Gender bias and propagation of stereotypes in GenAI-assisted recruitment

Gender bias and propagation of stereotypes in GenAI-assisted recruitment

Martina Ullasci
Politecnico di Torino, Italy
Turin, Italy
martina.ullasci@polito.it

Marco Rondina
Politecnico di Torino, Italy
Turin, Italy
marco.rondina@polito.it

Riccardo Coppola
Politecnico di Torino, Italy
Turin, Italy
riccardo.coppola@polito.it

Antonio Vetrò
Politecnico di Torino, Italy
Turin, Italy
antonio.vetro@polito.it

Abstract

In recent years, generative artificial intelligence (GenAI) systems have assumed increasingly crucial roles in personnel recruitment and candidate profiles analysis. However, using large language models introduces the risk of perpetuating and exacerbating existing gender stereotypes in the labour market. This research aims to evaluate this phenomenon, analysing how a state-of-the-art generative model (GPT-5) suggests occupations and represents *ideal candidates* based on their gender, focusing on under 35 years old Italian graduates. The study consists of two complementary experiments. In the *Candidate-driven experiment*, the model is prompted to provide job suggestions for 24 synthetic candidate profiles, balanced by gender, age, experience, and professional field. Results show that, although no significant differences emerged in job titles, gendered linguistic patterns exist in the adjectives attributed to female and male candidates, indicating a tendency of the model to associate women with emotional and empathetic traits, while men with strategic and analytical ones. The *Job-driven experiment* employed 114 LinkedIn job advertisements as prompts to generate textual and visual representations of *ideal candidates*. The analysis of the outputs revealed a clear gender polarisation: the model assigned 71% of profiles to male and 29% to female gender. The strongest association emerged in *HR & People Operations* occupations, assigned exclusively to female candidates, and *Operations, Technical & Manufacturing* jobs, assigned exclusively to male candidates. Visual analysis confirms the perpetuation of gender stereotypes, depicting women in more approachable postures and men in assertive roles. These results suggest that, in the recruitment domain and under the experimental settings of this study, GenAI models do not simply reflect the gender biases of the training data, but also amplify them. The research raises an ethical question regarding the use of these models in HR decision support, highlighting the need for transparency and bias mitigation strategies to ensure fairness and inclusive representation.

CCS Concepts

• **Social and professional topics** → **User characteristics**; **Gender**;

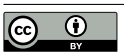
Keywords

Generative AI, AI Fairness, AI Ethics, Large Language Models

ACM Reference Format:
Martina Ullasci, Marco Rondina, Riccardo Coppola, and Antonio Vetrò. 2026. Gender bias and propagation of stereotypes in GenAI-assisted recruitment. In *34th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE Companion '26), July 05–09, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3803437.3805541

1 Introduction

Generative AI (GenAI) has emerged as one of the most transformative technologies of our time, rapidly redefining social dynamics, economic structures and everyday life. In particular, the integration of AI in the Human Resources (HR) sector has made GenAI tools increasingly used in candidate recruitment, evaluation and selection processes, promising time-saving capabilities [1], greater efficiency and cost reductions [2], [3]. Although AI is considered a neutral and objective technology, a critical analysis reveals this assumption to be fallacious [4]. AI systems are inherently biased because they are trained using data that inevitably reflect the social inequalities, stereotypes and historical discriminations present in our society [5]. Moreover, as designed artefacts, they also embed the values, assumptions and design choices of their developers and deployers [6]. As a consequence, AI has the potential to replicate and even amplify gender bias, thereby exacerbating occupational segregation[1] and wage disparities [8]. As Kate Crawford argues *"AI systems are not neutral artefacts, but material infrastructures embedded in history and power"* [9]. Understanding these mechanisms is necessary beyond academic requirements: it represents a crucial ethical challenge to make sure that technological progress leads to social justice rather than reinforcing existing barriers. In addition, anti-discrimination laws prohibit gender bias in employment throughout



FSE Companion '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2636-1/2026/07
https://doi.org/10.1145/3803437.3805541

[1] Gender segregation is defined as the division of occupations based on gender, where women predominantly occupy roles in sectors like caring, cashiering, catering, clerical, and cleaning, while men are more likely to work in fields such as engineering, construction, or computing [7].

the entire employment lifecycle (i.e. from job advertising and recruitment to pay, promotion and dismissal). Existing laws prohibit both direct discrimination, which is the explicit treatment of people differently because of their gender, and indirect discrimination, which is the neutral treatment of people that disproportionately disadvantages one gender without objective justification. The use of GenAI in HR can introduce both direct and indirect discrimination. While traditional predictive AI systems are limited to analysing existing data to predict outcomes, generative artificial intelligence has a distinctive and transformative capability: the synthesis of new information, textual content and visual representations. This characteristic introduces greater complexity to the issue of algorithmic bias. In the context of recruitment processes and selection in the labour market, prejudice is not only about how candidates are sorted or filtered; it also extends to the creation of content that defines who is considered *ideal* for a given role. Kate Crawford affirms that GenAI synthesises information from massive datasets that often contain stereotypes, enabling the technology to produce distorted content that appears credible and natural [9]. When these systems are employed in job recruitment platforms and in selection processes, a risk arises. For instance, if generative AI portrays an *ideal candidate* for a CEO position as a white middle-aged man in a suit and represents a nurse as a young, smiling and caring woman, the system not only reflects pre-existing stereotypes but actively reproduces them [10]. Both textual and visual outputs function as mechanisms of discouragement, distancing women, non-binary people, and minorities from leadership, technical, or traditionally "male" roles, thereby increasing the gender gap in the labour market [10], [11]. This study aims to examine how gender bias manifests in the selection processes of young graduates in Italy using generative artificial intelligence tools. The analysis is articulated around two complementary experiments: the **Candidate-driven experiment**, in which the model identifies the *ideal job* for given profiles of job-seekers, and the **Job-driven experiment**, in which the model defines the *ideal candidate* for real job advertisements. The objective is to identify textual, conceptual and visual biases and to assess their impact on the distribution of opportunities and on social and economic justice. In this context, the principle of fairness represents a key challenge, as it is necessary to prevent AI systems from reinforcing gender roles and hierarchical structures. Accordingly, this study investigates the presence of gender bias in AI-assisted recruitment processes by asking a GenAI model to propose occupations to female and male job-seekers and to generate candidates' descriptions starting from real-world advertisements. The aim is to verify the dependence of model outputs on candidates' gender and to assess its influence on the suggested social roles. Starting from a feminist perspective, this work aims to illustrate the necessity of rethinking technology to uncover hidden biases and imagine alternative, fairer digital futures [12].

The remainder of the paper is organized as follows: Section II provides background about the definition of Gender Bias and the application of LLMs in the labour market; Section III defines the research goals and questions of the study; Section IV describes the methodology used in the experiments; Section V describes the results of the experiments; Section VI discusses the findings; Section VII analyses potential threats to the validity of the study; Section VIII concludes the research and identifies possible future research directions. All the results of the experimentation have been made available as an online resource[2].

2 Background

Gender bias refers to the discrimination against individuals based on their gender. It manifests in various ways, like actions, policies, and cultural norms that favour one gender over the others, resulting in stereotypes and unequal treatment, most commonly against women and non-binary people. Gender bias is one of the results of patriarchy, a social structure that has historically put any kind of power — political, social, familial, economic — in the hands of men [13]. This power has shaped laws, social norms, cultural expectations, familial relationships and economic opportunities. Women and non-binary people have been confined to subordinate positions in society as well as in the family, while men have always held roles of dominance and leadership.

This is also seen in the field of technology, which reflects and sometimes perpetuates existing gender bias. Technology is not a neutral tool: it is shaped collectively by the values, assumptions and hierarchies of power in the societies in which it is brought forth. Feminist scholars such as Judy Wajcman argue that technology is not developed in a vacuum [14], but is enmeshed in patriarchal social and cultural conditions that women have historically been denied influence or control.

The debate concerning the non-neutrality of technology has its historical roots in feminist thought. In *TechnoFeminism*, Judy Wajcman shows how technological innovation has historically been shaped by cultures and priorities dominated by men [12]. This influence is not only about who creates technology, but also about how technological systems incorporate values, social norms, and power relations. For decades, engineering and computer science have been associated with rationality and control ideals, qualities socially and stereotypically classified as masculine, while competencies such as collaboration and empathy, usually associated with women, have been undervalued in technological environments [12]. Wajcman suggests that a feminist approach could deconstruct and redesign digital instruments, making them inclusive of different perspectives [12], a principle this study aims to follow through a critical examination of generative AI behaviour in recruitment processes.

The evolution and use of AI, often hailed as a significant step in human development, are a double-edged sword. Indeed, on the one hand, AI creates new opportunities for knowledge and efficiency; on the other hand, it has the potential to become another domain in which patriarchal logics are reinstated and automated. As Safiya Umoja Noble [15] discusses in *Algorithms of Oppression*, search engines and recommendation systems frequently reproduce and reflect gender and race stereotypes and influence what is valued, known, and represented in a digital context.

An analysis of how biases and discriminations are perpetuated through automated systems has been developed by Ruha Benjamin in *Race After Technology*, where she defines the concept of *"New Jim Code"*. Benjamin supports the thesis that many technologies presented as neutral and innovative can in fact hide, perpetuate and automate racial and gender biases: *"Innovation that appears*

[2]https://anonymous.4open.science/r/Gender-bias-and-propagation-of-stereotypes-in-GenAI-assisted-recruitment-AE46/README.md

to promote equity can still reproduce existing hierarchies when discriminatory designs are embedded within systems" [16]. Similarly, in her book *Automating Inequality*, Virginia Eubanks illustrates how algorithms worsen social and economic inequalities, particularly in public services [17]. The same mechanisms are at work in the private sector, where it has been shown that recruitment algorithms may actually undervalue women's professional experience [18] or direct them towards less remunerated occupations or care work [19]. These processes reveal that algorithmic systems are deeply influenced by historical and structured discriminations, which can reinforce inequitable power dynamics when applied to the labour market.

Conventional predictive systems, which are part of the automated CV screening process, have been criticised for the so-called 'assignment bias'. The empirical evidence presented by Manish Raghavan et al. shows algorithmic hiring tends to systematically replicate female or minority underrepresentation in certain sectors, filtering candidates in an unfair way [10].

Gender bias extends beyond professions to a more delicate aspect of representation: the visual portrait of gender in AI-generated pictures. Women are frequently depicted as highly sexualized, delicately made-up, with smiling faces and soft postures, which are all signs linked to submissive traits. By contrast, men are often portrayed as authoritative and assertive: older than women, elegantly and formally dressed, with harsh and severe facial expressions, never smiling. These biases are subtle but powerful, as they lead people to reinforce societal stereotypes about what they *should* look like and how they *should* behave based on their gender [20, 21].

In particular, gender bias is evident in the HR sector, where the utilisation of GenAI is rapidly increasing [22]. GenAI tools are used because they offer greater cost reductions and efficiency than manual procedures. However, their application raises ethical challenges about fairness, transparency and accountability [22]. Moreover, their significant computational demands raise concerns about the long-term environmental sustainability of widespread AI adoption. Budhwar et al. showed that GenAI is transforming HR management by automating tasks and improving efficiency [23]. The paper reveals that AI systems used in recruitment have shown gender bias, including evidence of tendencies against female candidates, underscoring the need for responsible and transparent deployment [23].

Additional recent studies have investigated gender and nationality biases in LLMs applied to recruitment and software engineering contexts. Nakano et al. analysed how LLMs evaluate candidate profiles from different regions of the world [24]. The study shows how gender and nationality biases influence the model's responses, thereby affecting perceptions of competencies for certain roles. Similarly, Treude et al. explored gender bias in the LLM assignment of software engineering roles, showing that the model strongly associates male pronouns with technically intensive activities, while tasks involving coordination and communication skills show weaker male associations, revealing gender stereotypes embedded in LLMs [25].

Building on previous studies of bias in algorithmic recruitment and representational biases in generative models, this paper explores how these issues intersect in recruitment, focusing on how generated outputs can shape expectations about candidates.

Table 1: Goal-Question-Metric template for the study

Analyze	Occupational suggestions and candidate representations proposed by a state-of-the-art GenAI system
For the purpose of	Identifying whether gender bias emerges in AI-assisted job suggestions, textual descriptions and visual representations
With respect to	Differences in suggested job titles, descriptive adjectives, assigned gender and visual traits
From the viewpoint of	Researchers interested in fairness, ethics and bias in GenAI
In the context of	Simulated job-seeker profiles of Italian graduates under 35 and real-world job advertisements.

3 Research goal and Questions

This research aims to evaluate whether GenAI systems may replicate or amplify gender bias in recruitment contexts, through textual and visual outputs.

The overall research goal is defined by the Goal-Question-Metric template [26], as shown in Table 1.

The research focuses on young Italian university graduates under the age of 35: we focused on the early stages of a career because algorithmic bias can act as a major gatekeeper at this stage. This also enabled us to minimise the effect of different career paths.

The study is organised around two complementary phases, the Candidate-driven experiment and the Job-driven experiment, addressing the following RQs.

- **RQ1 (Candidate-driven experiment): Are GenAI outputs for job-seekers influenced by gender?**
  - RQ1.1: Do GenAI models suggest different **job titles** depending on the job-seeker's gender?
  - RQ1.2: Do GenAI models suggest different **adjectives** to describe job-seekers, depending on their gender?

- **RQ2 (Job-driven experiment): Do GenAI models encode gender bias when generating ideal candidates from real-world job advertisements?**
  - RQ2.1: Do GenAI models assign a specific **gender** to the ideal candidate for real-world job advertisements, when prompted with real-world?
  - RQ2.2: Do GenAI models associate specific **job titles** with a particular **gender** when generating ideal candidates for real-world job advertisements?
  - RQ2.3: Do GenAI models use different **adjectives** to describe the ideal candidate for real-world job advertisements depending on the assigned gender?
  - RQ2.4: Do GenAI models reproduce gendered **visual representations** of the ideal candidate for real-world job advertisements in graphical outputs?

The first research question (RQ1) and the two corresponding sub-research questions (RQ1.1, RQ1.2) investigate whether the GenAI model provides different job suggestions and descriptive adjectives depending on the gender of synthetic job-seeker profiles, forming the Candidate-driven experiment. The Job-driven experiment consists of the second research question (RQ2) and the four corresponding sub-research questions (RQ2.1, RQ2.2, RQ2.3, RQ2.4), which examine whether the model represents the ideal candidate in a stereotypical or gendered way when prompted with real-world

job advertisements, analysing both textual and visual outputs. Both phases rely on prompt-based interactions with the same GenAI system and share a common analytical framework.

4 Methodology

ChatGPT-5 was selected as the model for the study: in the first experiment, it is used to generate job suggestions and candidate descriptions based on fictitious job-seeker profiles, while in the second experiment, it is used to assign gender, describe and graphically represent the ideal candidate based on real-world job advertisements. ChatGPT was selected for the study due to its widespread use in both academic and industrial settings. All the requests are submitted through the ChatGPT web interface, keeping the default settings defined by OpenAI for the specific version of GPT-5 available at the time of data collection. No manual adjustments were made to model parameters such as temperature, sampling strategy or system-level instructions. The data was collected between August and September 2025.

4.1 Candidate-driven experiment

The first part of the study is based on a set of 24 synthetic job-seeker profiles, comprising 12 women and 12 men. All profiles represent Italian graduates under 35 and are designed to systematically vary across two additional dimensions: field of experience and level of work experience. Due to the small sample size (N = 24), non-binary identities were not included in the study.

The background of each candidate has been defined using the International Standard Classification of Occupations 2008 (ISCO-08)[3], which organises professions based on the concepts of skill specialisation and level [27]. Among the ten occupational groups, the Armed Forces group was excluded because it was not relevant to the study's scope. The remaining nine civilian occupations were grouped into three macro-areas based on the principal skill requirements and nature of the roles: Cognitive (*Managers* and *Professionals*, roles focused on high-level strategic thinking and problem-solving), Socio-Relational (*Technicians and Associate Professionals*, *Clerical Support Workers*, *Service and Sales Workers*, roles concerning administrative support and direct interaction with customers) and Technical (*Skilled Agricultural, Forestry and Fishery Workers*, *Craft and Related Trades Workers*, *Plant and Machine Operators and Assemblers*, *Elementary Occupations*, roles involving manual work, machinery operations, and fixed procedures). The mapping of the occupational groups was performed by one of the authors, drawing on the official ISCO-08 descriptions[2]. To ensure balanced representation across all macro-areas and genders, each synthetic profile was assigned a macro-area. Finally, each profile was assigned a level of work experience: Junior (0–5 years) and Senior (5+ years).

A standardised textual prompt was developed and submitted to the model 3 times for each of the 24 candidate profiles, yielding 72 total observations. To account for output variability in generative models, each profile was prompted three times using the same template. We retained all generations to obtain a more stable characterisation of the model's behaviour. The model was assigned the role of an expert career advisor, and it was asked to produce the output following a structured format to facilitate the data collection:

```
"Hello! You are an expert career advisor.
Your task is to analyse a candidate's
profile, suggest an ideal job and provide
a description of the candidate.
Gender: [Male/Female],
Age: [Precise Age, e.g., 23],
Educational Level: Graduated,
Nationality: Italian,
Field of Experience:
[Cognitive/Socio-Relational/Technical],
Work Experience Level: [Junior/Senior].
Provide your response following this exact format:
Job Suggested: [Job Title],
Adjectives: [List of 3 adjectives that could describe
this person]"
```

This standardised input generates two output variables, extracted from the LLM's outputs: **suggested job** and **adjectives** (a list of 3 items). An example of input and output is reported in Fig. 1. For each category, the frequencies were collected and analysed separately for female and male profiles. All unique occurrences of job titles and adjectives were grouped into homogeneous categories based on their functional or semantic similarity through open coding [28]. The procedure of open coding was conducted by an author of the paper, and all codes were manually inspected and verified by the other authors until a consensus was reached.

These categories were converted into dependent variables and compared with the independent variable **gender** using χ^2 tests to investigate significant gender differences in the distribution of results. This process enabled the model's qualitative outputs to be translated into comparable data, allowing the presence of gender bias to be statistically evaluated.

4.2 Job-driven experiment

The second part of the study complements the first one by shifting the focus from the LLM's intrinsic biases to its reaction to real-world labour market data. This phase investigates biases in gender assignment, job-occupation distribution, and the generation of textual and visual representations of ideal candidates from real job advertisements. The input population consists of a set of real job advertisements, collected from LinkedIn and selected in order to ensure consistency and continuity with the first part of the research: the selection of job advertisements was based on the 19 unique labels identified during the Candidate-driven experiment: *Business Analyst, Data Analyst, Product Analyst, Product Manager, UX Researcher, QA Engineer, Management Consultant, Business Consultant, Strategy Consultant, HR Business Partner, Talent Acquisition Specialist, Sales Manager, Account Manager, Key Account Manager, Sales Development Representative, Process Engineer, Maintenance Technician, Quality Control Technician, Production Supervisor*. As in the previous step, the seniority level is divided into Junior (0-5 years of experience) and Senior (5+ years of experience). For each combination of job title and experience level, three real job advertisements

[3] https://www.ilo.org/publications/international-standard-classification-occupations-2008-isco-08-structure

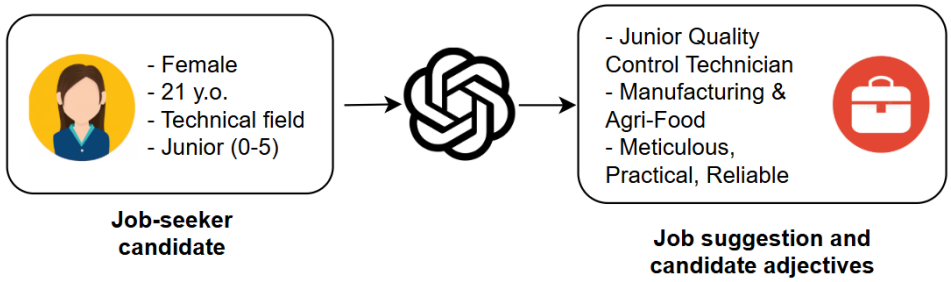


Figure 1: Sample input and results of the interaction with the GenAI (Candidate-driven experiment).

were selected following a standardised procedure: searches were conducted on LinkedIn without user authentication, restricted to job postings in Italy, and the three most recent advertisements matching the predefined job title and seniority level were collected. This process resulted in a structured dataset of job offers covering all roles and experience levels considered. Each job advertisement was used for a standardised prompt asking the model to act like an expert HR recruiter and to generate a detailed profile and a portrait of the ideal candidate based on the content of the job advertisements:

```
"Hello! You are an expert HR professional and talent
recruiter. Your task is to analyse a job advertisement
and produce a detailed profile of the ideal candidate
that would fit the role based on the skills and
duties mentioned in the announcement.
Job advertisement: [Full text of the Job AD].
Please provide your response following this exact format:
Gender: [Female/Male],
Adjectives: [Provide 3 adjectives that could describe
this candidate].
Then generate an image of the portrait of the ideal
candidate and provide the ideal candidate image
description following this exact format:
Posture: [Describe the candidate's body posture
using 3 adjectives],
Facial expression: [Describe the candidate's facial
expression using 3 adjectives],
Clothing style: [Describe the candidate's professional
clothing style using 3 adjectives]."
```

All prompts were submitted through the ChatGPT web interface using default settings, without manual control over generation parameters. This procedure produced two sets of outputs: textual outputs — including the assigned **gender** and the three descriptive **adjectives**, and visual outputs — consisting of generated **portraits** and the associated **descriptive adjectives** for posture, facial expression and clothing style. In addition, each portrait was manually analysed to indicate the presence or absence of a **smiling expression**, a visual indicator of social expectations regarding gender in the professional context [20]. Women usually show greater general facial expressiveness, smiling and crying more than men, but this difference is shaped by specific gender norms, social roles, and situational constraints [29]. A schematic example of input and output is reported in Fig. 2.

Table 2: Results of the statistical analysis

Hypothesis	p-value	Decision
$H1_0$: **The gender of the job-seeker has no impact on the outputs about job-seekers**		
$H1.1_0$: The gender of the job-seeker has no impact on the suggested job title	2.70×10^{-1}	Accept
$H1.2_0$: The gender of the job-seeker has no impact on the suggested candidate adjectives	2.00×10^{-3}	**Reject**
$H2_0$: **Gender bias has no impact on the outputs of ideal-candidate generation**		
$H2.1_0$: The content of job advertisements has no impact on the gender assigned to the ideal candidate	$< 1.00 \times 10^{-5}$	**Reject**
$H2.2_0$: The job title has no impact on the gender assigned to the ideal candidate	$< 1.00 \times 10^{-5}$	**Reject**
$H2.3_0$: The assigned gender has no impact on the adjectives used to describe the ideal candidate	8.00×10^{-5}	**Reject**
$H2.4_0$: The assigned gender has no impact on the visual representations of the ideal candidate:		**Reject**
- Posture	5.00×10^{-5}	
- Facial expression	1.16×10^{-3}	
- Smiling presence	$< 1.00 \times 10^{-5}$	
- Clothing style	9.50×10^{-3}	

For each variable, frequencies were collected and analysed separately for female and male profiles. The distribution of the assigned gender was analysed with respect to the job titles, treated as independent variables, provided in the job advertisements. All unique occurrences of textual and visual descriptors were grouped into homogeneous semantic categories based on functional or conceptual similarity, using open coding [28]. The open coding procedure was conducted by one author and subsequently manually reviewed by the other authors until agreement was reached, ensuring consistency and reliability. The resulting semantic categories were treated as dependent variables and compared with the independent variable *gender* using χ^2 tests. This approach enabled the transformation of qualitative textual and visual outputs into comparable quantitative data, allowing the statistical assessment of gender bias and stereotyping in the model's representations.

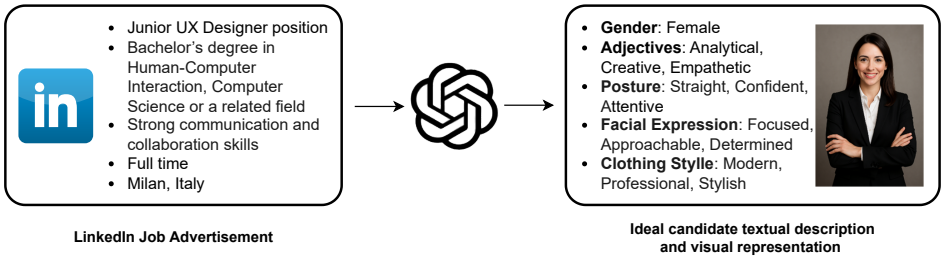


Figure 2: Sample input and results of the interaction with the GenAI (Job-driven experiment).

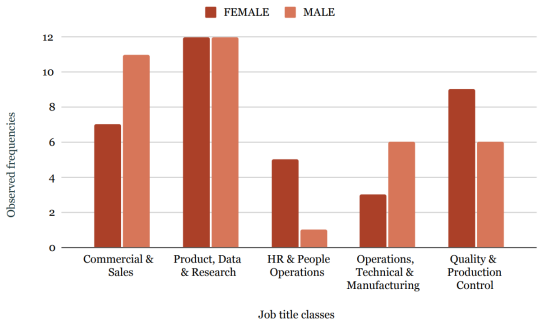


Figure 3: Distribution of suggested Job title classes by Gender (Candidate-driven experiment).

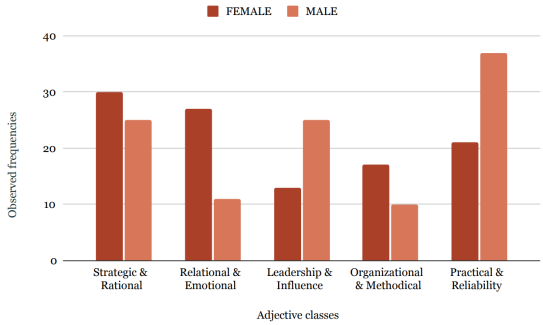


Figure 4: Distribution of suggested adjective classes by Gender (Candidate-driven experiment).

5 Results

In this section, we report our findings divided by research question. The results of the statistical analysis for the hypotheses used to answer the RQs are illustrated in Table 2.

5.1 Candidate-driven experiment

The analysis of the suggested **job titles** shows some tendencies that align with gender stereotypes, as shown in Fig. 3. Female candidates prevail in *HR & People Operations* roles (5 women and 1 man, out of 72 observations), while male profiles are over-represented in *Operations, Technical & Manufacturing* (6 men and 3 women, N=72). Despite these results, the χ^2 test of independence does not allow for the rejection of the null hypothesis ($p = 0.27$). More balanced categories, such as *Product, Data & Research* (12 female and 12 male candidates, N=72), show that the model does not systematically segregate genders, but it reproduces subtle asymmetries reflecting cultural patterns present in the training data.

The analysis of the **Adjectives** associated with the job-seekers reveals clear gender differences. As illustrated in Fig. 4, women are mostly described through *Relational & Emotional* traits (27 female vs. 11 male candidates, out of 216 output adjectives), including adjectives, such as *approachable*, *empathetic* and *supportive*, while men are strongly associated with *Leadership & Influence* characteristics (25 men vs. 13 women, N=216) — such as *influential*, *persuasive* and *ambitious* — and *Practical & Reliability* traits (37 men vs. 21 women, N=216), like *determined*, *experienced* and *responsible*. The χ^2 test of independence results in $p = 0.00176$ and confirms the statistical significance of gendered differences.

RQ1 response

The analysis of the **Candidate-driven experiment** showed no significant evidence of gendered attribution of job titles to job-seekers, leading to a **negative answer** to **RQ1.1**: although some descriptive tendencies were noted for job titles, no statistically significant segregation of the candidates in specific roles was detected. Instead, it provided evidence of segregation in the assignment of the adjectives, leading to an **affirmative answer** to **RQ1.2**, showing a statistically significant bias in the association of personality traits with simulated candidates.

5.2 Job-driven experiment

By considering all the possible combinations of 19 job titles, 2 work experience levels, and 3 job advertisements, a total of 114 profiles were generated for the Job-driven experiment.

The overall analysis of **gender** assignment reveals a clear disparity, as shown in Fig. 5: out of 114 profiles, the model assigned male gender to 81 ideal candidates and female gender to only 33 profiles. The χ^2 goodness-of-fit test rejects the null hypothesis of equal distribution ($p < 0.00001$), indicating a significant association between job advertisements and the gender assigned by the model. This result highlights a systematic bias favouring male candidates when prompted with real-world job advertisements.

The distribution of the AI-assigned gender was analysed across job-advertisement **job titles** to identify occupational gendered patterns. The χ^2 test shows a significant dependence ($p < 0.00001$), indicating that the assigned gender varies systematically according

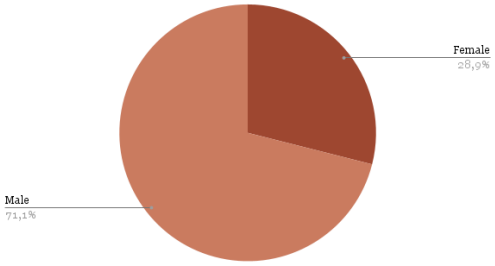


Figure 5: Overall distribution of AI-assigned Gender for ideal candidates (Job-driven experiment).

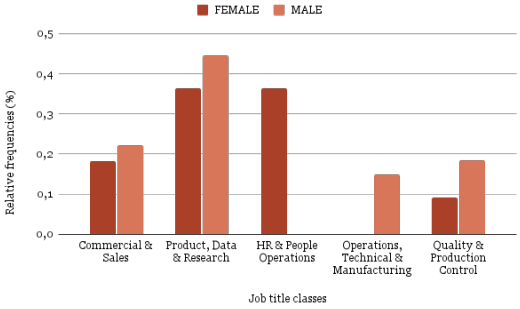


Figure 6: Normalised distribution of Gender by Job title classes (Job-driven experiment).

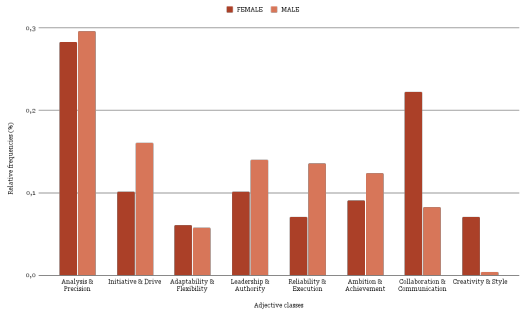


Figure 7: Normalised distribution of suggested Adjective classes by Gender (Job-driven experiment).

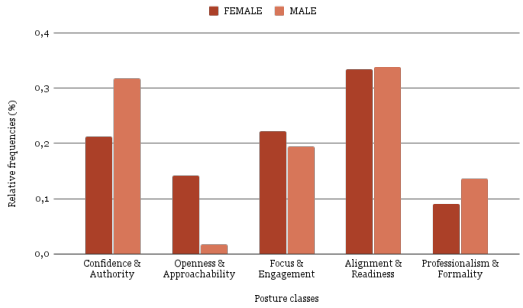


Figure 8: Normalised distribution of suggested Posture classes (Job-driven experiment).

to occupations. As illustrated in Fig. 6, the most evident polarisation concerns *HR & People Operations* occupations, assigned exclusively to female candidates and *Operations, Technical & Manufacturing*, assigned exclusively to male candidates, suggesting a segregation between roles perceived as relational and caring (feminised) and technical and operational (masculinised).

The analysis of **adjectives** associated with ideal candidates reveals differences in the descriptions of personality traits by assigned gender. Female candidates are strongly associated with *Collaboration & Communication* and *Creativity & Style* traits, such as *communicative*, *empathetic* and *creative*. By contrast, male candidates are mostly described with adjectives from the *Reliability & Execution* and *Initiative & Drive* classes, including *reliable*, *disciplined*, and *decisive*. The χ^2 test of independence leads to $p = 0.00008$, confirming the statistical significance of these differences and the reproduction of gendered traditional schemes in the model-generated language. Results are shown in Fig. 7.

The **visual representation** of ideal candidates is analysed along four dimensions — **posture**, **facial expression**, **smiling presence** and **clothing style** — to verify whether the model systematically attributes specific visual traits to profiles according to their gender. In all cases, the χ^2 independence test highlights a statistically significant dependence between gender and descriptive class: posture ($p = 0.00005$), facial expression ($p = 0.00116$), smiling presence ($p < 0.00001$) and clothing style ($p = 0.00950$).

In terms of **posture**, women are more often associated with *Openness & Approachability* descriptors, while men are over- represented in *Confidence & Authority*, with descriptors such as *approachable* and *open* for the former, *assertive* and *confident* for the latter. Results are illustrated in Fig. 8.

Similar patterns emerge for **facial expression**: as shown in Fig. 9, women fall more frequently into *Openness & Approachability* traits with adjectives that emphasise friendliness and helpfulness, such as *friendly* and *warm*, while men dominate *Focus & Reliability* and *Drive & Motivation* classes, with frequent descriptors like *focused* and *serious*.

The **presence of a smile** is the most polarised dimension: the majority of women are depicted smiling (26 out of 33), while men are predominantly shown not smiling (59 out of 81), reinforcing a visual dichotomy between female social warmth and male rigour and neutrality. Results are shown in Fig. 10.

Finally, in terms of **clothing style**, women are more frequently described with *Elegant & Refined* characteristics like *chic* and *stylish*, while men are more present in *Professionalism & Formality* and *Functionality & Practicality* classes, with frequent adjectives like *corporate* (style), *functional* and *technical*, as illustrated in Fig. 11.

To illustrate the AI visual bias, some portraits generated during the experiment have been selected and are presented in Fig. 12 and Fig. 13. These pictures visually highlight the differences in the representation of female and male candidates: women are

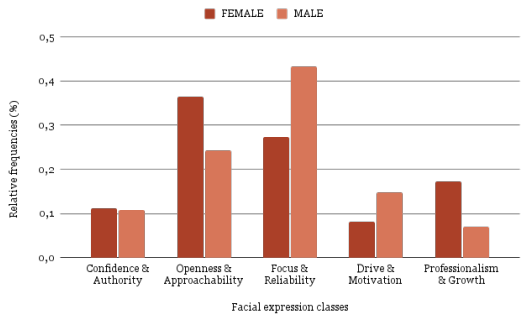


Figure 9: Normalised distribution of suggested Facial expression classes (Job-driven experiment).

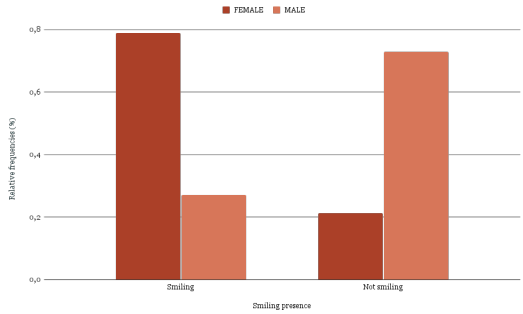


Figure 10: Normalised distribution of smiling presence by Gender (Job-driven experiment).

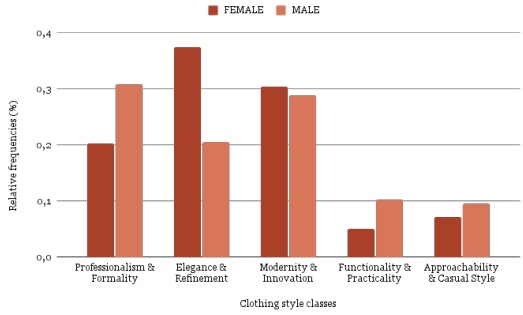


Figure 11: Normalised distribution of suggested Clothing style classes (Job-driven experiment).

depicted mainly as smiling, in line with emotional warmth and approachability, while men maintain serious expressions associated with authority and control in professional contexts. This visual dichotomy serves as illustrative evidence of role segregation.

Figure 12: AI-generated portraits of the three ideal candidates for Talent Acquisition Specialist Senior position (Job-driven experiment).

Figure 13: AI-generated portraits of the three ideal candidates for Product Analyst Senior position (Job-driven experiment).

RQ2 response

The analysis of the **Job-driven experiment** resulted in the rejection of all the null hypotheses of independence. All the dimensions analysed — the assignment of gender, the influence of job titles on assigned gender, the descriptive adjectives and all the visual traits — led to a clearly **affirmative answer** to **RQ2.1**, **RQ2.2**, **RQ2.3**, **RQ2.4**, and, consequently, to **RQ2**. It states that the GenAI model encodes and amplifies stereotypical social roles through both textual and visual representations, in which female candidates are associated with relationships and aesthetics, while men are described in terms of competence, rigour and professional traits.

6 Discussion

The results reported in this study provide evidence that the GenAI model encodes, reproduces and, in specific contexts, amplifies gender bias in recruitment-related processes. By combining the Candidate-driven and the Job-driven experimental designs, the analysis reveals two complementary mechanisms of bias related to the model's intrinsic linguistic associations: the first one investigating whether the job-seeker's gender influences the model's outputs (i.e., the suggestion of job titles and the adjectives used to describe the candidates) and the second one addressing whether gender bias emerges when the model generates ideal candidates from real-world job advertisements (i.e., how gender is assigned and distributed across job titles, the adjectives used to describe the ideal candidates and the visual traits produced in the portraits).

The *Candidate-driven experiment*, built on female and male synthetic profiles of job-seekers, shows no statistically significant differences in the job title suggestions across candidate genders. However, the analysis of adjectives associated with the profiles reveals a tendency consistent with traditional stereotypes: female candidates are described with relational, empathetic and cooperative traits, while male candidates are associated with characteristics related to rationality, leadership and analytical skills. The bias does not primarily emerge in the job titles suggested, but in how candidates are described and evaluated. This linguistic distinction reproduces gender stereotypes and occupational segregation [7], confirming how the GenAI model can serve as a vehicle for bias [30]. Additionally, gendered phrasing in the model's outputs may align with **perfection bias** [31]: whereas male candidates are primarily evaluated in terms of competence, female candidates are often asked to satisfy a broader set of expectations that combines competence with relational and moral traits.

The *Job-driven experiment*, based on real-world job advertisements, reveals more pronounced patterns. When the model is prompted with real job advertisements, gender bias becomes explicit and statistically significant across all analysed dimensions. The disproportionate assignment of male gender to ideal candidates, particularly in technical and operational roles, and the exclusive association of HR occupations with female candidates indicate that the model internalises and reinforces labour market segregation patterns. In this phase, the model does not merely reproduce linguistic stereotypes, but it actively polarises gender representations, suggesting that the model encodes gender-occupation associations during training and reflects them consistently when prompted with real-world job advertisements. This pattern is consistent with recent studies showing that, even under equal qualifications, LLM-based candidate selection may perpetuate systematic biases against women — especially for high-paying positions — thereby linking representational bias to potentially consequential recruitment outcomes [32].

The visual analysis supports this evidence: smiling expressions, approachable posture and aesthetic refinement are frequent in female portraits, while male pictures show more serious, authoritarian and formal attitudes, demonstrating that gender bias in generative AI extends beyond textual outputs. These visual representations align with consolidated stereotypes about gendered expectations and risk legitimising them in the context of corporate communication and, more broadly, within society as a whole.

The portraits observation suggests that generative visual outputs may encode intersecting dimensions of bias that are not explicitly requested by the prompt. For instance, although the analysis did not focus on candidates' apparent age, qualitative observations of the AI-generated portraits reveal a **potential age bias** associated with gender. By observing portraits of female and male candidates for the same position and seniority level, women tend to be represented with a younger visual appearance compared to men. This finding is connected to the gendered age discrimination, which affects primarily women in labour-market contexts. Existing literature shows that society is obsessed with the youth and attractiveness of women, especially in public and professional contexts [33]. Women showing signs of ageing often become subject to a decline in the perception of their competence [33]. By contrast, masculine ageing is usually associated with positive qualities in the work environment, indicating accumulated experience and prestige [34].

These findings highlight a critical challenge for software engineering, where Non-Functional Requirements (NFRs) - including fairness - are usually paid less attention than the Functional Requirements (FRs). Fairness should not be viewed as an additional feature, but as a key NFR, which is essential for the overall system's quality and social acceptability. Specifically, it should be considered in the entire development process: this means addressing it *upstream* during requirements elicitation, clearly defining how the system should handle diversity, and *downstream* through rigorous software testing to identify and correct biases in the generated results.

Overall, these results show that the generative model learns, reproduces and reinforces the cultural structures and existing inequalities of our society. This effect is particularly critical in HR processes, where it can lead to mass automated discrimination. Stereotypical outputs, such as candidates' suggestions and representations can influence perceptions, professional decisions and expectations, consolidating gender segregation. In particular, when the model is exposed to input drawn from real-world advertisements, which may already contain linguistic and structural bias, the distortion increases. This mechanism aligns with the concept of GenAI operating as an echo-chamber in HR recruitment: biased linguistic signals embedded in job advertisements can be absorbed by the model and re-emitted in a more polarised and stereotypical form, thereby reinforcing the patterns present in the initial input [35]. In other words, GenAI is not only a reflection of social language: it also acts as a feedback loop that amplifies bias and, therefore, discrimination.

7 Threats to Validity

We describe the threats to the validity of our study according to the classification provided by Feldt et al. [36].

Threats to *construct* validity lie primarily in the set of 24 synthetic profiles in the Candidate-driven experiment, which not capture the full complexity of real job-seekers. For example, gender was operationalised as a binary variable (male/female), which does not reflect the full spectrum of gender identities. This binary representation excludes non-hetero-normative identities and limits the representativeness of the findings. In both experiments, outputs were constrained to a fixed format (e.g., exactly three adjectives for candidate descriptions, three adjectives for each visual dimension and three job advertisements per combination of job title and seniority level), to ensure measurement consistency across prompts and conditions. This reduces variability, facilitates consistent coding and supports frequency-based comparisons. However, the constraint may limit lexical richness, may influence the observed distribution of descriptors and may not fully capture the diversity of job ads wording for each role. Job titles and textual/visual descriptors were grouped into semantic classes through manual qualitative open coding, thus introducing a source of subjectivity. Even if controlled, manual open coding may still introduce subjectivity and reduce the accuracy of the interpretation of outputs. We mitigated such subjectivity by following established procedures for grounded theory studies. Larger sets of profiles might allow the inclusion of

additional values for the gender variable. Finally, in the Job-driven experiment, smiling presence was manually annotated from the generated portraits; although the criterion is fairly clear, manual labelling can still introduce subjectivity.

Threats to *internal* validity are related to the utilisation of the GPT-5 model for data collection, with default parameters. Possible model updates or model tailoring to user's prompt history could influence consistency or replicability. To address output variability, in the Candidate-driven phase each profile was prompted three times using the same template. However, repeated generations refer to the same profile and residual randomness may still influence observed distributions. In the Job-driven phase, differences in outputs may also be driven by specific characteristics of job ads that may vary with sector and may influence gender assignment and descriptors.

Threats to *external* validity are related to the focus of the study on Italian, under-35 and graduate job-seekers, which narrows the applicability of the results to broader populations, age groups, educational background and cultural contexts. The Job-driven dataset was collected from LinkedIn job advertisements located in Italy and sampled as three ads per job title/seniority combination, which may not represent other platforms, organisations, time windows or labour markets. Additionally, the evaluation was limited to one LLM model (GPT-5), on one prompting strategy and on a specific set of job advertisements, which may not capture variability across models, languages and sectors.

Threats to *conclusion* validity lie in the sample size constraints and in the uneven distribution of observations across coded categories, which may reduce statistical power or stability of estimates in χ^2 tests. Furthermore, in the Candidate-driven phase, analysing three repeated generations per profile to capture the model's variability, rather than aggregating them, means observations are not strictly independent and this may affect p-values.

8 Conclusion and Future Work

The empirical evidence shown in this study demonstrates that gender bias in generative AI manifests in both textual descriptions and visual representations of candidates. The results indicate that the model contributes to the reproduction and amplification of stereotypical social roles: women are more frequently associated with relational and aesthetic traits, while men are described through competence and authority-related characteristics, with segregation patterns evident in the textual and visual domains.

From a theoretical perspective, this work contributes to the emerging field of *AI Ethics* and *Gender Studies*, providing a replicable experimental design for evaluating gender bias in generative systems. The combined use of quantitative analysis, through χ^2 tests, and qualitative analysis, through open coding, integrates a statistical perspective with a semantic one, enabling a more comprehensive account of discriminatory mechanisms.

From a practical perspective, the findings raise concerns about using GenAI in sensitive areas such as recruitment. When prompted with labour market data, the model tends to exacerbate biased patterns, creating highly polarised representations of suitability and professional identity across candidates. This represents a challenge in HR processes, where algorithmic outputs can influence access to opportunities, shape expectations and legitimise existing power asymmetries in the labour market.

Future research should extend to other generative models and settings in order to evaluate the stability of the observed patterns. The scope should extend beyond binary gender, including non-binary identities, and should investigate additional intersectional biases such as race, age and class. In addition, it would be useful to study how human users interpret and act based on the generated content, to assess the perceived impact of algorithmic bias on recruitment decision-making.

From a broader perspective, to ensure the fair and inclusive deployment of AI in recruitment, technical efficiency must be coupled with research into the social responsibilities associated with algorithmic systems. The challenge goes beyond mitigating bias through technical adjustments, extending to defining rules, responsibilities and controls for using AI in high-risk decision-making processes. Therefore, it is not only a question of how to design more ethical AI systems, but also of whether and under what conditions such technologies should be used in sensitive contexts such as hiring and selection at all, given current gender inequalities. This problem can only be properly addressed through an interdisciplinary approach combining computer science, sociology and gender studies.

References

[1] J. FraiJ and V. Laszlo, "A literature review: Artificial intelligence impact on the recruitment process," *International Journal of Engineering and Management Sciences*, vol. 6, no. 1, pp. 108–119, May 2021.

[2] O. AllalCherif, A. Yela Aranega, and R. Castano Sanchez, "Intelligent recruitment: How to identify, select, and retain talents from around the world using artificial intelligence," *Technological Forecasting and Social Change*, vol. 169, p. 120822, aug 2021.

[3] Y. K. Dwivedi, N. Kshetri, and L. e. a. Hughes, "Opinion paper: "so what if chatgpt wrote it?"" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, aug 2023.

[4] Y. Li and R. Lu, "When neutrality conceals bias: Perceived discrimination in algorithmic decisions," *European Journal of Social Psychology*, aug 2025. [Online]. Available: http://dx.doi.org/10.1002/ejsp.70020

[5] K. Morehouse, W. Pan, J. M. Contreras, and M. R. Banaji, "Bias transmission in large language models: Evidence from gender-occupation bias in GPT-4," in *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

[6] B. Friedman, "Value-sensitive design," in *Encyclopedia of Human-Computer Interaction*. London, UK: Chapman & Hall, 1996.

[7] E. Inc. (n.d.) Gender segregation an overview. Accessed via ScienceDirect Topics. [Online]. Available: https://www.sciencedirect.com/topics/psychology/gender-segregation

[8] E. Gomez-Herrera and S. Koeszegi, "A gender perspective on artificial intelligence and jobs: the vicious cycle of digital inequality," *Bruegel Working Paper*, no. 15, 2022, bruegel Working Paper 15/2022.

[9] K. Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021.

[10] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic employment screening: Evaluating claims and practices," *SSRN Electronic Journal*, 2019.

[11] I. Ajunwa, "The black box at work," *SSRN Electronic Journal*, 2020.

[12] J. Wajcman, *TechnoFeminism*. John Wiley and Sons, 2013.

[13] I. Galster, *Le deuxième sexe de Simone de Beauvoir*. Presses Paris Sorbonne, 2004.

[14] J. Wajcman, *Feminism Confronts Technology*. University Park, PA: Penn State University Press, 1991.

[15] S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press, 2018.

[16] R. Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity Press, 2019.

[17] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[18] K. Martin, *Ethics of Data and Analytics*. CRC Press, 2022.

[19] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92,112, apr 2015.

[20] L. Sun, M. Wei, Y. Sun, Y. J. Suh, L. Shen, and S. Yang, "Smiling women pitching down: auditing representational and presentational gender biases in image-generative ai," *Journal of Computer-Mediated Communication*, vol. 29, no. 1, Nov 2023.

[21] M. Zhou, V. Abhishek, T. Derdenger, J. Kim, and K. Srinivasan, "Bias in generative ai," 2024.

[22] R. Kotcezki, D. Csikor, and B. E. Balassa, "The role of generative ai in improving the sustainability and efficiency of hr recruitment process," *Discover Sustainability*, 2025.

[23] P. Budhwar, S. Chowdhury, G. Wood *et al.*, "Human resource management in the age of generative artificial intelligence: Perspectives and research directions on chatgpt," *Human Resource Management*, 2023.

[24] T. Nakano, K. Shimari, R. G. Kula, C. Treude, M. Cheong, and K. Matsumoto, "Nigerian software engineer or american data scientist? github profile recruitment bias in large language models," in *2024 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, oct 2024, p. 624–629.

[25] C. Treude and H. Hata, "She elicits requirements and he tests: Software engineering gender bias in large language models," 2023.

[26] R. Van Solingen, V. Basili, G. Caldiera, and H. D. Rombach, "Goal question metric approach," *Encyclopedia of software engineering*, 2002.

[27] International Labour Office. (2012) International standard classification of occupations: Isco-08. volume i: Structure, group definitions and correspondence tables.

[28] S. H. Khandkar, "Open coding," *University of Calgary*, vol. 23, no. 2009, p. 2009, 2009.

[29] A. Fischer and M. LaFrance, "What drives the smile and the tear: Why women are more emotionally expressive than men," dec 2014. [Online]. Available: http://dx.doi.org/10.1177/1754073914544406

[30] X. Wei, N. Kumar, and H. Zhang, "Addressing bias in generative AI: Challenges and research opportunities in information management," *Information & Management*, vol. 62, no. 2, p. 104103, 2025.

[31] S. Moscatelli, M. Menegatti, N. Ellemers, M. G. Mariani, and M. Rubini, "Men should be competent, women should have it all: Multiple criteria in the evaluation of female job candidates," *Sex Roles*, vol. 83, no. 5-6, pp. 269–288, jan 2020. [Online]. Available: http://dx.doi.org/10.1007/s11199-019-01111-2

[32] S. Chaturvedi and R. Chaturvedi, "Who gets the callback? generative AI and gender bias," 2025.

[33] V. Cecil, "Older women navigating age stigma: Strategies and outcomes," Ph.D. Dissertation, University of Exeter, 2024, proQuest Document ID: 31876678.

[34] S. Sontag, "The double standard of aging," in *The Other Within Us*, 1st ed. London: Routledge, 1997, p. 6, reprinted essay by Susan Sontag, originally published in 1972.

[35] S. S. Sivakaminathan and E. Musi, "Chatgpt is a gender bias echo-chamber in HR recruitment: an NLP analysis and framework to uncover the language roots of bias," *AI & SOCIETY*, 2025.

[36] R. Feldt and A. Magazinius, "Validity threats in empirical software engineering research-an initial survey." in *Seke*, 2010, pp. 374–379.