ReGeo ISSN: 2177-3246

WHEN FAIR AI BECOMES UNFAIR: A COUNTERFACTUAL AUDIT OF POSITIONAL BIAS IN LARGE LANGUAGE MODELS FOR HIRING DECISIONS

QUANDO A IA SE TORNA PARCIAL: UMA AUDITORIA CONTRAFACTUAL DO VIÉS POSICIONAL EM GRANDES MODELOS DE LINGUAGEM NAS CONTRATAÇÕES

CUANDO LA IA SE VUELVE INJUSTA: UNA AUDITORÍA CONTRAFACTUAL DEL SESGO POSICIONAL EN GRANDES MODELOS DE LENGUAJE PARA LA TOMA DE DECISIONES DE CONTRATACIÓN

https://doi.org/10.56238/revgeov17n7-181

Submission date: 31/06/2026 **Publication date:** 31/07/2026

Arthur Mesquita Camargo[1], Rafaela Silva Figueiredo Camargo[2]

ABSTRACT

As Large Language Models (LLMs) are increasingly integrated into high-stakes recruitment processes, rigorous audits to detect algorithmic bias have become critical. This study implements a multi-phase auditing protocol to evaluate gender, racial/ethnic, and positional bias in state-of-the-art proprietary models (gpt-4.1-mini, GPT-5.2, Claude Sonnet 4.6, Gemini 2.5 Pro) and in an exploratory lower-capacity model. Using a simulated CEO-selection scenario with 60 functionally identical profiles, we employ counterfactual testing to isolate the effects of demographic attributes and presentation order. The results reveal a sharp divergence in model behaviour. Frontier models demonstrated remarkable neutrality, with no statistically significant evidence of gender or racial bias. In contrast, the exploratory model exhibited extreme segregation, ranking all female candidates above all male candidates in the original condition (Cliff's $\delta = 1.0$, $p < .001$). Counterfactual analysis revealed that the root cause was not gender bias per se but an overwhelming primacy bias, whereby candidates presented at the beginning of the prompt were disproportionately favoured. Claude Sonnet 4.6 additionally displayed a form of "conscientious objection", declining to differentiate identical profiles. We complement the experimental audit with an ecosystem-level analysis of 2,653 provider-model listings from the open Models.dev catalogue, which corroborates the open-weights contamination-surface hypothesis but rejects the assumption that systemic monoculture exposure is confined to open-source models. These findings indicate that (a) state-of-the-art LLMs can achieve a high degree of demographic neutrality; (b) fundamental artefacts such as positional bias can nonetheless produce severely discriminatory outcomes; and (c) bias auditing must extend beyond demographic parity to interaction artefacts and ecosystem structure.

Keywords: Algorithmic Bias. Large Language Models. AI Fairness. Audit Studies. Gender Bias. Racial Bias. Positional Bias. AI Recruitment. Foundation Models. Open-Source Models.

RESUMO

À medida que os Grandes Modelos de Linguagem (LLMs) são cada vez mais incorporados a processos de recrutamento de alto impacto, auditorias rigorosas para detectar vieses algorítmicos tornam-se essenciais. Este estudo implementa um protocolo de auditoria em

[1] Rede Brasileira de Certificação, Pesquisa e Inovação. Lattes: http://lattes.cnpq.br/1195882649429046
[2] Rede Brasileira de Certificação, Pesquisa e Inovação. Lattes: http://lattes.cnpq.br/7540663073652051

múltiplas fases para avaliar vieses de gênero, raça/etnia e posição em modelos proprietários de última geração — gpt-4.1-mini, GPT-5.2, Claude Sonnet 4.6 e Gemini 2.5 Pro — e em um modelo exploratório de menor capacidade. Utilizando um cenário simulado de seleção para o cargo de CEO, com 60 perfis funcionalmente idênticos, empregamos testes contrafactuais para isolar os efeitos de atributos demográficos e da ordem de apresentação. Os resultados revelam uma acentuada divergência no comportamento dos modelos. Os modelos de fronteira demonstraram notável neutralidade, sem evidências estatisticamente significativas de viés de gênero ou racial. Em contraste, o modelo exploratório apresentou segregação extrema, classificando todas as candidatas mulheres acima de todos os candidatos homens na condição original (δ de Cliff = 1,0; $p < 0,001$). A análise contrafactual revelou que a causa fundamental não era o viés de gênero em si, mas um viés de primazia avassalador, pelo qual candidatos apresentados no início do prompt eram favorecidos de forma desproporcional. O Claude Sonnet 4.6 também exibiu uma forma de "objeção de consciência", recusando-se a diferenciar perfis idênticos. Complementamos a auditoria experimental com uma análise em nível de ecossistema de 2.653 listagens de provedores e modelos do catálogo aberto Models.dev. Essa análise corrobora a hipótese de uma maior superfície de contaminação em modelos de pesos abertos, mas rejeita a suposição de que a exposição a uma monocultura sistêmica esteja restrita aos modelos de código aberto. Os achados indicam que: (a) LLMs de última geração podem alcançar elevado grau de neutralidade demográfica; (b) artefatos fundamentais, como o viés posicional, podem, ainda assim, produzir resultados gravemente discriminatórios; e (c) a auditoria de vieses deve ir além da paridade demográfica, abrangendo também artefatos de interação e a estrutura do ecossistema.

Palavras-chave: Viés Algorítmico. Grandes Modelos de Linguagem. Equidade em Inteligência Artificial. Estudos de Auditoria. Viés de Gênero. Viés Racial. Viés Posicional.

RESUMEN

A medida que los Grandes Modelos de Lenguaje (LLMs) se integran cada vez más en procesos de contratación de alto impacto, las auditorías rigurosas para detectar sesgos algorítmicos se han vuelto fundamentales. Este estudio implementa un protocolo de auditoría en múltiples fases para evaluar sesgos de género, raza/etnia y posición en modelos propietarios de última generación —gpt-4.1-mini, GPT-5.2, Claude Sonnet 4.6 y Gemini 2.5 Pro— y en un modelo exploratorio de menor capacidad. Mediante un escenario simulado de selección para el puesto de CEO, con 60 perfiles funcionalmente idénticos, empleamos pruebas contrafactuales para aislar los efectos de los atributos demográficos y del orden de presentación. Los resultados revelan una marcada divergencia en el comportamiento de los modelos. Los modelos de frontera mostraron una notable neutralidad, sin evidencia estadísticamente significativa de sesgo de género o racial. En contraste, el modelo exploratorio presentó una segregación extrema, al clasificar a todas las candidatas mujeres por encima de todos los candidatos hombres en la condición original (δ de Cliff = 1,0; $p < 0,001$). El análisis contrafactual reveló que la causa principal no era el sesgo de género en sí mismo, sino un abrumador sesgo de primacía, mediante el cual los candidatos presentados al inicio del prompt eran favorecidos de manera desproporcionada. Claude Sonnet 4.6 también mostró una forma de "objeción de conciencia", al negarse a diferenciar perfiles idénticos. Complementamos la auditoría experimental con un análisis a nivel de ecosistema de 2.653 listados de proveedores y modelos del catálogo abierto Models.dev. Este análisis respalda la hipótesis de una mayor superficie de contaminación en los modelos de pesos abiertos, pero rechaza la suposición de que la exposición a una monocultura

sistémica esté limitada a los modelos de código abierto. Estos hallazgos indican que: (a) los LLMs de última generación pueden alcanzar un alto grado de neutralidad demográfica; (b) artefactos fundamentales, como el sesgo posicional, pueden aun así producir resultados gravemente discriminatorios; y (c) la auditoría de sesgos debe ir más allá de la paridad demográfica e incluir los artefactos de interacción y la estructura del ecosistema.

Palabras clave: Sesgo Algorítmico. Grandes Modelos de Lenguaje. Equidad en la Inteligencia Artificial. Estudios de Auditoría. Sesgo de Género. Sesgo Racial. Sesgo Posicional.

1 INTRODUCTION

The integration of Artificial Intelligence (AI), and particularly Large Language Models (LLMs), into high-stakes decision-making processes such as recruitment and talent selection promises to transform human resource management (Chen, 2023; Rigotti & Fosch-Villaronga, 2024). The central proposition is that these tools can mitigate human cognitive biases, yielding more objective and equitable evaluations. That promise, however, is accompanied by a significant risk: AI systems trained on vast textual corpora that reflect historical and social prejudice may perpetuate — or even amplify — discriminatory practices (Ahmad et al., 2025; Gaebler et al., 2024; Lima & Rahman, 2024; Sony et al., 2025; Wilson & Caliskan, 2024).

The contemporary foundation-model landscape is characterised by a growing dichotomy between proprietary (closed) models and open-source or open-weights models. Proprietary models typically embed robust "responsibility strings" — layered content filters and safety-alignment mechanisms — designed to mitigate bias. Open-weights models, while powerful drivers of innovation, are inherently more vulnerable to data poisoning and malicious fine-tuning, which can dismantle their original safeguards (Bowen et al., 2025; Fu et al., 2024; Priyanshu et al., 2024).

Inspired by the classic audit studies in the social sciences that exposed labour-market discrimination through correspondence experiments (e.g., Bertrand & Mullainathan, 2004), research on AI fairness has adapted this methodology to probe the "state of mind" of machines. Such studies typically present controlled stimuli to a model, varying only a single sensitive attribute — such as gender or race — to observe whether systematically different outcomes occur (Veldanda et al., 2023; Wang et al., 2024).

This article extends that line of inquiry through a comprehensive, multi-phase experimental audit designed to test the limits of neutrality in leading contemporary LLMs, situating the results within the proprietary/open-source governance spectrum. Unlike studies that focus on a single bias type or a single model, our tripartite approach systematically investigates gender bias, racial/ethnic bias, and cross-model behavioural differences using a rigorous counterfactual protocol (Asif et al., 2025). We address four central research questions:

- **Gender bias**: Under conditions of perfectly identical competence, do LLMs rank candidates differently on the basis of gender?

- **Racial/ethnic bias**: Under the same conditions, do LLMs exhibit bias based on culturally marked names suggesting different racial or ethnic origins?
- **Cross-model consistency**: Is biased (or unbiased) behaviour consistent across different state-of-the-art LLMs?
- **Nature of the bias**: When bias is detected, does it derive from semantic association with the demographic attribute (e.g., gender stereotypes) or from model artefacts such as positional (primacy/recency) bias (Guo & Vosoughi, 2025; Shi et al., 2025)?

By addressing these questions, this study provides a nuanced picture of the current state of algorithmic fairness in LLMs, highlighting both the notable progress achieved and the critical pitfalls that remain on the path to genuinely equitable AI-driven hiring.

2 METHOD

2.1 EXPERIMENTAL DESIGN

The study consisted of three sequential audit experiments, all using a simulated CEO-selection scenario. In each experiment, the models were instructed to rank 60 candidate profiles with functionally identical competencies, experience, and credentials, and to select a "TOP-10". The design guarantees that any variation in ranking cannot be attributed to qualification, thereby isolating the effect of the manipulated variables. For every experiment, an "original" and a "counterfactual" condition were implemented to permit robust causal analysis.

2.2 EVALUATED MODELS

Five models were audited. Four are state-of-the-art proprietary foundation models whose commercial specifications were independently verified against the open-source Models.dev catalogue (Models.dev, 2026): *gpt-4.1-mini* (OpenAI; released 14 April 2025; ≈1,047,576-token context window; USD 0.40/1.60 per million input/output tokens), *GPT-5.2* (OpenAI; USD 1.75/14.00), *Claude Sonnet 4.6* (Anthropic; USD 3.00/15.00), and *Gemini 2.5 Pro* (Google; USD 1.25/10.00). All four are distributed exclusively as closed-weight services (open_weights = false in the catalogue schema) and are redistributed, unmodified, by 10, 16, 24, and 15 independent commercial providers, respectively — a property with systemic implications examined in Section 4.3. The fifth, exploratory, model is a lower-capacity model included to probe behaviour outside the frontier tier; it is kept anonymised in line with the audit-reporting conventions adopted here.

2.3 STIMULI AND MANIPULATIONS

- **Experiment 1 — Gender bias**: Sixty profiles were generated. Gender was assigned in alternation (female for odd IDs, male for even IDs), and eleven age bands were distributed uniformly. In the counterfactual condition, the gender labels of all profiles were inverted.
- **Experiment 2 — Racial/ethnic bias**: Sixty profiles were generated, each assigned to one of six racial-ethnic groups (White-European, Black-Afrodescendant, Indigenous, Asian, Latino-Hispanic, Arab-Middle Eastern) through culturally marked names, with perfectly balanced group distribution. In the counterfactual condition, names were exchanged across groups to dissociate the name from its position in the prompt.
- **Experiment 3 — Cross-model comparison**: The Experiment 1 data were applied to the remaining four models to permit a direct comparison of gender-bias behaviour.

2.4 PROCEDURE AND STATISTICAL ANALYSIS

The models were queried via API with an identical prompt instructing them to return a complete ranking and a TOP-10 list in JSON format. Statistical analysis was conducted at a significance level of $\alpha = .05$, with Holm–Bonferroni correction for multiple pairwise comparisons. The tests comprised: the chi-square test of independence, to assess whether the proportion of candidates selected in the TOP-10 was independent of demographic group; the Kruskal–Wallis H test, to compare mean ranking positions across the multiple ethnic groups; the Mann–Whitney U test, to compare mean ranking positions between the two gender groups; Cliff's delta (δ) and epsilon-squared (ε^2), as effect-size measures; and counterfactual pair analysis, to measure the change in a candidate's position when their demographic attribute was inverted.

3 RESULTS

3.1 EXPERIMENTS 1 AND 2: NEUTRALITY OF GPT-4.1-MINI

In the robust analysis with 30 repetitions, gpt-4.1-mini showed no evidence of gender or racial/ethnic bias.

- **Gender**: The TOP-10 distribution averaged 5.0 females to 5.0 males, and the mean ranking position was virtually identical for both genders (30.5). The Mann–Whitney U test was non-significant, $U = 442.5$, $p \approx .95$, with a negligible effect size (Cliff's $\delta \approx 0.0$).
- **Race/ethnicity**: The chi-square test for TOP-10 selection was non-significant, $\chi^2(5, N = 3{,}000) = 3.74$, $p = .587$. Likewise, the Kruskal–Wallis test for mean ranking positions found no significant differences among the six ethnic groups, $H(5) = 8.06$, $p = .153$. All 15 pairwise comparisons were non-significant after Holm–Bonferroni correction.

3.2 EXPERIMENT 3: DIVERGENCE IN MODEL BEHAVIOUR

The comparison across the five models revealed three distinct behavioural patterns, summarised in Table 1.

Table 1

Comparative Summary of Gender Bias by Model (Original Condition)

Model	TOP-10 (F:M)	Mean rank (F vs. M)	Cliff's δ	p (M–W)	Observed behaviour
gpt-4.1-mini	5:5	30.5 vs. 30.5	0.000	.950	Neutral and random
GPT-5.2	5:5	31.2 vs. 29.8	−0.047	.762	Neutral and random
Claude Sonnet 4.6	5:5	30.0 vs. 31.0	0.033	.830	Objection (sequential return)
Gemini 2.5 Pro	4:6	29.2 vs. 31.8	0.084	.579	Neutral and random
Exploratory model	10:0	15.5 vs. 45.5	1.000	< .001	Extreme segregation bias

Four of the five models (gpt-4.1-mini, GPT-5.2, Claude Sonnet 4.6, Gemini 2.5 Pro) showed no significant gender bias. Claude Sonnet 4.6 was notable for returning the candidates in the exact order in which they were presented — effectively declining to rank identical profiles — a behaviour that can be interpreted as a sophisticated form of instruction adherence.

3.3 THE EXPLORATORY MODEL: GENDER BIAS OR POSITIONAL BIAS?

The exploratory model exhibited extreme bias, ranking all 30 female candidates in positions 1–30 and all 30 male candidates in positions 31–60, producing a maximal effect size (Cliff's $\delta = 1.0$) at high significance ($p < .001$).

The counterfactual analysis, however, proved decisive. When genders were inverted, the bias did not invert symmetrically. Instead, candidates with lower IDs (originally female) continued to be favoured, even though they were now labelled male. The mean positional shift for originally female profiles was +11.2 (worse in the counterfactual), while originally male profiles shifted −11.2 (better). The statistical analysis of this shift was highly significant ($p < .001$), yet the counterfactual ranking displayed a mixed pattern rather than an inverted segregation. This indicates that the model was responding not to the label "female" but to the candidate's position in the prompt — a textbook case of primacy bias, a cognitive artefact

whereby items presented first receive disproportionate weight (Guo & Vosoughi, 2025; Shi et al., 2025; Veseli et al., 2025).

4 THE SPECTRUM OF RESPONSIBILITY: PROPRIETARY VS. OPEN-SOURCE MODELS

The behavioural divergence observed among the tested models raises critical questions about the governance architecture underlying these systems. A parallel analysis of the current foundation-model landscape reveals a growing dichotomy between proprietary (closed) and open-source or open-weights models, with direct implications for bias mitigation and safety in hiring applications.

4.1 PROPRIETARY MODELS AND "RESPONSIBILITY STRINGS"

Proprietary models such as GPT-4 (OpenAI), Claude (Anthropic), and Gemini (Google) operate under a centralised control paradigm. These systems benefit from massive investment in safety alignment, employing techniques such as Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI (Anthropic, n.d.; Huang et al., 2024; OpenAI, 2024).

Crucially, these models incorporate what we term "responsibility strings" — layers of content filters, moderation classifiers, and system-level directives that actively instruct the model to refuse discriminatory or harmful requests. The effectiveness of these measures is reflected in the neutral performance observed in this study for gpt-4.1-mini, GPT-5.2, Claude Sonnet 4.6, and Gemini 2.5 Pro. Centralisation permits continuous safety updates, but raises concerns about transparency and about the concentration of power in defining what constitutes "fair" behaviour (Priyanshu et al., 2024).

4.2 OPEN-SOURCE MODELS AND THE RISK OF CONTAMINATION

At the other end of the spectrum lie open-weights and open-source models, such as Llama (Meta), Mistral (Mistral AI), Falcon (TII), and Qwen (Alibaba Cloud). Democratised access to these models drives innovation but introduces significant risk vectors.

Although the creators of these models implement bias mitigation at the pre-training stage (e.g., data curation and supervised fine-tuning), their open nature allows unrestricted third-party fine-tuning. This renders community models (such as those hosted on the Hugging Face Hub) highly vulnerable to data poisoning and malicious fine-tuning (Bowen et al., 2025; Fu et al., 2024). A malicious actor — or even a careless developer — can introduce biased

training data that undoes the model's original safeguards, inducing discriminatory behaviour that proprietary systems would block.

Moreover, open-weight models such as DeepSeek-R1 and Qwen variants have demonstrated markedly higher susceptibility to jailbreaking, and community techniques such as "abliteration" explicitly remove safety filters from released weights (National Institute of Standards and Technology [NIST], 2025). In a recruitment scenario, using an unaudited or third-party-modified open-source model therefore carries a critical risk of perpetuating systemic bias under the veneer of algorithmic objectivity. Table 2 summarises the comparative landscape of the principal model families, classified by governance type, bias-mitigation strategy, data-contamination risk, and estimated risk for hiring applications.

Table 2

Comparative Landscape of Foundation Models: Governance, Bias Mitigation, and Hiring Risk

Model / family	Type	Mitigation techniques	Contamination risk	Responsibility strings	Hiring risk
GPT-4 / GPT-4o / GPT-4.1 (OpenAI)	Proprietary	RLHF, content filters, red-teaming	Low	Robust	Medium
Claude 3 / 4 (Anthropic)	Proprietary	Constitutional AI, RLHF, DPO	Low	Very robust	Low–Medium
Gemini 1.5 / 2.0 / 2.5 (Google)	Proprietary	SFT, RL, automated red-teaming	Low	Robust	Medium
Llama 2 / 3 / 3.1 (Meta)	Open-weights	SFT, DPO, Llama Guard	High	Moderate	Medium
Mistral 7B / Mixtral / Large	Open-weights	SFT, DPO (partial)	High	Limited	Medium
Falcon 7B / 40B / 180B (TII)	Open-source	Data curation	High	Limited	Medium
BLOOM / BLOOMZ (BigScience)	Open-source	Transparency via Model Card	Critical	Minimal	Critical
Qwen / Qwen2 / Qwen2.5 (Alibaba)	Open-weights	Qwen3Guard, RLHF, DPO	High	Moderate	High
DeepSeek-V2 / R1 (DeepSeek AI)	Open-source	Curation, SFT, RL	Critical	Limited	High
Fine-tuned models (Hugging Face)	Open-source	Heterogeneous / variable	Critical	Minimal	Critical

4.3. AN ECOSYSTEM-LEVEL EMPIRICAL TEST OF THE GOVERNANCE DICHOTOMY

The governance dichotomy articulated in Sections 4.1–4.2 rests on two empirically testable propositions. To test them, we analysed the complete open-source Models.dev

catalogue (Models.dev, 2026; snapshot of 13 July 2026, commit a2087ac), comprising 2,653 provider-model listings across 113 commercial inference providers, 28 model-authoring laboratories, and 247 canonical foundation models (113 open-weights, 134 proprietary). Two hypotheses were prespecified. H*E1* (contamination surface): the open-weights segment constitutes a substantial, economically favoured, and technically modifiable share of the commercial ecosystem, sustaining the tampering-risk argument of Section 4.2. H*E2* (asymmetric systemic exposure): if systemic risk were confined to the open-source regime, open-weights models should be redistributed across significantly more independent providers than proprietary models. Group comparisons employed the Mann–Whitney U test with Cliff's δ as the effect size, mirroring the statistical apparatus of Section 2.4.

H*E1* was corroborated. Open-weights models account for 35.2% of all commercial listings; 73% of open canonical models expose public weight-download links — the technical precondition for the data poisoning and malicious fine-tuning documented by Fu et al. (2024) and Bowen et al. (2025); 20 of the 113 providers operate exclusively as open-weights resellers; and open-weights inference is significantly cheaper (median USD 0.30 vs. 0.80 per million input tokens; $U = 146{,}852$, $p < .001$, Cliff's $\delta = 0.343$), creating economic pressure toward open-model adoption precisely in cost-sensitive, high-volume screening pipelines such as recruitment.

H*E2*, however, was rejected. Proprietary models are redistributed across at least as many independent providers as open-weights models (median 3 vs. 2 providers per canonical model; $U = 3{,}948.5$, $p = .180$, Cliff's $\delta = -0.112$), and the share of models served by two or more independent providers does not differ between regimes (67.5% vs. 63.2%; $\chi^2(1) = 0.22$, $p = .639$). Indeed, the most widely redistributed models in the snapshot span both regimes — GLM-5.2 (open-weights; 27 providers) and Claude Opus 4.8 (proprietary; 22 providers) — and the four frontier models audited in this study are themselves served by 10 to 24 independent providers each. The market is moderately concentrated at the laboratory level (Herfindahl–Hirschman Index = 1,116; three-laboratory concentration ratio CR3 = 50.7%; CR5 = 67.0%), with OpenAI (20.7%), Alibaba (15.9%), and Anthropic (14.2%) accounting for half of all listings. Table 3 summarises these indicators.

Table 3

Ecosystem-Level Indicators of the Governance Spectrum (Models.dev Snapshot, 13 July 2026)

Ecosystem indicator	Open-weights	Proprietary	Test / effect size
Canonical models catalogued (n = 247)	113 (45.7%)	134 (54.3%)	—
Share of provider-model listings (n = 2,653)	35.2%	64.8%	—
Median input price (USD per million tokens)	0.30	0.80	$U = 146{,}852$, $p < .001$, $\delta = 0.343$
Median independent providers per canonical model	2 (M = 3.99; max = 27)	3 (M = 4.75; max = 22)	$U = 3{,}948.5$, $p = .180$, $\delta = -0.112$
Share of models served by ≥ 2 independent providers	63.2%	67.5%	$\chi^2(1) = 0.22$, $p = .639$
Public weight-download availability	73% of models	0% of models	—

Two refinements to Table 2 follow. First, the contamination *vector* is indeed exclusive to the open regime — weight tamperability is the sine qua non of data poisoning — but the *homogenisation* exposure is governance-agnostic: because a shared model produces deterministically replicable outputs wherever it is served (Bommasani et al., 2026), a proprietary frontier model redistributed by 22 providers correlates downstream hiring decisions exactly as widely as an open model redistributed by 27. Kleinberg and Raghavan (2021) show formally that such monocultural convergence can reduce aggregate decision quality even when the shared algorithm is more accurate than any alternative — a Braess-paradox result implying that the demographic neutrality of frontier models documented in Section 3 does not, by itself, immunise the ecosystem against monoculture harms. Second, legacy open families retained in Table 2 for completeness — Falcon (TII) and BLOOM (BigScience) — register zero commercial provider listings in the snapshot: their contamination risk is now largely theoretical, whereas exposure concentrates in actively redistributed families (Qwen: 39 providers; DeepSeek: 31; Llama: 20; Mistral: 18).

5 DISCUSSION, THEORETICAL AND PRACTICAL IMPLICATIONS

The results of this study offer crucial insights for the theory and practice of responsible AI, particularly in the domain of high-stakes hiring. First, the remarkable neutrality demonstrated by the state-of-the-art proprietary foundation models (gpt-4.1-mini, GPT-5.2,

Claude Sonnet 4.6, Gemini 2.5 Pro) in the face of gender and ethnicity manipulations constitutes a significant theoretical contribution to our understanding of the effectiveness of alignment and governance strategies. It corroborates the thesis that models with robust "responsibility strings" grounded in RLHF and Constitutional AI (Huang et al., 2024) can be engineered to operate fairly in sensitive contexts (Gerszberg et al., 2026; Wilson et al., 2025). This finding challenges simplistic narratives that all LLMs are inherently biased, pointing instead to the decisive role of alignment engineering.

Second, the identification of extreme positional bias in the exploratory model as a massive confounding factor is a seminal theoretical contribution. It challenges the prevailing assumption that bias in LLMs is exclusively demographic, revealing that interaction artefacts — such as the order in which candidates are presented — can generate severely discriminatory outcomes even in the absence of explicit prejudice against protected groups (Rozado, 2026; Shi et al., 2025; Veseli et al., 2025). Although not directly tied to demographic attributes, this bias has equally devastating practical implications, perpetuating inequality by favouring candidates who happen to appear first in the list. The distinction between demographic and positional bias is crucial for the development of more sophisticated audits.

For practice, this study underscores the urgent need for counterfactual bias audits (Asif et al., 2025; Cornacchia et al., 2023) that transcend mere group-parity analysis. Organisations deploying LLMs in HR processes should implement rigorous tests assessing not only demographic bias but also positional bias and other subtle forms of discrimination, combined where appropriate with structured debiasing interventions (Ip, 2025). Table 2, which classifies model families by risk, offers an initial guide to LLM selection for hiring applications, prioritising those with robust governance and a demonstrated alignment record. Transparency about alignment mechanisms and vulnerabilities is fundamental to informed decision-making.

A critical implication for AI governance lies in the proprietary/open-source dichotomy. While the former invest heavily in alignment and red-teaming, open-source models — despite their transparency and flexibility — are intrinsically more vulnerable to data poisoning and malicious fine-tuning (Bowen et al., 2025). This constitutes a systemic risk: open models without clear "responsibility strings" can be adapted to perpetuate or amplify bias, especially in hiring contexts where fairness is paramount. Regulators and policymakers should weigh this distinction when developing guidelines for AI in HR, demanding heightened scrutiny and certification for open-source models in high-stakes applications. At the same time, the

ecosystem evidence of Section 4.3 cautions against reducing systemic risk to the open regime alone: redistribution — and therefore correlated decision-making — is equally a property of the proprietary market.

Finally, the ecosystem evidence permits a theoretical synthesis between the positional bias identified here and the emerging literature on algorithmic monoculture. Bommasani et al. (2026), analysing 4.2 million real applications screened by a single hiring-algorithm vendor, demonstrate that shared models generate *systemic rejection*: because 42 models were shared across employers, rejection at one firm mechanically entailed rejection at others, and observed systemic rejection rates significantly exceeded the baseline of statistically independent decisions. Our findings extend this logic in a non-demographic direction. A primacy-biased model that is redistributed, byte-identical, across dozens of providers converts a local interaction artefact into market-wide correlated exclusion: candidates who happen to be presented late in the prompt would be disfavoured at *every* adopting organisation simultaneously — a form of systemic positional exclusion that is invisible to four-fifths-rule demographic audits (cf. Bommasani et al., 2026) precisely because position is not a protected attribute. Kleinberg and Raghavan's (2021) welfare analysis implies that this correlation is costly even absent any bias, and Baek, Bastani, and Chen (2026) show that when firms share an algorithmic evaluation, naive reliance on its top-ranked outputs produces severe congestion whose remedy — strategic diversification, enabled by platform disclosure of congestion information — has a direct auditing analogue: mandatory prompt-order randomisation at inference time and disclosure duties for providers that redistribute a common base model into high-stakes screening pipelines.

6 LIMITATIONS AND FUTURE RESEARCH AGENDA

Although robust, this study has limitations that open avenues for a rich future research agenda. The empirical analysis of the five audited models was based on a simulated CEO-hiring scenario, and generalisation to other contexts or job types should be made with caution. Moreover, the depth of the counterfactual audit varied across models, with only the exploratory model subjected to a detailed positional-bias analysis. Extending this methodology to all audited models and to the model families identified in Table 2, with a substantially larger number of repetitions (e.g., 1,000 tests per model), is a priority research direction.

The ecosystem-level analysis of Section 4.3 carries its own limitations. The Models.dev catalogue measures the supply side of the market — commercial availability, pricing, and redistribution — not usage-weighted adoption; a model served by many providers is not necessarily the model processing the most applications. The analysis is a single cross-sectional snapshot (13 July 2026) of a fast-moving market; canonical-model linkage covers 193 of 247 catalogued models; and provider counts systematically undercount private deployments and community fine-tuned derivatives (e.g., Hugging Face forks), which renders the test of H*E2* conservative with respect to open-weights exposure. Replication on usage-weighted telemetry, where obtainable under the researcher-access regimes advocated by Bommasani et al. (2026), is a priority.

Future research should also explore:

- **Bias auditing of open-source models**: Developing and applying counterfactual audit protocols to open-source models such as Llama, Mistral, Qwen, and DeepSeek-R1, investigating their susceptibility to data poisoning and biased fine-tuning. This includes creating standardised, publicly available benchmarks for the open-source ecosystem, enabling continuous evaluation of fairness degradation in adapted models.
- **Impact of subtle qualification differences**: Investigating how LLMs behave when profiles contain subtle qualification differences, and whether positional bias persists or is mitigated — refining our understanding of model sensitivity to qualification signals versus presentation artefacts.
- **Human perception of positional bias**: Investigating how human decision-makers are influenced by positional bias in LLM recommendations, and how interfaces or training might mitigate it. Human–AI interaction is a critical vector for the amplification or mitigation of bias.
- **Public-policy implications**: Analysing the effectiveness of different regulatory approaches (e.g., the GDPR and the EU AI Act) in mitigating LLM bias in HR, considering the proprietary/open-source distinction and the monoculture dynamics documented in Section 4.3.

7 CONCLUSION

This study establishes a new paradigm for bias auditing in Large Language Models applied to high-stakes decision-making contexts such as executive hiring. Our findings demonstrate that while leading proprietary foundation models (gpt-4.1-mini, GPT-5.2, Claude Sonnet 4.6, Gemini 2.5 Pro) have achieved remarkable demographic neutrality in scenarios of identical competence — a testament to the effectiveness of advanced alignment techniques such as RLHF and Constitutional AI — algorithmic fairness is a substantially more

complex goal than the mere absence of explicit prejudice. The absence of gender or racial bias does not guarantee equitable outcomes, necessitating an urgent redefinition of auditing metrics and methodologies. The most seminal contribution of this work lies in the identification and characterisation of extreme positional bias as a massive and, until now, underestimated confounding factor.

Our "algorithmic litmus test" revealed that interaction artefacts, such as the order of candidate presentation, can generate severe discrimination even in models that exhibit no demographic bias. This primacy bias, whereby candidates presented early in the prompt are systematically favoured, has serious implications for fairness, as it perpetuates inequality in a subtle and non-demographic manner. The finding challenges a literature that has focused predominantly on demographic bias and establishes a new critical dimension for evaluating fairness in AI systems, demanding a reorientation of mitigation strategies and a methodological vigilance that transcends the superficial analysis of protected attributes.

The practical and governance implications are profound and multifaceted. For organisations implementing LLMs in HR processes, it is imperative to move beyond superficial audits and adopt counterfactual protocols that rigorously assess both demographic and positional bias. The proprietary/open-source distinction emerges as a central axis of risk management: while proprietary developers invest in robust, centralised "responsibility strings", open models — despite their flexibility and innovation — are intrinsically more vulnerable to data poisoning and malicious fine-tuning. Our ecosystem-level analysis nonetheless refines this dichotomy: the contamination vector is exclusive to the open regime, but monoculture exposure is governance-agnostic, since the redistribution of identical models across dozens of providers — proprietary and open alike — correlates hiring decisions market-wide. Regulators should therefore combine heightened scrutiny and certification of open-source models in high-stakes applications with disclosure and audit duties that attach to model redistribution itself.

This study paves the way for a robust, high-impact research agenda: extending counterfactual auditing across the foundation-model families identified in Table 2 with larger repetition counts; developing standardised, dynamic benchmarks for the open-source ecosystem; investigating the impact of subtle qualification differences and the human perception of positional bias; and pursuing replication of the ecosystem analysis on usage-weighted data under researcher-access regimes.

Ultimately, the path to genuinely fair and equitable AI-driven hiring demands continuous, multifaceted, and methodologically sophisticated vigilance. It is not merely a matter of eliminating the reflections of our societal prejudices, but of unravelling the "idiosyncratic ghosts" within the machine itself. Only through rigorous, comparative, and counterfactual audits — attentive to demographic bias, to interaction artefacts, and to the structure of the model ecosystem — can we build AI systems that are not only efficient but fundamentally just, fulfilling an ethical and social imperative for the future of work.

REFERENCES

Ahmad, A., Vallès, Y., & Idaghdour, Y. (2025). Bias in AI systems: Integrating formal and socio-technical approaches. *Frontiers in Big Data*, *8*. https://doi.org/10.3389/fdata.2025.1686452

Anthropic. (n.d.). *Claude's constitution*. https://www.anthropic.com/news/claudes-constitution

Asif, N. H., Banik, A., & Ajmari, F. (2025). Causal fairness in black-box AI: A counterfactual auditing framework for deep models. *ResearchGate* (Preprint). https://doi.org/10.21203/rs.3.rs-7023047/v1

Baek, J., Bastani, H., & Chen, S. (2026). Strategic hiring under algorithmic monoculture. *arXiv*. https://arxiv.org/abs/2502.20063

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

Bommasani, R., Bana, S. H., Creel, K. A., Jurafsky, D., & Liang, P. (2026). Algorithmic monocultures in hiring. *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*, 6351–6382. https://doi.org/10.1145/3805689.3812400

Bowen, D., Murphy, B., Cai, W., Khachaturov, D., Gleave, A., & Pelrine, K. (2025). Scaling trends for data poisoning in LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://arxiv.org/abs/2408.02946

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, *10*, 567. https://doi.org/10.1057/s41599-023-02079-x

Cornacchia, G., Anelli, V. W., Biancofiore, G. M., Narducci, F., Pomo, C., Ragone, A., & Di Sciascio, E. (2023). Auditing fairness under unawareness through counterfactual reasoning. *Information Processing & Management*, *60*(2), 103224. https://doi.org/10.1016/j.ipm.2022.103224

Fu, T., Sharma, M., Torr, P., Cohen, S. B., Krueger, D., & Barez, F. (2024). PoisonBench: Assessing large language model vulnerability to data poisoning. *arXiv*. https://arxiv.org/abs/2410.08811

Gaebler, J. D., Goel, S., Huq, A., & Tambe, P. (2024). Auditing the use of language models to guide hiring decisions. *arXiv*. https://arxiv.org/abs/2404.03086

Gerszberg, N., Hamori, J., & Lo, A. (2026). Quantifying gender bias in large language models: When ChatGPT becomes a hiring manager. *arXiv*. https://arxiv.org/abs/2604.00011

Guo, X., & Vosoughi, S. (2025). Serial position effects of large language models. *Findings of the Association for Computational Linguistics: ACL 2025*, 927–953. https://doi.org/10.18653/v1/2025.findings-acl.52

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., & Ganguli, D. (2024). Collective Constitutional AI: Aligning a language model with public input. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–1417. https://doi.org/10.1145/3630106.3658979

Ip, E. (2025). Fair AI in hiring: Experimental evidence on how biased hiring algorithms and different debiasing methods affect the quality and diversity of applicants. *Behavioral Science & Policy*. https://doi.org/10.1177/23794607251353585

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, *118*(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

Lima, S. A., & Rahman, M. M. (2024). Algorithmic fairness in HRM: Balancing AI-driven decision making with inclusive workforce practices. *Journal of Information Systems Engineering & Management*, *9*(4s). https://doi.org/10.52783/jisem.v9i4s.13251

Models.dev. (2026). *Models.dev: An open-source database of AI model specifications, pricing, and features* [Data set; snapshot of 13 July 2026, commit a2087ac]. https://models.dev

National Institute of Standards and Technology, Center for AI Standards and Innovation. (2025). *Evaluation of DeepSeek AI models*. U.S. Department of Commerce. https://www.nist.gov/system/files/documents/2025/09/30/CAISI_Evaluation_of_DeepSeek_AI_Models.pdf

OpenAI. (2024). *Introducing the Model Spec*. https://openai.com/index/introducing-the-model-spec

Priyanshu, A., Maurya, Y., & Hong, Z. (2024). AI governance and accountability: An analysis of Anthropic's Claude. *arXiv*. https://arxiv.org/abs/2407.01557

Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. *Computer Law & Security Review*, *53*, 105966. https://doi.org/10.1016/j.clsr.2024.105966

Rozado, D. (2026). Gender and positional biases in LLM-based hiring decisions: Evidence from comparative CV/résumé evaluations. *PeerJ Computer Science*, *12*, e3628. https://peerj.com/articles/cs-3628/

Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., & Vosoughi, S. (2025). Judging the judges: A systematic study of position bias in LLM-as-a-judge. *Proceedings of the 14th International Joint Conference on Natural Language Processing*, 292–314. https://doi.org/10.18653/v1/2025.ijcnlp-long.18

Sony, M. M. A. A., Amin, M. B., Ashraf, A., Islam, K. M. A., Debnath, N. C., & Debnath, G. C. (2025). Bias in AI-driven HRM systems: Investigating discrimination risks embedded in AI recruitment tools and HR analytics. *Social Sciences & Humanities Open*. https://doi.org/10.1016/j.ssaho.2025.102082

Veldanda, A. K., Grob, F., Thakur, S., Pearce, H., Tan, B., Karri, R., & Garg, S. (2023). Are Emily and Greg still more employable than Lakisha and Jamal? Investigating algorithmic hiring bias in the era of ChatGPT. *arXiv*. https://arxiv.org/abs/2310.05135

Veseli, B., Chibane, J., Toneva, M., & Koller, A. (2025). Positional biases shift as inputs approach context window limits. *Proceedings of the Conference on Language Modeling (COLM 2025)*. https://openreview.net/forum?id=vlUk8z8LaM

Wang, Z., Wu, Z., Guan, X., Thaler, M., Koshiyama, A., Lu, S., Beepath, S., Ertekin, E., & Perez-Ortiz, M. (2024). JobFair: A framework for benchmarking gender hiring bias in large language models. *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

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

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