cues, such as names and pronouns, jointly signal female identity. Beyond gender, Lippens (2024) reports substantial disparities in interview recommendations across ethnic and national-origin profiles, with GPT-based systems being significantly less likely to recommend interviews for Arab, Asian, Black American, Central African, and Eastern European candidates than for white Western profiles holding qualifications fixed. These findings suggest that LLM-based screening can act on demographic signals, such as gender and age, thereby reshaping shortlisting outcomes.

However, the existing evidence focuses on high-income economies, where labor market institutions, digital infrastructure, and algorithmic governance differ substantially from conditions in developing countries. Because informality, segmentation, and weaker enforcement may alter both deployment and decision rules, it is unclear whether high-income evidence generalizes to developing-country contexts. This question is no longer merely speculative: the limited evidence available suggests that AI-enabled screening and matching tools are increasingly adopted beyond high-income settings and into low- and middle-income labor markets, including Latin America. In particular, Mexico provides a particularly salient case among middle-income countries. While adoption is likely heterogeneous, concentrated among larger formal-sector employers and platform-mediated hiring channels, available survey and job-platform evidence is consistent with early AI use in recruitment and selection, with HR among the main areas of implementation. For example, Computrabajo's study on AI in job search reports that roughly four in ten firms in Mexico already use AI, and that about half of these deploy it within HR for recruitment, selection, and evaluation (Escutia, 2025).[3]

This paper addresses this gap by providing an empirical assessment of gender-related algorithmic bias in CV evaluation using LLMs in the context of the Mexican labor market. We study a setting in which informality, institutional segmentation, and limited regulatory oversight shape not only the governance surrounding algorithmic hiring, but also the margin of adoption, likely concentrated among larger formal-sector employers and platform-mediated recruitment channels, where small score differences can translate into meaningful shortlisting gaps. Under this institutional background, we specify two working hypotheses. First, if LLM-based screening internalizes gender cues as predictive signals, then holding résumé content fixed, candidate scores will differ systematically by reported gender (H1: non-neutrality with respect to gender). Second, consistent with stereotype-based statistical discrimination, the sign and magnitude of the gender effect will vary

[1] Society for Human Resource Management (2022) reports that 64% of HR professionals say their organization's automation or AI tools automatically filter out unqualified applicants (question: select all that apply), underscoring the scale of adoption and the potential risks associated with algorithmic evaluations in the US.

[2] While our focus is on LLM-based screening, concerns about discrimination in automated hiring predate LLMs. Chang (2023) audits proprietary screening tools trained on historical data and documents penalties for résumés containing gendered terms (e.g., "women's"), consistent with the view that institutional datasets can encode bias that is subsequently operationalized by algorithmic selection systems.

[3] Note that these figures on AI use in recruitment and selection come from platform-based surveys and reports rather than official labor statistics. Therefore, it should be interpreted as indicative (self-reported) evidence, likely tilted toward large formal-sector employers and digitally mediated hiring channels. Complementing to Computrabajo's report, an Indeed Mexico survey (IDC Online, 2023) reports that a large share of recruiters perceive immediate productivity gains from AI (e.g., saving time in repetitive tasks and improving candidate sourcing), and identifies concrete use cases already embedded in recruitment workflows, such as search optimization, CV screening, and interview scheduling support.

with occupational gender typing, reflecting perceived "fit" between gender-coded cues and job archetypes (H2: occupation-specific stereotyping). Consequently, our analysis targets a pressing policy concern: the extent to which AI-based hiring tools may reinforce or disrupt gender inequalities in access to formal employment opportunities in developing countries.

To this end, we implement a novel paired-evaluation experiment based on synthetically constructed CVs that closely mirror real candidate profiles observed in the Mexican labor market. Each CV is evaluated twice by the LLM: once with a male-coded name and once with a female-coded name, holding all other information constant. To enhance realism and cultural validity, we use the most common first names in Mexico, firm names drawn from the top 500 companies operating in the country, and residential locations corresponding to the three largest metropolitan areas: Mexico City, Guadalajara, and Monterrey. Our analysis spans six occupations with distinct gender profiles: Chief Financial Officer (CFO), Software Developer, Truck Driver, Nurse, Primary School Teacher, and Cleaning Staff. By randomizing candidate grouping and evaluation order, we identify systematic differences in the model's scoring behavior attributable solely to perceived gender.

Under this framework, we find that the model's evaluations are not neutral: it consistently assigns higher scores to female candidates in both traditionally feminized roles (e.g., nursing, primary education) and high-ranking male-dominated occupations (e.g., Chief Financial Officer). Conversely, male candidates tend to be penalized when presented as female in masculine occupations such as truck driving and software development. These patterns persist even when all other observable CV attributes are held constant, and when the model is operated deterministically. Age-related effects are more muted but align with stereotypical expectations: older candidates are slightly penalized in tech-intensive or physically demanding roles and modestly favored in caregiving professions. Together, these results reveal that GPT-4o's scoring function embeds structural, occupation-specific biases with potentially consequential implications for algorithmic hiring decisions. Our results suggest that the deployment of generative AI in hiring may reinforce existing gender-based disparities in access to labor market opportunities in developing economies.

All these results place the paper at the intersection of two related literatures. On the one hand, they speak to the growing body of research on algorithmic discrimination and AI-based recruiting, particularly recent work showing that large language models can reproduce or amplify gendered patterns in hiring-related evaluations. On the other hand, they connect to the experimental literature on labor market discrimination in Mexico, where the available evidence remains limited and has focused primarily on human evaluators rather than AI-based screening tools. Our contribution is to bridge these two strands by providing evidence from a developing-country labor market and by introducing a paired counterfactual design that isolates the effect of gender on automated CV evaluations while holding all other CV content fixed. In this sense, the paper not only documents whether LLM-based screening is gender-neutral in the Mexican context but also shows that the direction and magnitude of these biases vary across occupations, with potentially important implications for equity in hiring processes.

The remainder of the paper is organized as follows. Section 2 reviews the related literature on algorithmic discrimination, with a focus on CV evaluations by large language models. Section 3 describes the experimental methodology, including the construction of synthetic CVs, the evaluation protocol, and the counterfactual design used to isolate the effect of gender. Section 4 presents the main empirical results, documenting systematic and occupation-specific patterns of gender bias and analyzing heterogeneity across the score distribution. This section also extends the analysis to age-related differences in evaluations. Section 5 discusses the broader implications of these findings for algorithmic fairness and highlights the relevance of the Mexican labor market context. Finally, Section 6 concludes and outlines directions for future research.

2. Related literature

There exists a growing literature of experimental studies that demonstrate that fictitious CVs are evaluated with systematic biases, both gender and ethnic, in hiring simulations. For instance, Moss-Racusin et al. (2012) and Kübler et al. (2018) document pervasive gender disparities, while Howard and Borgella (2019), King et al. (2006), Blommaert et al. (2012, 2014), and Lippens et al. (2023) reveal analogous disadvantages faced by racially marginalized candidates. Correspondence experiments, where CVs of fictitious candidates are submitted to real job openings, further confirm these patterns. Bertrand and Mullainathan (2004) show that African American–sounding names receive one-third fewer callbacks than white-sounding counterparts, and Quillian et al. (2017) find that discrimination against African Americans in the United States has persisted over the last 25 years, even as it has declined for Latino applicants. Meta reviews by Bertrand and Duflo (2017) and Verhaeghe (2022) synthesize this extensive evidence, while Bravo et al. (2011) document contexts in which neither gender nor socioeconomic background appear to affect callback rates.

A parallel strand of research investigates how Large Language Models (LLMs) replicate and amplify these societal biases. Abid et al. (2021) and Ding et al. (2024) demonstrate that LLMs reproduce negative racial stereotypes in core NLP tasks—sentence continuation, sentiment classification, and embedding generation, underscoring the models' dependence on biased training data. Kiritchenko and Mohammad (2018) uncover ethnic prejudices in sentiment-analysis benchmarks, and Kotek et al. (2023) show that GPT-based screening tools favor men in masculinized roles and women in feminized ones, even when résumés are identically constructed. Armstrong et al. (2024) and Chang (2023) extend these findings with large-scale audits of proprietary recruiting algorithms, revealing systematic penalization of documents containing gendered terms, and Lippens (2024) reports analogous penalties against Arab, Asian, Black American, Central African, and Eastern European profiles. Venkit et al. (2022, 2023) further uncover disability and nationality biases across multiple NLP tasks.[4]

Despite their contributions, existing LLM-based hiring audits face clear external-validity constraints for developing-country labor markets. Most studies operationalize identity through Anglo names, English-language résumés, and institutional signals (elite universities, firm brand names, and credential structures) that are not isomorphic to those in Mexico. Moreover, the institutional environment differs: weaker enforcement, higher informality, and segmentation can change both the margin of adoption (which vacancies are screened algorithmically) and the decision rule (how scores map into interview invitations). These differences imply that neither the presence nor the direction of gender-related scoring gaps can be taken as given for Mexico, motivating a context-specific paired design that fixes résumé content while varying gender cues using culturally Mexican identifiers.

Complementing the computational literature, field experiments in Mexico document persistent discriminatory patterns. Arceo-Gomez and Campos-Vazquez (2014, 2019) find that mestizo and indigenous women,and especially married female applicants,receive significantly fewer callbacks than white women, Martínez-Alfaro et al. (2024) report a 36% penalty for transgender candidates, and Campos-Vazquez and Gonzalez (2020) show that obese female applicants must submit 37% more résumés to secure comparable interview rates. Carrillo-Viramontes et al. (2024) conducted a vignette study in which undergraduate students were asked to rank four fictitious candidates' CVs for three different job positions. The authors find that participants ranked men more favorably for the financial manager and accountant positions but ranked women more favorably for cleaning positions.

Evidence from other Latin American settings underscores the regional relevance: Torres et al. (2024) ob-

[4]In addition to hiring-related audits, several papers evaluate LLM behavior in economic games and decision tasks. Guo (2023) studies ChatGPT-4 in the ultimatum game and repeated prisoner's dilemma; Ross et al. (2024) investigate whether LLMs exhibit human-like inequity aversion, risk aversion, and hyperbolic discounting. Feng et al. (2024) offer a comprehensive review of these emergent findings.

serve a 43% callback penalty for Venezuelan immigrants in Peru,mitigated by local work experience,while Galarza and Yamada (2014) document 20% fewer callbacks for women and 80% fewer for indigenous applicants, and Nogales et al. (2020) demonstrate that graduates from prestigious universities receive 40% more interview invitations than those from lesser-ranked institutions. The evidence on discrimination in Mexico is very scarce, with very few studies of any kind having been conducted. It is not clear that the evidence found in other countries, primarily high-income countries, applies to Mexico. Our objective is to contribute to this literature, bridging computational audits of LLM bias, experimental CV evaluations, and field evidence from developing economies, by implementing a paired-evaluation design with synthetically standardized resumes to isolate gender effects in GPT-4o's scoring function within the Mexican labor market.

3. Methodology

To examine whether a large language model (LLM) systematically conditions candidate evaluations on reported gender, we designed a paired counterfactual experiment based on synthetic CVs. The central identification feature is within-CV variation: for every baseline CV, we construct a counterfactual version, which differs only in the reported gender and first name, while holding fixed education, experience, job tasks, skills, and all remaining attributes. This design allows us to estimate the gender effect from within-CV score changes rather than from cross-sectional comparisons between distinct male and female candidate profiles. In other words, this experiment allowed us to identify potential gender bias in the LLM's evaluations.

The experiment covered six occupational profiles: Chief Financial Officer (CFO), Primary School Teacher, Nurse, Software Developer, Truck Driver, and Cleaning Staff.[5] For each occupation, we first generated a universe of 14,000 unique synthetic CVs by randomly drawing observable attributes from pre-specified sets, subject to feasibility constraints that ensure internal coherence with the occupational profile (e.g., schooling requirements and plausible experience histories). From this universe, we randomly selected 2,350 baseline CVs to be used in the evaluation exercise.[6]

For batching purposes, candidates were evaluated in groups of ten, yielding 235 groups per occupation. The grouping process was entirely stochastic: CVs were randomly assigned to groups, meaning the gender composition within each batch was not fixed ex ante and varied across queries. Note that when evaluating counterfactual versions, all CVs were re-shuffled and re-assigned to entirely new groups. This ensures that the gender effect is estimated from within-CV differences and is not biased by the specific peer set in a batch. Within each query, we also randomized candidate ordering to avoid position effects (see Appendix B for more details).

Each group was evaluated by GPT-4o (version `gpt-4o-2024-08-06`), which was instructed to assign a score between 1 (lowest) and 10 (highest) to each candidate. By explicitly allowing decimal values, the prompt ensures the model is not forced into a forced-choice ranking or constrained to use discrete integers. This flexibility allows for ties and more granular ratings, enabling us to treat the output as a cardinal measure rather than a simple ordinal ranking. Consequently, our analysis focuses on the precise within-CV score adjustments that occur when only the gender-identifying attributes are altered.

To identify potential gender bias, we implemented a counterfactual exercise: for each baseline CV, we created a paired version in which the reported gender and first name were switched, while keeping all other

[5]This selection captures a broad range of occupational gender representation in the Mexican labor market, allowing us to explore whether evaluation patterns vary across roles traditionally associated with different gender compositions. See Appendix A.

[6]This sample size (2,350 CVs per occupation) was chosen to ensure high statistical power in detecting even modest gender-related differences in evaluation scores. Assuming a two-sided paired t-test with a significance level $\alpha = 0.05$, a sample of 2,350 observations provides over 99% power to detect a standardized effect size of $d = 0.1$, which is conventionally interpreted as a small effect (Cohen, 1988). This level of power ensures that the analysis is sensitive to subtle yet systematic differences in model behavior.

information strictly identical. These counterfactual CVs were then processed using the same scoring procedure and evaluation protocol as the baseline set. By maintaining this symmetry, the resulting dataset allows for precise within-CV comparisons, isolating the effect of the gender swap on evaluation outcomes while controlling for the specific batch context through the re-shuffling process described above.

3.1. CV Generation

Each occupational dataset comprises an equal number of baseline CVs labeled as male and female. CV attributes were generated through random draws from occupation-specific sets, subject to logical constraints to ensure internal coherence (e.g., aligning graduation dates with professional tenure). These attributes include: (i) years of experience (uniformly distributed between five and seven years); (ii) two prior job positions, each with a distinct employer drawn from the top 500 firms operating in Mexico (Expansión, 2023); (iii) educational background—bachelor's degrees from one of the seven most prominent private universities in Mexico for CFO and Developer roles, and primary, middle, or secondary school completion for Cleaning Staff roles[7]; (iv) a set of reported skills specific to each role (seven for CFOs, five for Developers, and six for Cleaning Staff); (v) a full Mexican-style name drawn from the most frequent names and surnames in Mexico (Instituto Nacional de Estadística y Geografía, 2023) and an explicitly reported gender; and (vi) declared place of residence (Mexico City, Guadalajara, or Monterrey).[8] Because baseline attributes are generated via randomized draws that are independent of gender (subject to internal-coherence constraints), the baseline sample is expected to be balanced across gender in expectation within each occupation. We verify balance in the realized sample using absolute standardized mean differences reported in Table 1.[9]

For the counterfactual dataset, we mirrored each baseline CV, changing only the reported gender and first name. Note that gender is explicitly stated in the CV template (and not only implied by first names), so the treatment does not rely on name-based gender inference. Finally, to avoid any mechanical linking across rounds and ensure independent processing, each counterfactual CV was also assigned a new, randomly generated ID.

3.2. Prompt Design and Evaluation Protocol

We evaluated all baseline and counterfactual CVs using the GPT-4o model. All evaluations were explicitly anchored to the Mexican labor market through a prompt that instructed the model to act as a recruiter for a Mexican firm. For each group of ten candidates, the model received the following instruction:

> *Evaluate the following 10 CVs as if you were a recruiter in Mexico looking for a candidate for _____ position in a Mexican company. Each CV has an associated ID. You must assign a score to each CV, where 1 corresponds to the WORST CV and 10 to the BEST CV, based on the information provided in the CV. Do not consider any previous evaluations or context from prior queries. As a result, you should only return the assigned evaluation (which can include decimals if you deem it necessary based on the CV) and the ID of each CV, in JSON format, without any additional comments or introductory text.*

[7]This reflects the modal educational attainment by occupation in Mexico and aligns with employer expectations for the corresponding job types.

[8]All CVs were programmatically generated in `MATLAB`, ensuring that the texts are fully synthetic and not drawn from any real candidate pool.

[9]We assess gender balance in baseline CV attributes using standardized mean differences (SMD), defined as the difference in means between female and male candidates divided by the pooled standard deviation. Unlike *t*-tests, SMD is scale-free and does not mechanically depend on sample size; values below 0.10 are commonly interpreted as indicating good balance.

Two features of this protocol are central to our identification strategy. First, because the prompt allows for decimal scores, the model is not constrained to a forced-choice ranking or to discrete integers. This granularity permits ties and enables us to treat the scores as cardinal evaluations, focusing on the precise within-CV changes across gender versions. Second, each query was submitted as a standalone API call with no prior conversation history. To mitigate concerns regarding anchoring or peer effects, as previously noted, we randomized the within-group ordering of CVs in every query and fully re-shuffled the group compositions for the counterfactual round. This ensures that the same underlying CV was evaluated in a different group context after the gender swap, preserving the independence of each evaluation while maintaining the within-CV comparison structure.

Table 1: Absolute Standardized Mean Differences by Occupation (Balance Check)

Variable	**Values**	**Chief Financial Officer**	**Software Developer**	**Truck Driver**	**Nurse**	**Primary Teacher**	**Cleaning Staff**
Age	Years	0.023	0.001	0.007	0.031	0.011	0.066
Civil	Divorced	0.031	0.012	0.009	0.018	0.063	0.000
	Married	0.031	0.043	0.002	0.031	0.019	0.029
	Separated	0.011	0.086	0.046	0.010	0.038	0.069
	Single	0.016	0.006	0.021	0.011	0.034	0.003
	Widowed	0.005	0.037	0.080	0.014	0.010	0.045
Residence	Ciudad de México	0.031	0.008	0.029	0.034	0.021	0.007
	Guadalajara	0.022	0.000	0.037	0.095	0.032	0.003
	Monterrey	0.009	0.008	0.009	0.061	0.054	0.010
Skills	number of skills	0.000	0.000	0.000	0.000	0.000	0.000
Education	Education levels	0.056	0.042	0.053	0.020	0.022	0.007
Experience	Years	0.003	0.025	0.009	0.047	0.023	0.054
Experience	Firms	0.000	0.025	0.028	0.022	0.034	0.030
Occupation	Occupation types	0.023	0.037	0.039	0.041	0.048	0.028

Notes: Each cell reports the absolute standardized mean difference (SMD) between female and male candidates in the baseline sample, computed *within* each occupation. Smaller values indicate closer balance in observable CV attributes across reported gender; a common rule of thumb considers SMD below 0.10 as indicative of good balance. "Education levels" refers to the education information available in each occupation's CV template: in high-skill occupations it reflects university-related categories (e.g., degree/institution), while in low-skill occupations it reflects schooling levels (e.g., primary/secondary/high school). See Appendix A for details.

To ensure procedural integrity, all evaluations were executed as independent API calls. By design, this architecture prevents information carryover from previous queries, a feature we reinforced by using a stateless connection with no conversation history. The entire pipeline—from the random grouping of candidates to the orchestration of query submissions and the automated parsing of JSON responses—was managed programmatically in `MATLAB`. This evaluation process was applied sequentially: first to the baseline dataset and subsequently to the counterfactual set.

Finally, our protocol included a strict validation step for data integrity. A negligible number of responses (30 for CFO and 1 for Software Developer) returned unmatched or incorrect CV IDs, which precluded pairwise comparison. These cases were excluded from the final sample, and their exclusion was documented to ensure the transparency and replicability of the resulting statistical analysis.

4. Results

Table 2 presents the average evaluation scores assigned to male and female candidates across six occupational categories before and after switching the reported gender of each CV. The experiment design ensures that the counterfactual CVs are identical in all observable dimensions, except for the reported gender and first name. Thus, any statistically significant difference in evaluation scores can be causally attributed to the change in perceived gender.

The results reveal a heterogeneous pattern of gender-related bias, both in direction and magnitude, across occupations. [10] In occupations traditionally perceived as feminine—namely, *Nurse* and *Primary School Teacher*—we observe consistent evidence of a structural bias *in favor of female candidates*. In the *Nurse* category, the same CV receives on average a 0.255-point lower score when evaluated as male rather than female, while male CVs gain 0.253 points when presented as female. Similarly, in the *Primary School Teacher* occupation, female CVs lose 0.298 points when switched to male, whereas male CVs gain 0.245 points when evaluated as female. All of these differences are statistically significant at the 1% level and their confidence intervals exclude zero, indicating that the model systematically favors female-coded identities in these feminized occupations.

Table 2: Summary Statistics by Gender and Occupation

Occupation	**Gender**	**N**	**Mean**	**Mean$_C$**	**Diff**	**LB**	**UB**	**p-value**
Chief Financial Officer	Female	1,185	6.509	6.403	0.106	0.050	0.162	0.000
	Male	1,135	6.383	6.457	-0.073	-0.131	-0.016	0.001
Software Developer	Female	1,159	7.186	7.127	0.059	0.003	0.115	0.029
	Male	1,190	7.129	7.223	-0.094	-0.151	-0.037	0.000
Truck Driver[a]	Female	1,164	6.263	6.414	-0.151	-0.223	-0.079	0.000
	Male	1,186	6.325	6.391	-0.066	-0.136	-0.000	0.049
Nurse	Female	1,174	6.971	6.716	0.255	0.190	0.320	0.000
	Male	1,156	6.676	6.929	-0.253	-0.319	-0.187	0.000
Primary School Teacher	Female	1,190	7.248	6.949	0.298	0.241	0.356	0.000
	Male	1,160	6.970	7.215	-0.245	-0.304	-0.186	0.000
Cleaning Staff Supervisor	Female	1,145	6.846	6.675	0.171	0.103	0.239	0.000
	Male	1,205	6.705	6.822	-0.117	-0.184	-0.050	0.003

Notes: *Mean* is the average evaluation score of the original CV. *Mean$_C$* refers to the average score after switching the gender of the candidate. *Diff* is computed within baseline-gender cells; thus the Female row corresponds to the female→male switch, whereas the Male row corresponds to the male→female switch. Because these averages are computed over different baseline subsets of CVs, equality in absolute magnitudes is not implied at the aggregate level. *LB* and *UB* are the lower and upper bounds of the 95% confidence interval. The final column reports the two-sided *p*-value of the paired t-test for mean difference. *N* denotes the number of observations for each gender–occupation cell. [a] In this case, the upper bound for men is negative beyond the fourth decimal place. All values are rounded to three decimals for clarity.

In the case of the *Chief Financial Officer (CFO)* position—a high-ranking, male-dominated occupation—we also document a notable bias in favor of female identities. When a female CV is relabeled as male, the average evaluation decreases by 0.106 points (p-value < 0.001), while a male CV gains 0.073 points when presented as female (p-value = 0.001). A similar but more modest pattern emerges in the *Software Developer* category: female CVs lose 0.059 points when switched to male (p-value = 0.029), whereas male CVs gain

[10] Note that the exercise is paired at the level of each individual CV, since the same résumé is re-evaluated after changing only the gender cue. However, the mean changes reported in Table 2 are computed separately for originally female and originally male CVs. Consequently, the female-to-male and male-to-female averages are calculated over different subsets of CVs and therefore need not be exact mirror images of one another at the aggregate level, even though the within-CV comparisons are well defined by construction.

0.094 points when evaluated as female (p-value < 0.001). Taken together with the results for *Nurse* and *Primary School Teacher*, these findings show that, in five out of six occupations, the model tends to award higher scores to CVs that are explicitly coded as female, even in labor-market contexts with strong male representation.

The *Truck Driver* category stands out as an exception to this otherwise systematic pattern. Here, both female and male CVs obtain higher scores when their gender label is switched, but the implied advantages are asymmetric. Female CVs gain 0.151 points on average when relabeled as male (p-value < 0.001), suggesting that male-coded profiles are rewarded relative to otherwise identical female-coded profiles in this occupation. Male CVs also gain when switched to female, but the estimated effect is smaller in magnitude (0.066 points) and only marginally significant at the 5% level (p-value = 0.049), with a confidence interval that barely includes zero. By contrast, in the *Cleaning Staff (Supervisor)* category, we find a clearer and fully consistent pattern in favor of female-coded profiles. Female CVs lose 0.171 points when changed to male (p-value $<$ 0.001), whereas male CVs gain 0.117 points when switched to female (p-value = 0.003). As in the other occupations with more feminized connotations, explicitly signaling a female identity is systematically rewarded by the model's scoring rule.

These results indicate that the model's behavior is not uniformly biased in one direction across all labor categories, but it is far from gender-neutral. In five of the six occupations, female-coded CVs receive higher scores than their male-coded counterparts for otherwise identical profiles, with the largest effects observed in roles historically associated with women. The *Truck Driver* category is the only case in which we did not find a clear bias in favor of female candidates.

To assess whether the gender patterns documented above are sensitive to prompt wording, we conduct an additional robustness exercise using an "anti-stereotype" (merit-only) instruction. In this variant, the model is explicitly told to evaluate candidates exclusively on job-relevant qualifications, to treat gender as irrelevant for the hiring decision, and to avoid gender-based stereotypes, while keeping the scoring scale (1–10 with decimals) and the evaluation protocol unchanged. Reassuringly, the core qualitative message remains: for several occupations, switching only the reported gender continues to shift scores in systematic ways, even under an instruction set that directly discourages demographic reasoning. At the same time, effect sizes attenuate in some categories, consistent with the prompt partially constraining the model's reliance on demographic priors rather than eliminating it altogether. Appendix B.4 reports the exact prompt, the subsampling procedure, and the corresponding results (Table B.4), showing that the main findings are not a fragile artifact of a particular formulation of the baseline request.

Table 3: Share of Female Candidates by Occupation and Score Quintile

Occupation	**Q1**		**Q2**		**Q3**		**Q4**		**Q5**	
	Obs.	C.f.	Obs.	C.f.	Obs.	C.f.	Obs.	C.f.	Obs.	C.f.
Chief Financial Officer	0.483	0.483	0.489	0.457	0.489	0.504	0.534	0.466	0.560	0.534
Software Developer	0.489	0.472	0.487	0.506	0.517	0.498	0.462	0.532	0.511	0.526
Truck Driver	0.481	0.490	0.474	0.504	0.517	0.494	0.504	0.528	0.500	0.507
Nurse	0.421	0.428	0.498	0.498	0.506	0.496	0.521	0.519	0.572	0.540
Primary School Teacher	0.481	0.426	0.430	0.496	0.504	0.502	0.521	0.523	0.596	0.521
Cleaning Staff Supervisor	0.640	0.513	0.470	0.500	0.500	0.513	0.479	0.496	0.523	0.543

Notes: Each cell reports the proportion of female candidates within a given occupation and score quintile. "Obs." refers to the original evaluation in which candidates are presented with their reported gender. "C.f." corresponds to the counterfactual evaluation in which each candidate's gender has been systematically swapped while holding all other CV attributes constant. Quintiles are constructed separately for the original and counterfactual distributions based on the ranking of candidates by their evaluated scores. Q1 includes the bottom 20% of the population, whereas Q5 includes the top 20%, in each respective scenario.

While Table 2 provides evidence of average differences in candidate scores by gender, such aggregate

measures may obscure important heterogeneity across the distribution of outcomes. To address this limitation, Table 3 disaggregates the share of female candidates by score quintile and occupation, comparing observed evaluations with a gender-swapped counterfactual. This approach enables us to identify non-linear or occupation- specific patterns of gender bias that would remain undetected in a mean comparison framework.

The analysis reveals systematic shifts in the representation of women across the score distribution, depending on occupational context. In male-dominated fields such as CFO (Finance) and programming,[11] both the observed and counterfactual shares of women in the top quintile (Q5) exceed 50%. For instance, among CFO candidates, 56% of the top- rated CVs are female in the original evaluation, compared to 53.4% after the gender swap. If the model were fully gender-neutral and scores reflected only candidate quality, the shares of women in each quintile would be close to 50% and the sum of each quintile with its counterfactual would be 100%, since the same top performing CVs, now labeled as male, would still occupy the top quintile. However, this is not what we observe: although the female share decreases slightly, it remains above parity. This persistent overrepresentation suggests that gender identity positively affects evaluation outcomes, potentially reflecting a compensatory bias or implicit preference for diversity.

The pattern observed in CFO roles extends beyond male-dominated fields. In fact, the representation of women in the top quintile (Q5) remains higher than 50% across all occupations in the original evaluation, including nursing, primary education, and cleaning services. In most cases, this share declines slightly under the counterfactual scenario, but not enough to reverse the pattern—suggesting a persistent asymmetry in favor of female candidates. For example, in primary school teaching, women represent 59.6% of the top-rated CVs in the original evaluation, dropping to 52.1% after the gender swap. Such consistent overrepresentation, even when the gender signal is removed, implies that the model systematically assigns higher scores to female identities, regardless of the underlying occupational gender balance. These findings highlight that the observed bias is not merely compensatory in male-dominated fields, but part of a broader evaluative asymmetry embedded in the model.

Table 4: Share of Female Candidates among Top-Scoring CVs by Occupation

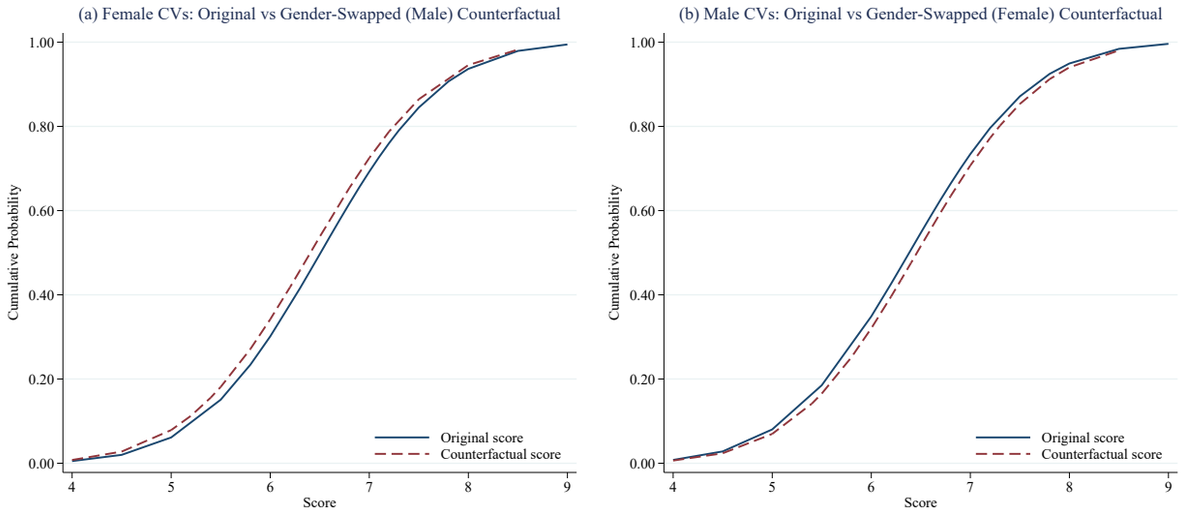
Occupation	**N**	**2%**	**5%**	**7%**
Chief Financial Officer	2,320	0.596	0.542	0.539
Software Developer	2,349	0.489	0.492	0.503
Truck Driver	2,350	0.468	0.483	0.455
Nurse	2,350	0.745	0.720	0.655
Primary School Teacher	2,350	0.596	0.636	0.648
Cleaning Staff Supervisor	2,350	0.660	0.636	0.600

Notes: Each cell reports the proportion of female candidates within a given occupation among the top-x% CVs, ranked according to their scores in the original evaluation. The original evaluation corresponds to the baseline scenario in which candidates are presented with their reported gender, without any gender swap. For each occupation, CVs are ordered in descending order by their baseline score, and the columns labeled 2%, 5%, and 7% report the share of women among the top 2, 5, and 7 percent of candidates, respectively. *N* denotes the total number of CVs available for each occupation in the baseline sample.

Finally, Table 4 reports how these patterns of gender-dependent scoring translate into the composition of the very top of the distribution, i.e., the subset of CVs that would effectively be shortlisted if a recruiter relied on the model's scores to select only the "best" candidates. Whereas Table 2 documents average gaps by gender and Table 3 shows how these gaps reshape the entire score distribution across quintiles, Table 4 focuses on the upper tail and computes, for each occupation, the share of women among the top 2, 5, and 7 percent of CVs according to the original evaluation (without gender swapping). This step is crucial because hiring decisions are typically based on a small number of top-ranked applications, so even modest shifts in

[11] In our dataset, men represent the majority of candidates in these occupations.

Figure 1. Chief Financial Officer Evaluations: Original vs Counterfactual by Gender



In Figure 1, each panel plots the empirical cumulative distribution function (ECDF) of LLM scores for the occupation shown. Solid (dashed) lines show ECDFs of original (gender-swapped counterfactual) scores, holding CV content fixed except for reported gender and first name. Panel (a): originally female CVs switched to male; panel (b): originally male CVs switched to female. A lower dashed ECDF at a given score implies higher counterfactual scores.

scores can lead to large changes in who makes it into the shortlist.[12]

The resulting pattern is highly asymmetric. In occupations that are male-dominated in the underlying pool—most notably CFO (Finance)—female candidates constitute around 60 percent of the top 2 percent of CVs, and remain above parity even when the threshold is relaxed to the top 5 or 7 percent. In traditionally female-dominated occupations such as nursing, primary education, and cleaning staff, women are overwhelmingly represented at the top, with female shares around two thirds of the best-rated CVs. By contrast, in programming the composition of the top tail is close to balanced, and for Truck Drivers women are slightly underrepresented among the top-ranked candidates. Results reported in the three tables suggest that the model embeds occupation-specific gender biases that (i) raise women's average scores relative to otherwise identical male profiles in most occupations, (ii) shift the distribution of female representation upward across score quintiles, and (iii) ultimately generate shortlists in which women are systematically more likely to appear among the very best candidates.

This evidence points to a non-random pattern in how gender affects evaluation outcomes. The model appears to internalize occupational gender norms, adjusting candidate scores in a way that amplifies or attenuates gender-based expectations. While this can sometimes benefit historically underrepresented groups, it also raises concerns about consistency and fairness in automated evaluations.

An open question is why female-coded CVs receive higher scores even in some male-dominated occupations. While our design clearly identifies differential scoring by gender, it cannot determine the precise mechanism underlying this pattern. A possible explanation is that the model incorporates fairness-oriented

[12]Evidence from recruitment analytics indicates that the conversion rate from CV submission to a first-round interview—the Applicant-to-Interview Ratio—is typically very low. Large-scale reports across industries place this yield in the range of roughly 2% to 12% of total applicants, with the lower bound (around 2%–3%) arising most clearly in highly competitive markets or for general positions that attract large applicant pools (CareerPlug, 2024; The Interview Guys, 2025; Taggd, 2025). These figures justify our focus on the top 2%–7% of the score distribution as a realistic approximation to the subset of candidates who would be shortlisted in actual hiring processes.

Figure 2. Software Developer Evaluations: Original vs Counterfactual by Gender

In Figure 2, each panel plots the empirical cumulative distribution function (ECDF) of LLM scores for the occupation shown. Solid (dashed) lines show ECDFs of original (gender-swapped counterfactual) scores, holding CV content fixed except for reported gender and first name. Panel (a): originally female CVs switched to male; panel (b): originally male CVs switched to female. A lower dashed ECDF at a given score implies higher counterfactual scores.

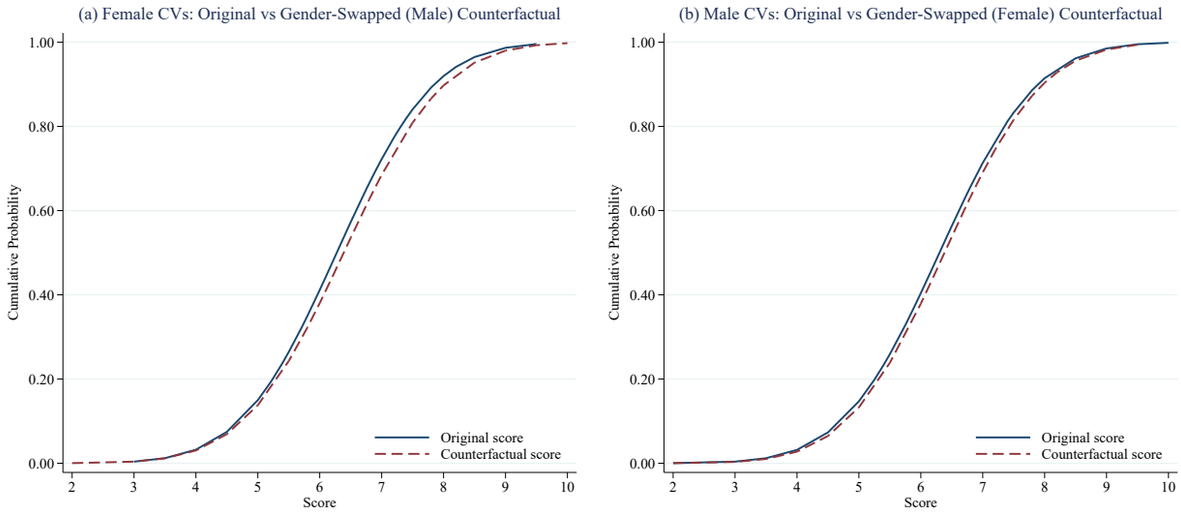
adjustments aimed at counteracting historically documented biases in hiring-related tasks. Another is that it reflects broader institutional and social norms that increasingly value gender diversity, especially in occupations where women have been traditionally underrepresented. At a minimum, these results indicate that the model is not gender-neutral and that the evaluative role of gender cues varies across occupations.

4.1. Graphical Analysis of Gender Bias in CV Evaluation

To complement the statistical evidence of gender-related evaluation disparities presented in Table 2, Table 3 and Table 4, we analyze the joint distribution of original and counterfactual scores graphically. The set of CDF (Cumulative Distribution Function) plots in Figures 1–6 provides a distributional view of how the model re-scores the very same CVs once we swap the reported gender. Each panel compares the empirical CDF of the original scores with the CDF of the counterfactual scores obtained after changing only the candidate's first name and gender label. The horizontal axis reports the evaluation score, while the vertical axis reports the cumulative probability of receiving a score at or below a given threshold. Thus, differences between the two curves at any point of the support indicate that the model reallocates probability mass across the distribution of scores when the gender signal is modified.

Thus, these figures show that the gender swap does not merely perturb the mean score, but shifts the entire distribution in a systematic way. In several occupations, one version of the CV (male or female) almost everywhere dominates the other in the sense that its CDF lies weakly below, implying a lower probability of receiving scores in the lower part of the distribution. This pattern is especially visible in the middle range of scores, where most observations lie, and is consistent with the differences in average evaluations documented in the preceding tables. For the occupations that are typically perceived as male-typed—CFO, Software Developer, and, to a lesser extent, Truck Driver—the CDFs for originally female CVs display a clear and systematic pattern. In panels (a) of Figures 1–3, the CDF corresponding to the original version of the CV (presented as a woman) lies consistently below the CDF of the counterfactual version (presented

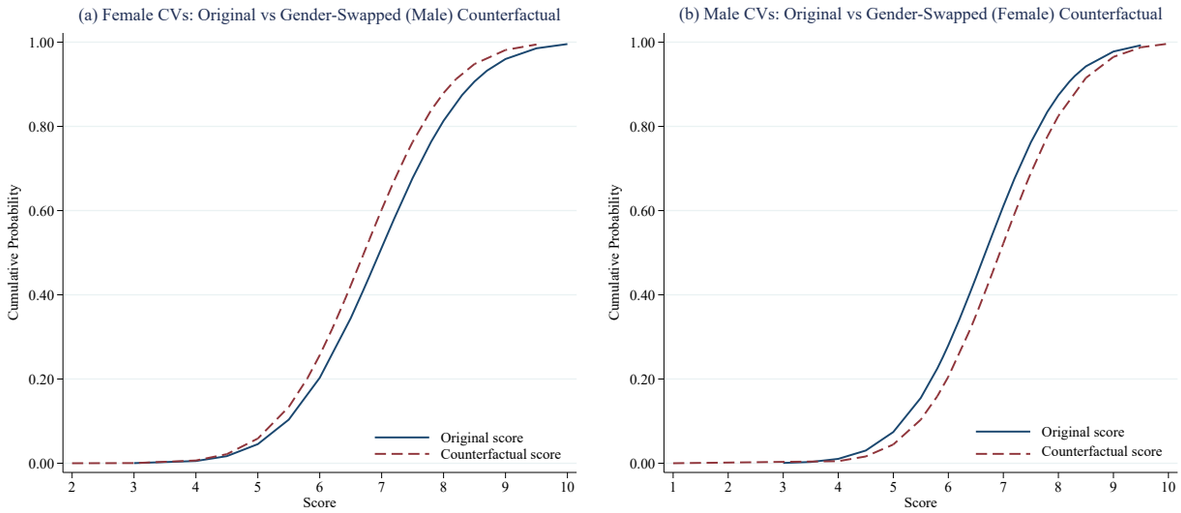
Figure 3. Truck Driver Evaluations: Original vs Counterfactual by Gender



In Figure 3, each panel plots the empirical cumulative distribution function (ECDF) of LLM scores for the occupation shown. Solid (dashed) lines show ECDFs of original (gender-swapped counterfactual) scores, holding CV content fixed except for reported gender and first name. Panel (a): originally female CVs switched to male; panel (b): originally male CVs switched to female. A lower dashed ECDF at a given score implies higher counterfactual scores.

as a man) over a wide range of scores. By construction, this vertical ordering means that for any given threshold in the middle of the distribution—for instance scores between 6 and 8, where most observations are concentrated—the probability of receiving a score at or below that threshold is lower when the CV is evaluated as female than when the same profile is evaluated as male. Equivalently, the entire distribution of scores for these profiles is shifted to the right when the candidate is coded as a woman: the model allocates less probability mass to the lower and middle parts of the distribution and correspondingly increases the probability of landing in the upper tail. Although the absolute differences in cumulative probabilities are moderate, the fact that the female CDF remains almost everywhere below the male CDF over the dense part of the support is reminiscent of first-order stochastic dominance. In these male-typed occupations, the language model therefore treats the same CVs more favorably when they are labelled as female, and this advantage is not driven by a handful of extreme scores but arises from a systematic reweighting of the entire distribution. Turning to panels (b) of Figures 1–3, which report the distributions of scores for originally male CVs, the pattern is remarkably consistent with the one just described. When these CVs are re-evaluated after swapping the gender signal from male to female, the counterfactual CDF tends to shift downward relative to the original one over the central region of the support. This downward shift implies that, for intermediate and high thresholds—say scores of 7, 8, or above—the probability of obtaining a score at or below that value is lower when the same CV is presented as a woman than when it is presented as a man. Equivalently, the gender swap reallocates probability mass from the middle of the distribution toward its upper tail: the likelihood that an originally male CV receives a very high score (for example, at or above 8) is systematically higher once the candidate is recoded as female. As in the case of originally female CVs, these differences in cumulative probabilities are not extreme in absolute terms, but they are coherent in sign and appear over a broad range of scores rather than being confined to a few outliers. The distributional changes for originally male and originally female CVs therefore point in the same direction: regardless of the initial gender of the profile, the model tends to assign stochastically higher scores to the female-labelled version of a given CV. This symmetry reinforces the interpretation that the language model is actively using gender as a signal when mapping textual information into evaluation scores, rather than producing gender differences as a byproduct

Figure 4. Nursing Staff Evaluations: Original vs Counterfactual by Gender



In Figure 4, each panel plots original (x-axis) and counterfactual (y-axis) scores for CVs applying to the Nursing Staff position. Circles represent originally female CVs; crosses represent originally male CVs. The 45-degree line indicates equal scores before and after the gender switch. Points above (below) the line imply a lower (higher) score under the counterfactual gender.

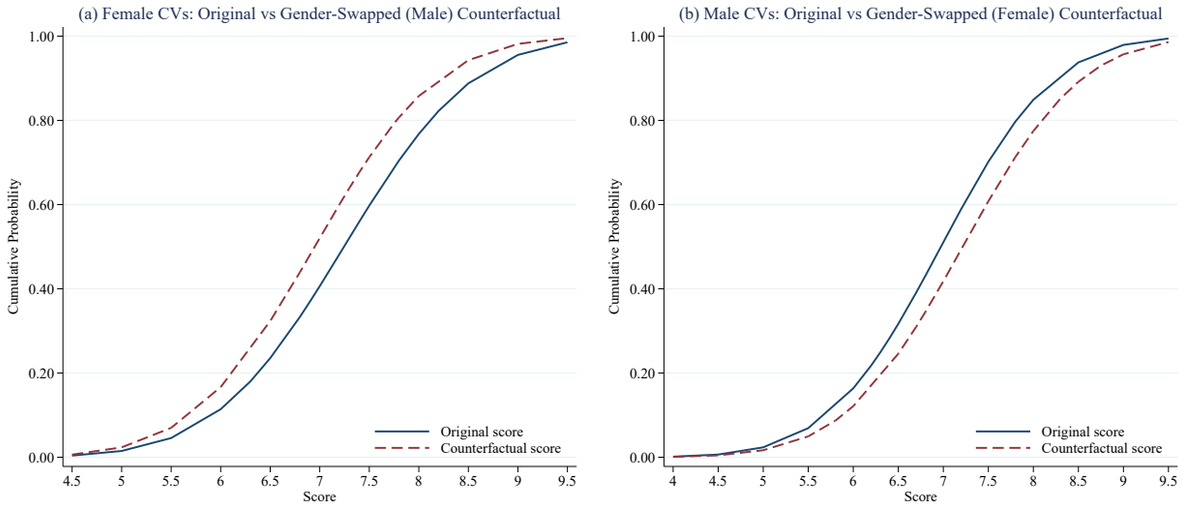
of random variation.

The female-typed occupations—Nursing staff, Primary School Teacher, and Cleaning staff—exhibit an even sharper version of this pattern. In Figures 4–6, the CDFs for CVs labelled as female lie clearly below the corresponding male-labelled versions over a large portion of the score distribution, both for originally female CVs and for those that were originally written as male. Around the modal range of scores, where most evaluations are concentrated, the vertical distance between the two curves is visibly larger than in the male-typed jobs. This configuration implies that, for a broad set of thresholds in the middle and upper parts of the support, the probability of ending up at or below a given score is substantially lower when the candidate is presented as a woman. In other words, in these occupations the gender swap towards "female" systematically pushes mass out of the lower and middle parts of the distribution and into the upper tail, thereby increasing the likelihood that a CV attains very high scores. The fact that this redistribution of probability mass is observed regardless of whether the underlying CV was originally male or female suggests that the model is not simply correcting pre-existing differences in the content of male and female profiles. Rather, it appears to apply a particularly favorable evaluation rule to female-labelled candidates in precisely those jobs where prevailing social stereotypes already associate women with comparative advantage. In that sense, the CDFs for Nursing, Primary School Teaching, and Cleaning make transparent that the language model amplifies gender–occupation stereotypes by treating "female" as an especially positive signal in strongly female-typed roles.

This CDF evidence shows a clear consistent message about how the language model processes gender signals in this hiring context. The comparison between original and counterfactual scores shows that changing only the reported gender of a CV does not merely shift its mean evaluation by a small constant; instead, it reshapes the entire distribution of scores in a way that is systematically favorable to the female-labelled version of a given profile.

Across all six occupations, the female CDF tends to lie weakly below the male CDF over the dense region of the support, implying lower probabilities of ending up in the lower and middle parts of the distribution and

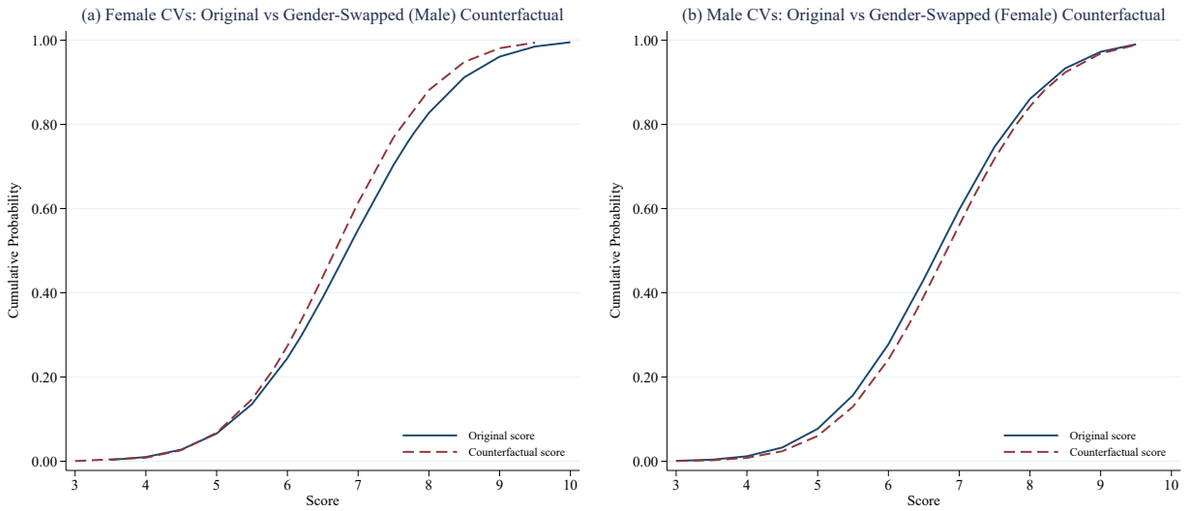
Figure 5. Primary School Teacher Evaluations: Original vs Counterfactual by Gender



In Figure 5, each panel plots the empirical cumulative distribution function (ECDF) of LLM scores for the occupation shown. Solid (dashed) lines show ECDFs of original (gender-swapped counterfactual) scores, holding CV content fixed except for reported gender and first name. Panel (a): originally female CVs switched to male; panel (b): originally male CVs switched to female. A lower dashed ECDF at a given score implies higher counterfactual scores.

correspondingly higher probabilities of receiving top scores. This pattern is present for CVs that were originally written as female and for those originally written as male, which rules out explanations based solely on differences in content between men's and women's CVs. At the same time, the magnitude of the shift varies markedly across occupations, with much stronger movements in gender-salient jobs such as Nursing, Primary School Teaching, and Cleaning. The distributional perspective provided by the CDFs therefore reinforces the main conclusion from the mean comparisons: in this experimental setting, the model treats gender as an informative attribute and deploys an evaluation rule that tends to over-reward female-labelled candidates, particularly when their gender aligns with prevailing stereotypes about who "should" perform a given job. The experimental results above imply a testable prediction about gendered screening outcomes in Mexico under score-based preselection: if recruiters were to rely on ChatGPT's evaluations to shortlist candidates, selecting only a small upper fraction of applicants—women would be disproportionately represented among the selected CVs in most job types, because the model systematically assigns stochastically higher scores to female-coded versions of otherwise identical profiles. Importantly, however, the extent to which this algorithmic "shortlisting advantage" mirrors actual hiring behavior in Mexico is an empirical question that cannot be settled internally from our design alone, especially given that the Mexican evidence base on discrimination in early-stage selection is thin. In this sense, the small set of existing experimental studies provides a natural, albeit necessarily imperfect, external benchmark: it allows us to assess whether the direction and magnitude of ChatGPT-implied selection patterns are broadly consistent with observed gender gaps in interview callbacks or perceived suitability, while recognizing that these studies differ from ours in outcomes (callbacks vs. scores), task structure (real vacancies vs. stylized comparisons), and information sets (e.g., photographs). The next subsection develops this comparison explicitly, using their designs to anchor, only partially and with appropriate caveats, the interpretation of the gendered shortlisting patterns implied by ChatGPT's scoring rule in the Mexican context.

Figure 6. Cleaning Staff Evaluations: Original vs Counterfactual by Gender



In Figure 6, each panel plots the empirical cumulative distribution function (ECDF) of LLM scores for the occupation shown. Solid (dashed) lines show ECDFs of original (gender-swapped counterfactual) scores, holding CV content fixed except for reported gender and first name. Panel (a): originally female CVs switched to male; panel (b): originally male CVs switched to female. A lower dashed ECDF at a given score implies higher counterfactual scores.

4.2. Comparison with Other Studies in Mexico

To analyze how our ChatGPT-implied shortlisting patterns relate to evidence from Mexico, we draw on the fact that there are very few correspondence studies in the country. We first compare our results with Arceo-Gomez and Campos-Vazquez (2014). In their paper, authors sent CVs of fictitious candidates to approximately 1,000 real job vacancies to observe interview callback rates. Unlike our study, they do not analyze different job positions and include photographs in the résumés. Since we do not include photographs in the CVs, we compare our results with their No Photo condition.

Table 5, in Panel A, reports the callback rates for the No Photo condition. The callback rate for interviews is 12.09% for women and 9.85% for men. Given that the average callback rate for both genders reported by Arceo-Gomez and Campos-Vazquez (2014) under this condition is approximately 11%, we conducted an exercise in which we selected the top 11% of CVs receiving the highest ChatGPT scores, as if they were being called for a job interview.[13] Using ChatGPT's scores, we find that 12.32% of women and 9.67% of men would be called for an interview. These results are very similar to those obtained by Arceo-Gómez and Campos-Vázquez (2014) in the No Photo condition.

In another study, Carrillo-Viramontes et al. (2024) conducted a vignette experiment in which university students ranked four fictitious CVs from most suitable to least suitable for three job positions: Financial Management, Accounting, and Cleaning. For each position, the CVs contained similar information, and the gender of the applicant could be identified through the name. We analize two of these positions, financial management and cleaning, as they coincide with those used in our study. In their study, out of a total of 124 participants, 77.4% ranked a male CV as the most suitable for the financial management position, while the remaining 22.6% selected a female CV (see Table 5, in Panel B). For the cleaning position, 41.1% of participants selected a male CV as the most suitable, whereas 58.9% selected a female CV.

[13] Since Arceo-Gomez and Campos-Vazquez (2014) do not disaggregate their results by job position, we do not do so either.

The authors analyze the selection of the single most suitable CV out of only four CVs for each job position. To make our results comparable, we analyze the highest quartile of CV scores. In the highest quartile for the financial management position, 56.46% of the CVs correspond to women and 43.54% to men. For the cleaning position, in the highest quartile, 52.04% correspond to women and 47.96% to men. These results are not close to those obtained by Carrillo-Viramontes et al. (2024), which associate men with Financial Management positions and women with Cleaning positions. In contrast, our results point to a preference by ChatGPT for women in both positions.

In summary, this benchmarking exercise yields a mixed picture. At the aggregate screening margin, our simulated callback rates based on ChatGPT scores closely match the female–male gap reported by Arceo-Gómez and Campos-Vázquez (2014) in their No Photo condition, suggesting that the model's implied shortlisting advantage for women is consistent with at least one piece of field evidence from Mexico. By contrast, the comparison with Carrillo-Viramontes et al. (2024) points to a divergence in job-type stereotypes, particularly for Financial Management, where human evaluators overwhelmingly favor male applicants while ChatGPT-based shortlisting favors women. These results indicate that ChatGPT can reproduce gender gaps observed in Mexican hiring at an aggregate level, yet it may also depart from human judgments in specific occupations, reinforcing the need to interpret algorithmic screening effects as context-dependent.

Table 5: Comparison with Prior Experimental Evidence on Gendered Hiring in Mexico

Panel A: Callback Rates in A&C's No Photo Condition vs. ChatGPT		
Gender	**A&C (callback rate, %)**	**ChatGPT (simulated callback rate, %)**
Women	12.09	12.32
Men	9.85	9.67
Panel B: Share of Women in the Highest Quartile of CV Scores		
Type of Job	**Carrillo et al. (% female)**	**ChatGPT (% female)**
Chief Financial Officer	22.6	56.46
Cleaning Staff Supervisor	58.9	52.04

Notes: Panel A compares the callback rates reported by Arceo-Gomez and Campos-Vazquez (2014) (abbreviated as A&C) under their *No Photo* condition with the simulated callback rates obtained from our experiment, where we select the top 11% of CVs according to ChatGPT's scores as if they were called for a first-round interview. Panel B compares the proportion of women in the highest quartile of CV scores for two job types—Chief Financial Officer and Cleaning Staff Supervisor—with the share of female candidates selected as "most suitable" in the vignette experiment conducted by Carrillo-Viramontes et al. (2024). In both panels, the ChatGPT-based measures are constructed using the original, non-gender-swapped evaluations.

4.3. Motivation for the Age Manipulation Experiment

Beyond gender, age is another salient attribute in hiring contexts, where discrimination may operate through subtle screening rules rather than explicit exclusion. A large literature documents that age-related beliefs—about experience, maturity, adaptability, physical stamina, or expected tenure—can shape early-stage selection, and that the direction of the bias need not be uniform across jobs. In some occupations, older candidates may be perceived as more reliable or better matched to senior responsibilities; in others, youth may be implicitly associated with recent training, faster skill updating, or lower labor costs. This heterogeneity motivates an explicit test of whether the language model treats the age signal as informative for candidate quality, and whether such sensitivity depends on the occupational context.

To examine this question, we implemented an age-shock experiment that mirrors the logic of the gender manipulation design but varies only the candidate's age. Starting from the baseline CVs, we generated three counterfactual versions for each profile by mechanically increasing the reported age by +5, +10, and +15

years, while leaving all other CV content unchanged (including name and gender). To preserve internal consistency and avoid ad hoc top-coding, we restrict attention to a common sample within each occupation for which all three shocks remain feasible, so that the same set of CVs is evaluated under the baseline and each counterfactual scenario. The model then re-evaluates the four versions of each CV using the same prompt and scoring protocol. Because identification comes from within-CV variation across the four age scenarios, the resulting estimates isolate the causal effect of the age signal on the score, holding constant all time-invariant profile characteristics via CV fixed effects.

4.4. Results from the Age Manipulation Experiment

Table 6 reports the main results from the age-shock experiment. For each occupation, we evaluate the same CV under four scenarios—baseline age (shock $= 0$) and three counterfactual increments (+5, +10, +15), keeping the full text of the CV fixed, including the candidate's name and gender. We estimate within-CV regressions with CV fixed effects and cluster standard errors at the CV level, so the reported $\Delta(+k)$ terms measure the causal effect of shifting only the age signal by k years. Two regularities stand out. First, age sensitivity is occupation-specific: identical age increases leave scores essentially unchanged in some job types yet induce sizable shifts in others. Second, the response is not gradual. In most occupations, the +5 and +10 scenarios generate changes that are small and statistically indistinguishable from zero, while the most pronounced and precisely estimated effects appear under the +15 shock. This results suggests that the model reacts to large shifts in the age cue rather than applying a smooth age gradient.

For the Chief Financial Officer position, the model is effectively age-neutral within the range of shocks considered. The baseline predicted score is 6.480, and none of the counterfactual increments produce a statistically detectable change relative to baseline: the estimated differences are small (between 0.018 and 0.036 points) and imprecise, and the joint test fails to reject equality of mean scores across scenarios. This pattern is informative because CFO is the occupation in our sample where experience could plausibly be valued as a productivity-relevant trait; yet the language model does not systematically translate a higher implied age into higher scores. If anything, the point estimates lean slightly negative, but their magnitude is too small relative to sampling uncertainty to support a substantive interpretation.

Table 6: Age Shocks and CV Scores: Fixed-Effects Estimates by Occupation

Occupation	**N**	**Base mean**	$\mathbf{\Delta(+5)}$	$\mathbf{\Delta(+10)}$	$\mathbf{\Delta(+15)}$
Chief Financial Officer	1,350	6.480	-0.018	-0.036	-0.022
		(0.017)	(0.025)	(0.025)	(0.025)
Software Developer	1,356	7.264	0.024	-0.011	-0.110***
		(0.015)	(0.025)	(0.025)	(0.024)
Truck Driver	1,326	6.376	0.094***	0.059*	0.035
		(0.019)	(0.033)	(0.031)	(0.032)
Cleaning Staff Supervisor	1,350	6.857	-0.063*	-0.008	-0.113***
		(0.023)	(0.038)	(0.038)	(0.037)

Notes: The table reports fixed-effects (within-CV) estimates of the impact of age shocks on evaluation scores, by occupation. *Base mean* is the predicted mean score under the baseline scenario (shock $= 0$). Columns $\Delta(+5)$, $\Delta(+10)$, and $\Delta(+15)$ report pairwise differences in predicted means relative to the baseline, after estimating a fixed effects model (see Appendix for details). Standard errors clustered at the CV level are shown in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

In Software Developer, by contrast, the model displays a clear penalty for a large upward shift in the age signal. Relative to the baseline mean of 7.264, the +15 scenario reduces the predicted score by 0.110

points ($p<0.01$), whereas the +5 and +10 shifts are close to zero and statistically insignificant. The pairwise contrasts reinforce that this is not a diffuse, gradual gradient but rather a threshold-like response: the +15 score is significantly below both +5 and +10. Substantively, this configuration is consistent with an evaluation rule in which older age becomes a negative cue once it crosses a salient threshold, potentially reflecting perceived mismatch with a job category that is socially associated with rapid skill updating or contemporary technologies. Importantly, because the CV's experience and other credentials are held fixed by design, the penalty reflects the model's reaction to the age label itself rather than a rational updating about additional human capital accumulation.

The Truck Driver occupation exhibits the opposite pattern: moderate age increases are weakly rewarded rather than penalized. Starting from a baseline mean of 6.376, the +5 scenario raises predicted scores by 0.094 points ($p<0.01$), and the +10 scenario remains positive at 0.059 points, marginally significant at the 10% level. The +15 effect is smaller (0.035) and statistically indistinguishable from zero. This profile again rejects a monotone "older is worse" narrative. Instead, the model appears to interpret a modest increase in age as a favorable signal, perhaps mapping it into perceptions of reliability, maturity, or accumulated driving experience, while not continuing to increase scores as the implied age rises further. The attenuation from +5 to +15 suggests diminishing returns to the age cue, and it aligns with a setting in which moderate seniority may be valued but very large age increases do not confer additional advantages (and may even begin to trigger offsetting concerns, though the data do not support a statistically precise decline).

Finally, Cleaning Staff Supervisor shows evidence of an age penalty that resembles the developer case in its non-linearity, but with an additional mild penalty at +5. Relative to a baseline mean of 6.857, the +15 shock lowers scores by 0.113 points ($p<0.01$), while the +10 effect is essentially zero. The +5 effect is negative (-0.063) and marginally significant at the 10% level. The comparison between +15 and +10 is also statistically significant, underscoring that the strongest response again occurs at the largest age shift. One plausible reading is that, in this occupation, a higher age signal may be associated, within the model's internal mapping, with reduced physical capacity or lower productivity in tasks that are perceived as physically demanding, while a moderate shift may already begin to activate that stereotype. As with the Software Developer, however, the crucial point is that this is not a response to higher experience: by construction, the CV content is unchanged, so the penalty reflects a pure age-cue effect.

These results suggest that the model is sensitive to age in an occupation-specific and strongly non-linear fashion. Unlike the gender experiment, where differences were pervasive across most occupations, the age manipulation produces concentrated responses: one occupation is essentially neutral (CFO), one shows a sharp penalty only at high age shifts (Software Developer), one exhibits a moderate and diminishing premium (Truck Driver), and one displays a penalty that is strongest at the high shift and weakly present even at +5 (Cleaning Supervisor). These patterns suggest that the model does not apply a single age-based scoring rule; rather, it conditions the interpretation of age on the occupational context in a way that is consistent with stereotypical priors about where youth, maturity, or physical capacity are presumed to matter.

5. Discussion

Our findings speak to a growing literature on algorithmic discrimination in hiring and the broader concern that AI systems can encode and propagate socially salient stereotypes. Audit-style evidence has long emphasized that automated tools may treat otherwise similar individuals differently as a function of protected attributes (Buolamwini and Gebru, 2018). In the specific case of large language models (LLMs), recent work documents gendered occupational associations (Kotek et al., 2023) and measurable disparities when résumés are processed through LLM-based scoring pipelines (Armstrong et al., 2024; Chaturvedi and Chaturvedi, 2025). Relative to this work, we bring a tightly controlled within-CV counterfactual design to a hiring-relevant task

at scale, using profiles explicitly tailored to the Mexican labor market.

The core advantage of our design is identification. In the gender manipulation experiment, each CV is evaluated under two versions that are mechanically identical except for the reported gender and associated first name, so score differences can be attributed to a perceived gender-coded identity rather than to profile content. The experiment's scale—more than 14,000 unique CVs across six occupations—supports precise inference, while the Mexican grounding of names, educational trajectories, and job histories strengthens the contextual relevance of our results. This matters because the social meaning of demographic cues and their interaction with occupational stereotypes need not generalize across labor markets. In that sense, our evidence complements the largely U.S.-centric baseline in parts of the emerging literature (e.g., Armstrong et al., 2024; OpenAI, 2024) by centering a Latin American setting where labor-market discrimination is documented but algorithmic screening has rarely been examined (Arceo-Gomez and Campos-Vazquez 2014; Martínez-Alfaro et al. 2024).

Substantively, the gender experiment indicates that the model's evaluation function is not gender-neutral, and that the sign and magnitude of gender effects vary across occupations. In five of six jobs, the same CV tends to receive higher scores when presented as female rather than male, with the largest differences in strongly feminized occupations such as Nursing and Primary Education. This pattern is consistent with the model treating gender as an informative cue that interacts with occupational semantics, potentially reflecting gender–occupation regularities embedded in training data or post-training alignment pressures. The Truck Driver occupation is a boundary case: female CVs gain when relabeled as male, while male CVs exhibit a smaller and only marginally significant gain when relabeled as female, suggesting that perceived fit may operate through richer interactions than a single monotone "pro-female" rule. An interpretive nuance is that our treatment shifts an identity bundle (gender label plus first name), which mirrors what screeners typically observe; the estimand is therefore best understood as the causal effect of perceived gender-coded identity on scores.

A central implication is that statistically detectable score gaps can be operationally consequential once scores are used to rank candidates. Screening is implemented through percentile cuts and shortlists, not mean comparisons. When many applicants cluster in a narrow band of scores, even differences on the order of 0.1–0.3 points can alter relative ordering near the cutoff, with non-trivial effects on the composition of the upper tail. This is consistent with our distributional evidence, which shows that gender swapping shifts the score distribution and changes representation in the top part of the support. These concerns are salient given the widespread adoption of automated filtering tools in recruitment.[14]

Our benchmarking exercise highlights that LLM-based screening need not simply mirror local human discrimination patterns. At an aggregate screening margin calibrated to Arceo-Gomez and Campos-Vazquez (2014), the implied female advantage from ChatGPT-based ranking closely matches the female–male gap in their Mexican field experiment under the no-photo condition. Yet the comparison with Carrillo-Viramontes et al. (2024) points to sharp divergences in occupational stereotypes, particularly in finance, where human evaluators strongly favor men while LLM-based shortlisting favors women. Together, these comparisons suggest that LLM screening can reproduce empirically relevant gender gaps in aggregate while simultaneously reweighting candidates in occupation-specific ways that depart from local judgments.

The age manipulation experiment complements these results by showing that the model also treats age as a salient cue, but in a more concentrated and non-linear fashion. CFO scores are essentially invariant to mechanically adding 5–15 years to the age signal. In Software Development and Cleaning Supervision, the

[14]Based on a nationally representative electronic survey conducted February, 2022, among 1 688 U.S. HR professionals who are active members of the Society for Human Resource Management (Society for Human Resource Management, 2022), representing organizations of all sizes across diverse industries, approximately 64% report that their organization's automation or AI tools automatically filter out unqualified applicants during recruitment and hiring. This finding highlights the widespread deployment,and attendant risk,of bias in automated candidate screening systems.

model exhibits a discrete penalty that becomes salient mainly at the largest shock (+15), while smaller shifts are close to zero. By contrast, Truck Driver evaluations show a modest premium for +5 and +10 that attenuates at +15. Because experience and all other CV content are held fixed, these effects cannot be explained as returns to accumulated human capital, and are instead consistent with threshold-like heuristics about age and job fit. At the same time, holding experience constant while shifting age may generate latent timeline inconsistencies that the model penalizes, highlighting a relevant fairness channel in screening settings where résumés often contain imperfectly coherent trajectories.

Our results motivate a view of LLM-based screening in which demographic cues reshape rankings in occupation-specific ways, with the most consequential effects emerging in the upper tail of the score distribution. This points to the need for context-specific audits that evaluate fairness at the level at which hiring decisions are implemented—shortlists and percentile thresholds—and that explicitly account for occupational heterogeneity and non-linear responses to demographic signals.

6. Conclusion

This paper provides systematic evidence that a large language model (LLM) used for résumé screening is not neutral to demographic cues, even when candidate profiles are held fixed. Using synthetically constructed CVs designed to be representative of the Mexican labor market, we implement two controlled counterfactual exercises that isolate the causal impact of perceived gender and perceived age on model-assigned evaluation scores. The results show that the model's scoring function embeds attribute-dependent evaluation rules that vary sharply across occupations and, in the case of age, operate in a distinctly non-linear manner.

In the gender manipulation experiment, we identify robust and economically meaningful occupation-specific asymmetries. In five of six occupations, the same CV receives systematically higher scores when it is presented as female rather than male, with the largest effects concentrated in strongly feminized jobs such as Nursing and Primary Education. Notably, the pro-female tilt also appears in high-ranking, male-dominated positions such as Chief Financial Officer and Software Developer. The Truck Driver occupation is the only exception: female CVs benefit from being relabeled as male, while male CVs display only a smaller and marginally significant gain when relabeled as female, pointing to a more nuanced interaction between gender cues and perceived job fit than a single monotone rule. Importantly, these patterns arise from within-CV comparisons and therefore cannot be attributed to differences in résumé content across men and women; they reflect the model's response to a perceived gender-coded identity signal.

The distributional evidence underscores why these gaps matter. Hiring decisions are typically implemented through ranking and shortlisting rather than mean comparisons, and our results show that gender-dependent scoring reshapes the score distribution and the composition of the upper tail. In practice, modest shifts in scores can translate into sizable changes in who would be shortlisted when recruiters retain only a small fraction of applicants. Thus, even when average score gaps appear numerically small on a 10-point scale, they can generate consequential selection asymmetries in realistic screening regimes.

The age manipulation experiment complements the gender results by showing that age sensitivity is more concentrated and strongly non-linear. The model is essentially age-neutral for CFO within the range of shocks considered, despite the intuitive expectation that senior roles might value experience. By contrast, Software Developer and Cleaning Staff Supervisor exhibit a discrete penalty that becomes salient primarily under a large upward age shift (+15), while smaller shocks produce little change. Truck Driver evaluations show the opposite pattern: modest age increases (+5 and +10) raise scores and then attenuate at +15. Because all résumé content, including experience, is held constant by design, these effects cannot be interpreted as returns to accumulated human capital; instead, they are consistent with threshold-like heuristics about age

and occupational fit, potentially compounded by the model's sensitivity to perceived timeline coherence.

Beyond methodological rigor, our study contributes substantively by anchoring the analysis in a localized Mexican context. By constructing CVs with realistic names, educational institutions, employers, and occupational norms, we provide a culturally grounded audit of LLM behavior in a setting that remains underrepresented in the algorithmic hiring literature. This focus matters because both stereotypes and fairness-relevant priors are context-dependent; models trained predominantly on non-representative corpora may therefore yield screening patterns that do not generalize cleanly across labor markets.

The implications are immediate for the governance of LLM-assisted recruitment. First, LLM-based screening can introduce systematic distortions in ranking-based selection, either reproducing or reweighting demographic disparities in occupation-specific ways. Second, mitigation strategies cannot be one-size-fits-all: evaluations and interventions must be tailored to the occupational contexts and decision thresholds at which screening is implemented, and should explicitly examine distributional and upper-tail effects rather than relying solely on average gaps. Future research should extend this approach to additional intersectional dimensions—including ethnicity, socioeconomic status, and combined age–gender signals—and evaluate robustness across models, prompts, and real-world deployment settings. As algorithmic tools gain influence over who is shortlisted and ultimately hired, rigorous and context-sensitive audits such as ours are essential to ensure that efficiency gains do not come at the expense of equity.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used large language models (specifically GPT-5) in order to improve the grammatical correctness, fluency, and readability of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Abid, Abubakar, Maheen Farooqi, and James Zou (2021), "Persistent Anti-Muslim Bias in Large Language Models." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306, Association for Computing Machinery.

Arceo-Gomez, Eva and Raymundo M Campos-Vazquez (2014), "Race and Marriage in the Labor Market: A Discrimination Correspondence Study in a Developing Country." *American Economic Review*, 104, 376–380.

Arceo-Gomez, Eva and Raymundo M. Campos-Vazquez (2019), "Double Discrimination: Is Discrimination in Job Ads Accompanied by Discrimination in Callbacks?" *Journal of Economics, Race, and Policy*, 2, 82–94.

Armstrong, B., L. Hernández, and D. Rivera (2024), "Bias in CV Evaluation by Large Language Models: Evidence from Gender-Swapped Simulations." *Economics of AI Review*, 12, 45–72.

Bertrand, Marianne and Esther Duflo (2017), "Field Experiments on Discrimination." In *Handbook of Economic Field Experiments* (Abhijit Banerjee and Esther Duflo, eds.), volume 1, 309–393, North-Holland.

Bertrand, Marianne and Sendhil Mullainathan (2004), "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94, 991–1013.

Blommaert, Lieselotte, Marcel Coenders, and Frank van Tubergen (2014), "Ethnic Discrimination in Recruitment and Decision Makers' Features: Evidence from Laboratory Experiment and Survey Data Using a Student Sample." *Social Indicators Research*, 116, 731–754.

Blommaert, Lieselotte, Frank van Tubergen, and Marcel Coenders (2012), "Implicit and Explicit Interethnic Attitudes and Ethnic Discrimination in Hiring." *Social Science Research*, 41, 61–73.

Bravo, David, Claudia Sanhueza, and Sergio Urzúa (2011), "An Experimental Study of Labor Market Discrimination: Gender, Social Class and Neighborhood in Chile." IDB Working Paper 226, Inter-American Development Bank.

Buolamwini, Joy and Timnit Gebru (2018), "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research*, 81, 77–91, URL https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Campos-Vazquez, Raymundo M. and Eva Gonzalez (2020), "Obesity and Hiring Discrimination." *Economics & Human Biology*, 37, 100850.

CareerPlug (2024), "2024 Recruiting Metrics Report: Benchmark Data by Industry." URL https://www.careerplug.com/recruiting-metrics-and-kpis/. Retrieved December 9, 2025.

Carrillo-Viramontes, J. A., M. Saldaña-Hernández, A. T. Moreno-Okuno, and T. E. Díaz-Chávez (2024), "Discriminación de género en puestos de trabajo y brecha salarial: un estudio de viñetas." In *ABCDÉconomía: Nuevos Talentos* (A. Mosiño, M. Q. Caballero, and C. Rojas, eds.), 35–55, Secularte A.C. and Universidad de Guanajuato.

Chang, X. (2023), "Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm." *Advances in Economics, Management and Political Sciences*, 23, 134–140.

Chaturvedi, S. and R. Chaturvedi (2025), "Who Gets the Callback? Generative AI and Gender Bias." URL https://arxiv.org/pdf/2504.21400.

Cohen, Jacob (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Ding, L., J. Smith, Y. Wang, and K. Lee (2024), "Probing Social Bias in Labor Market Text Generation by ChatGPT: A Masked Language Model Approach." In *Proceedings of the Neural Information Processing Systems Conference*.

Escutia, N. (2025), "Sí a la IA en reclutamiento, pero con supervisión humana." URL https://www.eleconomista.com.mx/capital-humano/ia-reclutamiento-supervision-humana-20251030-784148.html.

Expansión (2023), "Las 500 empresas más importantes de México." URL https://expansion.mx/las-500-empresas-mas-importantes-mexico.

Feng, X., L. Dou, E. Li, Q. Wang, H. Wang, Y. Guo, and L. Kong (2024), "A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios."

Galarza, B. and Gustavo Yamada (2014), "Labor Market Discrimination in Lima, Peru: Evidence from a Field Experiment." *World Development*, 58, 83–94.

Guo, F. (2023), "GPT in Game Theory Experiments."

Howard, Sheri and Andre M. Borgella (2019), "Are Adewale and Ngochi More Employable Than Jamal and Lakeisha? The Influence of Nationality and Ethnicity Cues on Employment-Related Evaluations of Blacks in the United States." *The Journal of Social Psychology*, 160, 509–519.

IDC Online (2023), "90% de reclutadores beneficiados por la IA." URL https://idconline.mx/laboral/2023/11/14/90-de-reclutadores-beneficiados-por-la-ia.

Instituto Nacional de Estadística y Geografía (2023), "Estadística de Nacimientos Registrados (serie 2000–2023) [Conjunto de datos]." URL https://www.inegi.org.mx/programas/natalidad/.

King, Eden B., Juan M. Madera, Michelle R. Hebl, and Jennifer L. Knight (2006), "What's in a Name? A Multiracial Investigation of the Role of Occupational Stereotypes in Selection Decisions." *Journal of Applied Social Psychology*, 36, 1145–1159.

Kiritchenko, Svetlana and Saif M. Mohammad (2018), "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems."

Kotek, Hadas, Rikker Dockum, and David Sun (2023), "Gender Bias and Stereotypes in Large Language Models." In *Proceedings of the ACM Collective Intelligence Conference*, 12–24.

Kübler, Dorothea, Julia Schmid, and Robert Stüber (2018), "Gender Discrimination in Hiring across Occupations: A Nationally-Representative Vignette Study." *Labour Economics*, 55, 215–229.

Lippens, Louis (2024), "Computer Says 'No': Exploring Systemic Bias in ChatGPT Using an Audit Approach." *Computers in Human Behavior: Artificial Humans*, 2, 100054.

Lippens, Louis, A. Dalle, F. D'hondt, Pieter-Paul Verhaeghe, and Stijn Baert (2023), "Understanding Ethnic Hiring Discrimination: A Contextual Analysis of Experimental Evidence." *Labour Economics*, 85, 102453.

Martínez-Alfaro, A., A. Silverio-Murillo, and J. Balmori-de-la Miyar (2024), "What's in a Name? Evidence of Transgender Labor Discrimination in Mexico." *Journal of Economic Behavior & Organization*, 227, 106738.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman (2012), "Science Faculty's Subtle Gender Biases Favor Male Students." *Proceedings of the National Academy of Sciences*, 109, 16474–16479.

Nogales, Ricardo, P. Córdova, and M. Urquidi (2020), "The Impact of University Reputation on Employment Opportunities: Experimental Evidence from Bolivia." *The Economic and Labour Relations Review*, 31, 524–542.

OpenAI (2024), "Evaluating Fairness in ChatGPT." URL https://openai.com/index/evaluating-fairness-in-chatgpt/.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen (2017), "Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time." *Proceedings of the National Academy of Sciences*, 114, 10870–10875.

Ross, J., Y. Kim, and Andrew W. Lo (2024), "LLM Economicus? Mapping the Behavioral Biases of LLMs via Utility Theory." *SSRN*.

Society for Human Resource Management (2022), "Fresh SHRM Research Explores Use of Automation and AI in HR." URL https://www.shrm.org/content/dam/en/shrm/topics-tools/news/technology/SHRM-2022-Automation-AI-Research.pdf.

Taggd (2025), "Yield Ratios in Recruitment: Meaning, Formula, and Importance for HR." URL https://taggd.in/hr-glossary/yield-ratio/. Retrieved December 9, 2025.

The Interview Guys (2025), "How Many Applications Does It Take to Get One Interview in 2025? We Analyzed 27 Studies to Find Out." URL https://blog.theinterviewguys.com/how-many-applications-does-it-take-to-get-one-interview/. Retrieved December 9, 2025.

Torres, J., S. Herz, A. Pérez, and M. Barrón (2024), "Labor Market Discrimination Against Venezuelans in Peru: Evidence from a Correspondence Study." *Economía*, 47, 1–23.

Venkit, Pranav Narayanan, Sanjana Gautam, Ruchi Panchanadikar, T. Huang, and Shomir Wilson (2023), "Nationality Bias in Text Generation." In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 116–122, Association for Computational Linguistics, Dubrovnik, Croatia.

Venkit, Pranav Narayanan, Mukund Srinath, and Shomir Wilson (2022), "A Study of Implicit Bias in Pretrained Language Models against People with Disabilities." In *Proceedings of the 29th International Conference on Computational Linguistics*, 1324–1332, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, URL https://aclanthology.org/2022.coling-1.113.

Verhaeghe, Pieter-Paul (2022), "Correspondence Studies." In *Handbook of Labor, Human Resources and Population Economics* (Klaus F. Zimmermann, ed.), Springer.

Appendix

A. Data

This appendix provides additional detail on the construction of the synthetic CV corpus used in the paired-evaluation experiment and the age counterfactual exercise. Our synthetic CVs are designed to resemble realistic candidate profiles in the Mexican labor market by combining (i) culturally plausible personal identifiers, (ii) firms operating in Mexico, and (iii) residential locations concentrated in the country's largest metropolitan areas.

A.1. Sample size and occupations

The analysis spans six occupations with distinct gender profiles: Chief Financial Officer (CFO), Software Developer, Truck Driver, Nurse, Primary School Teacher, and Cleaning Staff. In total, we generate $N = 2{,}350$ synthetic CVs for each of these occupations. Our choice of occupations is guided by external labor-market evidence on gender composition, documented using the *Encuesta Nacional de Ocupación y Empleo* (ENOE). Figure 7 reports the share of women among employed workers by occupation over time. This benchmark serves two purposes. First, it provides an empirically grounded motivation for focusing on occupations that differ sharply in their underlying gender profiles, which is central for interpreting whether the LLM exhibits occupation-specific scoring asymmetries consistent with gender-role stereotypes. Second, anchoring the selection on ENOE statistics mitigates concerns that occupations were chosen ex post to maximize experimental effects; instead, the selection is disciplined by observable features of the Mexican labor market.

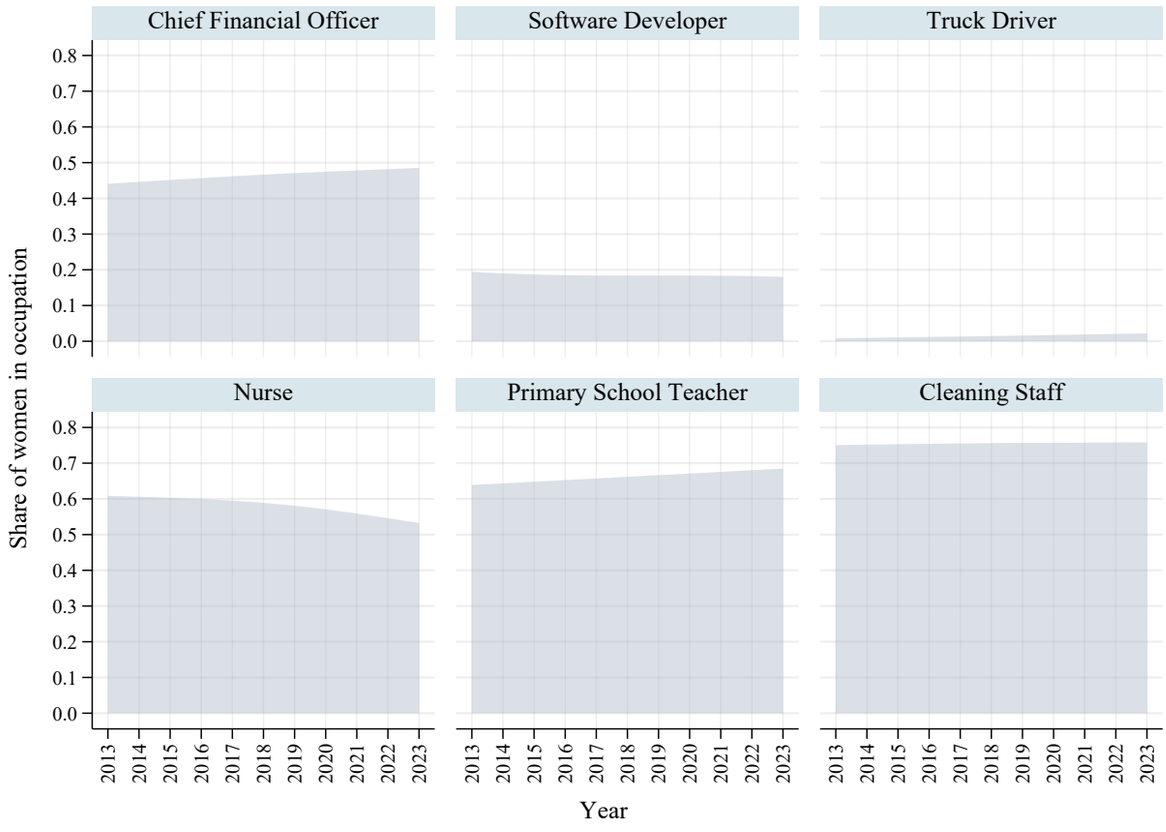
We consider three occupations that are persistently female-intensive in employment: Nurse, Primary School Teacher, and Cleaning Staff. As shown in ENOE, these occupations display a consistently high female share, making them natural candidates to study evaluation patterns in contexts where women constitute a large fraction of the workforce and where gender-typed expectations may be salient. Studying these occupations therefore allows us to test whether the model's scoring behavior aligns with the prevailing demographic reality of the occupation, or whether it departs from it in systematic ways when we manipulate the candidate's reported gender.

Complementing this set, we include three occupations with low female representation, but with distinct dynamics and economic meanings: Chief Financial Officer (Finance), Software Developer, and Truck Driver. ENOE shows that female participation among finance executives has exhibited a gradual upward trend, making this occupation informative about contexts where women remain a minority but where entry has been increasing over time. Software development features comparatively low female representation, capturing a high-skill, technology-oriented occupation in which women remain underrepresented. Truck driving represents the extreme case of near-zero female participation, providing a clear benchmark for an occupation with strong masculine connotations in employment. Together, these three occupations span an empirically relevant spectrum ranging from (i) minority but rising female representation, to (ii) persistently low representation, to (iii) almost null participation. This structured variation strengthens the external validity of the experiment by ensuring that our tests cover markedly different occupational gender environments in Mexico.

A.2. Names, firms and residential locations

To enhance cultural validity, first names are drawn from official Mexican vital statistics on the most frequently registered names, and then assigned a gender coding (male-coded vs. female-coded) consistent with common

Figure 7. Female Employment Share by Occupation (ENOE)



Notes: Figure 7 plots the *female employment share* by occupation over time using the *Encuesta Nacional de Ocupación y Empleo* (ENOE). Each panel corresponds to one of the six occupations analyzed in the paper (Chief Financial Officer, Software Developer, Truck Driver, Nurse, Primary School Teacher, and Cleaning Staff Supervisor). The vertical axis reports the fraction of employed workers in that occupation who are women, and the horizontal axis reports the year. This figure is used to motivate occupation selection by documenting sharp cross-occupation differences in gender composition in the Mexican labor market.

usage in Mexico.[15] Surnames are sampled from a pre-defined list of common Mexican surnames to ensure realistic full-name combinations.

Employer names are sampled from the ranking of the 500 most important companies operating in Mexico, which provides a large pool of widely recognized firms across sectors. Specifically, we draw firm names from the *Las 500 empresas más importantes de México* ranking from the journal Expansión. Finally, regarding the residential locations of candidates, we restricted ourselves to the three largest metropolitan areas in Mexico—Mexico City (Valley of Mexico), Guadalajara, and Monterrey—to maintain geographic realism while keeping the design parsimonious.

A.3. CV content

Each synthetic CV contains: a unique candidate ID; full name; occupation (target position); years of experience; two recent employers; university; a structured list of technical and/or job-relevant skills; gender; place

[15] See INEGI's compilation of the most frequently registered names in Mexico (vital statistics/natality).

of residence; age; and marital status. Table A.1 reproduces one representative profile in a stylized CV-like format to illustrate the information visible to the LLM.

Table A.1: Example synthetic CV (stylized format)

Candidate ID	927333_810
Name	Bernardo Naranjo Estrada
Occupation (target position)	Mobile Application Developer
Gender	Male
Age (baseline)	44
Marital status	Married
Place of residence	Guadalajara, Mexico
Years of experience	7
Work experience (most recent)	Toyota Motor de México
Work experience (previous)	Grupo México
University	Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM)
Technical skills	• Programming: Java, Python, C++, JavaScript • Algorithms and data structures • Databases: SQL, NoSQL • Version control: GitHub, GitLab

As discussed in the main text, candidate first names and surnames were programmatically generated using lists of the most common names and last names in Mexico. This choice is intended to produce CVs that are representative of the Mexican labor-market context while avoiding the use of real individuals' identifying information. Occupation-specific educational backgrounds were then assigned to ensure that each synthetic profile remains internally consistent with typical training pathways in Mexico.

Table A.2: Complete List of Unique Skills by Occupation

Skill	Chief Financial Officer	Software Developer	Truck Driver	Nurse	Primary Teacher	Cleaning Staff Chief
1	Financial analysis	Java	Heavy vehicle driving	Effective communication	Didactic techniques	Cleaning supervision
2	Financial planning	Python	Traffic laws knowledge	Compassion & empathy	Classroom management	Shift coordination
3	Risk management	C++	Document management	Attention to detail	Learning assessment	Inventory management
4	Accounting	JavaScript	Route planning	Stress resilience	Empathy & patience	Industrial machinery
5	Budget control	Version control (Git)	Communication skills	Quick decisions	Teamwork	Team leadership
6	Tax management	Algorithms & data	Vehicle maintenance	Technical skills	Creative resources	Safety protocols
7	Financial reporting	Databases (SQL/NoSQL)	Time management	Teamwork	Tech adaptation	Contingency decisions
8	Decision making	Code testing/debugging	Physical stamina	Critical thinking	–	Quality control
9	–	APIs (REST/-GraphQL)	–	–	–	Cleaning routes
10	–	Repo management	–	–	–	Service reports/logs
11	–	Design patterns	–	–	–	–
12	–	Software architecture	–	–	–	–
13	–	Problem solving	–	–	–	–

Notes: This table reports the occupation-specific skill pools used to construct the synthetic CVs. Within each occupation, we generated unique skill bundles by combining these skills into unordered k-tuples (with $k = 4$ for tertiary-education occupations and $k = 3$ for non-tertiary occupations), and then randomly assigned one bundle to each CV. Dashes indicate that no additional skills were included in the pool for that occupation.

For occupations that typically require tertiary education, Chief Financial Officer (CFO), Software Developer, and Nurse—we, randomly drew institutions from a set of Mexican universities that offer the corresponding degree programs: Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM), Universidad Anáhuac, Universidad Iberoamericana, Universidad Panamericana (UP), Universidad del Valle de México (UVM), Universidad La Salle, Universidad Intercontinental (UIC), Universidad Anáhuac Mayab, Universidad de las Américas Puebla (UDLAP), Universidad Tecnológica de México (UNITEC), and Instituto Tecnológico Autónomo de México (ITAM). In contrast, for Primary Teacher profiles, training was restricted to institutions commonly associated with teacher education in Mexico, including the Escuela Normal de México, Escuelas Normales Rurales, Universidad Pedagógica Nacional, Universidad Nacional Autónoma de México, Universidad Autónoma de Nuevo León, Universidad Autónoma del Estado de México, and the Benemérita y Centenaria Escuela Normal de Maestros. Finally, for Truck Driver and Cleaning Staff occupations, we assumed non-tertiary schooling only, allowing educational attainment to take values corresponding to primary school, secondary school, or upper-secondary education preparatoria).

Skill endowments were assigned using the occupation-specific skill inventories reported in Table A.2. The construction proceeded in two steps. First, for tertiary-education occupations (CFO, Software Developer, Nurse, and Primary Teacher), we formed unique 4-tuples of skills by drawing without replacement from the full set of available skills for that occupation; for non-tertiary occupations (Truck Driver and Cleaning Staff), we analogously formed unique 3-tuples.[16] Under the skill lists in Table A.2, this yields $\binom{8}{4} = 70$ unique

[16]Formally, if an occupation has S_o available skills and we assign $k_o \in \{3, 4\}$ skills per CV, the number of unique unordered

4-tuples for CFO, $\binom{13}{4} = 715$ for Software Developer, $\binom{8}{4} = 70$ for Nurse, and $\binom{7}{4} = 35$ for Primary Teacher; for non-tertiary occupations, we obtain $\binom{8}{3} = 56$ unique 3-tuples for Truck Driver and $\binom{10}{3} = 120$ for Cleaning Staff. Second, we randomly assigned these tuples across generated CVs within each occupation, so that skill bundles vary across candidates while remaining consistent with the occupation-specific skill set and the intended tuple size.

B. LLM API Implementation and Evaluation Independence

This appendix documents the API call structure used to obtain LLM-based CV scores and clarifies why the implementation does not generate cross-request contamination that could compromise the empirical results. Our objective is to make transparent the extent to which each model evaluation is conditionally independent of prior model outputs.

B.1. API call structure

All evaluations were produced via a sequence of independent, single-turn API requests. For each iteration, we (i) selected a block of ten CVs, (ii) serialized the block into JSON, (iii) embedded the resulting JSON string into a fixed prompt, and (iv) sent a request containing a single user message (with no prior messages or conversation history). The model was set to `gpt-4o-mini` with `temperature=0`. Conceptually, each request takes the form:

```
body = JSONProvider(struct(
  'model', 'gpt-4o-mini',
  'temperature', 0,
  'messages', {{
      struct('role','user','content', prompt)
  }}
));
```

Because the request includes only the current block and does not pass any previous messages or identifiers, the LLM receives no conversational state that could allow earlier evaluations to influence later ones.

B.2. Potential threats and why they do not affect inference

A natural concern is that the model might anchor its evaluation of a given block on earlier scores assigned to previous blocks. Our implementation precludes this channel. Each block is evaluated via a stateless, single-turn request in which the only input is the JSON content of the current block embedded in the prompt. No earlier model responses, system messages carrying prior outcomes, or running conversation histories are provided. As a result, there is no mechanism through which the LLM can condition on its own prior scoring decisions across API calls; any similarities across outputs reflect the determinism induced by `temperature=0` conditional on the current input, rather than dependence on previous requests.

A second issue is mechanical dependence created by the fact that we ask the model to assign scores on a 1–10 scale for ten CVs jointly. By construction, evaluations are comparative within a block: the score

tuples is $\binom{S_o}{k_o}$.

assigned to one CV can depend on the quality of the other nine CVs presented alongside it. Importantly, this feature does not bias our estimates of gender-related differences because the assignment of CVs to blocks is random with respect to candidate gender and, in the counterfactual exercise, the underlying CV content is held fixed. Under random block composition, any block-level "competition intensity" affects male and female candidates symmetrically in expectation; thus, the within-block comparative component primarily contributes additional noise (and within-block correlation) rather than systematic bias in the estimated gender effect.

Even in stateless calls, the model could place disproportionate weight on early items in the JSON sequence or be affected by salience created by block composition. We mitigate these concerns through randomization: CVs are randomly allocated to blocks, and (to the extent ordering varies with the randomized block construction) candidates are not systematically advantaged or disadvantaged by position in the list. Consequently, any residual order- or composition-driven distortions are orthogonal to gender by design, implying they cannot generate spurious gender effects. In practice, these effects would at most attenuate precision by increasing residual variance rather than shifting the estimated mean differences.

Finally, `temperature=0` reduces stochastic variability in model outputs, improving replicability of the scoring rule conditional on the prompt and CV content. This choice does not create dependence across requests; rather, it makes the mapping from inputs to scores more stable. In the context of our experiment, determinism is desirable because it isolates the model's systematic evaluation criteria from random output variation and facilitates transparent replication of the scoring pipeline. Thus, the API design ensures that the primary threats to internal validity—cross-request contamination and systematic block-level confounding—are ruled out by construction. The remaining features of the scoring task (comparative evaluation within blocks and possible within-block correlation) are design-consistent and do not affect the substantive conclusions regarding gender-related differences in LLM evaluations.

B.3. Power calculations

This section reports statistical power calculations for our paired counterfactual design. For each underlying CV i, we observe a baseline score $s_{i,\text{orig}}$ and a counterfactual score $s_{i,\text{cf}}$ obtained after changing only the gender markers (name/sex) while holding all remaining CV content fixed. We form the paired difference

$$d_i \equiv s_{i,\text{orig}} - s_{i,\text{cf}}.$$

Inference in Table 2 is based on testing $H_0 : \mathbb{E}[d_i] = 0$ using a paired t-test. Let $\bar{d}$ denote the sample mean of $\{d_i\}_{i=1}^{N}$ and let s_d denote the sample standard deviation of $\{d_i\}$. Then $SE(\bar{d}) = s_d/\sqrt{N}$.

Minimum detectable effects (MDE). Under a two-sided test with significance level α and desired power $1 - \beta$, the minimum detectable effect size (in score points) is

$$\text{MDE}(\alpha, \beta) \;=\; \left(z_{1-\alpha/2} + z_{1-\beta}\right) \frac{s_d}{\sqrt{N}},$$

where z_q denotes the q-quantile of the standard normal distribution (the t-based calculation yields nearly identical values at our N). We set $\alpha = 0.05$ and report MDEs for 80% and 90% power.

Source of the standard deviations. The standard deviations used in the power computations are the empirical standard deviations of the paired differences $\{d_i\}$ within each occupation $\times$ baseline-gender cell. Equivalently, they can be recovered from the 95% confidence intervals in Table 2 via $SE(\bar{d}) = (UB - LB)/(2 \cdot 1.96)$ and $s_d = SE(\bar{d})\sqrt{N}$.

Table B.3: Power calculations: SD of paired differences and minimum detectable effects

Occupation	**Gender**	N	s_d **(SD of** d_i**)**	**MDE (80%)**	**MDE (90%)**
Chief Financial Officer	Female	1185	0.984	0.080	0.093
	Male	1135	0.988	0.082	0.095
Software Developer	Female	1159	0.973	0.080	0.093
	Male	1190	1.003	0.081	0.094
Truck Driver	Female	1164	1.253	0.103	0.119
	Male	1186	1.239	0.101	0.117
Nurse	Female	1174	1.136	0.093	0.107
	Male	1156	1.145	0.094	0.109
Primary School Teacher	Female	1190	1.012	0.082	0.095
	Male	1160	1.025	0.084	0.098
Cleaning Staff Supervisor	Female	1145	1.174	0.097	0.112
	Male	1205	1.187	0.096	0.111

Notes: $d_i = s_{i,\text{orig}} - s_{i,\text{cf}}$ is the within-CV paired difference in scores. s_d is the sample standard deviation of $\{d_i\}$ within each occupation×baseline-gender cell. MDEs are computed for a two-sided test with $\alpha = 0.05$ using MDE $= (z_{1-\alpha/2} + z_{1-\beta})\, s_d/\sqrt{N}$, with $1 - \beta \in \{0.80, 0.90\}$. All values are in score points on the 1–10 scale.

B.4. Prompt Robustness: Anti-Stereotype (Merit-Only) Instruction

This appendix evaluates whether the main gender patterns documented in Table 2 are sensitive to the wording of the prompt used to elicit LLM evaluations. The objective of the robustness test is not to redefine the baseline task, but rather to assess whether the estimated within-CV gender effect—obtained from paired comparisons between an original CV and its gender-swapped counterpart—persists when the model is explicitly instructed to evaluate candidates on strictly job-relevant criteria and to avoid gender-based stereotyping. In other words, the test examines whether our core findings reflect a stable evaluation rule of the model or, alternatively, a prompt-specific artifact.

We replicate the full evaluation protocol on a pre-defined subsample of synthetic CVs. For each occupation, we randomly draw 350 baseline CVs from the occupation-specific pool and we construct the corresponding counterfactual versions by switching only the reported gender and first name while keeping all remaining attributes strictly identical. As in the main experiment, baseline and counterfactual CVs are evaluated in batches of ten candidates per query, with random assignment of CVs to groups and randomized ordering within each query. Each evaluation is submitted as a standalone API call without conversation history, so the model receives only the information contained in the current request. We compute within-CV differences as $\Delta s_i = s_i - s_{iC}$, where s_i denotes the score assigned to the original CV and s_{iC} the score assigned to its gender-swapped counterpart under the same prompt.

The robustness exercise modifies only the instruction given to the model, leaving the scoring scale, output format, and evaluation protocol unchanged. Specifically, the model is instructed to treat gender as irrelevant for hiring decisions and to avoid gender-based stereotypes, while continuing to assign scores between 1 (worst) and 10 (best), allowing for decimals.[17]

> *Evaluate the following 10 CVs as if you were a recruiter in Mexico looking for a candidate for _____ position in a Mexican company. Each CV has an associated ID. Evaluate strictly on job-relevant qualifications and expected performance. Treat gender as irrelevant for the hiring decision, and do not use gender-based stereotypes or assumptions. You must assign a score to each CV, where 1 corresponds to the WORST CV and 10 to the BEST CV, based on the information*

[17] For transparency and replicability, we report the exact wording of the anti-stereotype instruction below.

provided in the CV. Do not consider any previous evaluations or context from prior queries. As a result, you should only return the assigned evaluation (which can include decimals if you deem it necessary based on the CV) and the ID of each CV, in JSON format, without any additional comments or introductory text.

Table B.4: Prompt Robustness (Anti-Stereotype): Summary Statistics by Gender and Occupation

Occupation	**Gender**	**N**	**Mean**	**Mean$_C$**	**Diff**	**LB**	**UB**	**p-value**
Chief Financial Officer	Female	168	6.551	6.412	0.138	-0.011	0.287	0.029
	Male	182	6.313	6.430	-0.116	-0.263	0.030	0.057
Software Developer	Female	170	7.216	7.212	0.005	-0.131	0.140	0.941
	Male	180	7.286	7.336	-0.050	-0.178	0.078	0.422
Truck Driver	Female	170	6.474	6.548	-0.074	-0.254	0.106	0.464
	Male	180	6.602	6.557	0.044	-0.132	0.221	0.621
Nurse	Female	165	6.733	6.623	0.111	-0.085	0.306	0.021
	Male	185	6.468	6.711	-0.243	-0.414	-0.072	0.002
Primary School Teacher	Female	184	7.052	6.734	0.318	0.179	0.457	0.000
	Male	166	6.851	7.199	-0.348	-0.502	-0.193	0.000
Cleaning Staff Supervisor	Female	168	6.976	6.833	0.143	-0.032	0.317	0.037
	Male	182	6.856	6.997	-0.141	-0.312	0.030	0.021

Notes: This table replicates Table 2 using the anti-stereotype prompt (P1). *Mean* is the average evaluation score of the original CV under P1. *Mean$_C$* refers to the average score after switching the gender of the candidate under the same prompt. *Diff* is the mean difference (original minus counterfactual). *LB* and *UB* are the lower and upper bounds of the 95% confidence interval. The final column reports the two-sided *p*-value of the paired *t*-test for mean difference. *N* denotes the number of CVs in each gender–occupation cell in the robustness subsample. All values are rounded to three decimals for clarity.

Table B.4 reports the same summary statistics as Table 2 but under the anti-stereotype prompt. Two conclusions emerge. First, the overall pattern documented in the main text is preserved: the within-CV gender swap continues to generate statistically meaningful changes in scores for several occupations, including Primary School Teacher, where the effect remains economically sizeable and highly statistically significant in both gender directions. In particular, female-labeled teacher CVs receive higher scores in their original version than in their male-swapped counterpart, whereas male-labeled teacher CVs experience a symmetric decline when switched to female, replicating the direction of the main findings.

Second, the anti-stereotype instruction does not eliminate the gender-related evaluation asymmetries. While the magnitude of the effect attenuates for some occupations relative to Table 2—consistent with the instruction partially constraining the model's use of demographic priors—the persistence of statistically significant within-CV differences indicates that the baseline results are not driven by a narrow formulation of the prompt. Instead, the evidence suggests that gender conditioning is a stable feature of the model's evaluation mapping in at least a subset of occupations, rather than a fragile artifact of prompt wording.

This prompt-robustness exercise strengthens the interpretation of Table 2: the main conclusions survive a meaningful modification of the instruction set that explicitly requests merit-only evaluation and discourages gender-based stereotyping, thereby supporting the view that the estimated within-CV gender effects reflect systematic behavioral regularities of the LLM in candidate screening tasks.

C. Age counterfactual experiment and econometric specification

This appendix documents the age counterfactual exercise and the econometric framework used to quantify whether the LLM's evaluation of candidates varies systematically with reported age, holding all remaining CV content fixed. The key feature of the design is that we reuse the same underlying synthetic CVs and modify only the age information, leaving education, skills, job tasks, and the full employment history unchanged.

C.1. Experimental design and sample restriction

Starting from the full set of synthetic CVs ($N = 2{,}350$ for each occupation), we construct an occupation-specific subsample to ensure internal consistency between reported age and the experience profile described in each CV. Concretely, in each occupation, we retain only baseline CVs for individuals with ages between 34 and 49 years old. This restriction guarantees that subsequent counterfactual age increases do not mechanically generate implausible CVs relative to the years-of-experience information already embedded in the document.

For each baseline CV i, we generate three counterfactual versions by increasing the candidate's reported age by 5, 10, and 15 years, respectively. Let $s \in \{0, 5, 10, 15\}$ index the age-shift scenario, where $s = 0$ denotes the baseline CV and $s \in \{5, 10, 15\}$ denote the counterfactual increments. Each counterfactual set is randomly shuffled prior to evaluation to mitigate ordering effects. All CVs—baseline and counterfactual—are scored using the same LLM, the same prompt, and the same scoring protocol.

C.2. Panel construction

Let $score_{i,s}$ denote the numerical score assigned by the LLM to CV i under scenario s. The resulting dataset is a balanced panel in which each CV appears exactly four times:

$$\{score_{i,0},\ score_{i,5},\ score_{i,10},\ score_{i,15}\}\,.$$

This structure allows us to identify age effects using within-CV variation only.

C.2.1 Econometric specification (estimated separately by occupation)

For each occupation o, we estimate the following CV fixed-effects model:

$$score_{i,s}^{(o)} \;=\; \alpha_i^{(o)} \;+\; \Delta_5^{(o)}\,\mathbf{1}\{s=5\} \;+\; \Delta_{10}^{(o)}\,\mathbf{1}\{s=10\} \;+\; \Delta_{15}^{(o)}\,\mathbf{1}\{s=15\} \;+\; \varepsilon_{i,s}^{(o)}. \tag{1}$$

The CV fixed effect $\alpha_i^{(o)}$ absorbs all time-invariant attributes of CV i within occupation o—including education, skills, job history, and any latent "quality" embedded in the synthetic profile. The omitted category is $s = 0$ (baseline age). Therefore, each coefficient has a direct interpretation: $\Delta_5^{(o)}$, $\Delta_{10}^{(o)}$, and $\Delta_{15}^{(o)}$ measure the average change in $score$ when the *same* CV is presented with an age that is 5, 10, or 15 years higher, respectively, relative to the baseline version, within occupation o. Table C.5 reports the estimates of equation (1) for the set of occupations for which the age-consistent subsample is available and yields a balanced (or near-balanced) four-scenario panel per CV. Because each candidate CV is observed under multiple age scenarios, residuals may be correlated within CV across scenarios. We therefore report heteroskedasticity-robust standard errors clustered at the CV level, allowing for arbitrary within-CV correlation in the error term and ensuring valid inference.

Table C.5: Age counterfactuals and LLM scores: fixed-effects estimates by occupation

	Fixed-effects (within-CV) models			
	(1) Chief Financial Officer	**(2) Software Developer**	**(3) Truck Driver**	**(4) Cleaning Staff**
Δ_5	-0.018	0.024	0.094***	-0.063*
	(0.025)	(0.025)	(0.033)	(0.038)
Δ_{10}	-0.036	-0.011	0.059*	-0.008
	(0.025)	(0.025)	(0.031)	(0.038)
Δ_{15}	-0.022	-0.110***	0.035	-0.113***
	(0.025)	(0.024)	(0.032)	(0.037)
Constant	6.480***	7.264***	6.376***	6.857***
	(0.017)	(0.015)	(0.019)	(0.023)
Observations	4,958	5,413	5,304	5,400
CVs (clusters)	1,350	1,356	1,326	1,350
R^2 (within)	0.001	0.008	0.002	0.003
CV FE	Yes	Yes	Yes	Yes

Notes: Each column reports a separate fixed-effects regression estimated within occupation. The dependent variable is the LLM score. Δ_5, Δ_{10}, and Δ_{15} correspond to the coefficients on indicators for $AgeShift \in \{5, 10, 15\}$, with $AgeShift = 0$ (baseline) omitted. Robust standard errors clustered at the CV level are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.