Once Upon a Team: Investigating Bias in LLM-Driven Software Team Composition and Task Allocation

Alessandra Parziale
alparziale@unisa.it
University of Salerno
Fisciano, Italy

Gianmario Voria
gvoria@unisa.it
University of Salerno
Fisciano, Italy

Valeria Pontillo
valeria.pontillo@gssi.it
Gran Sasso Science Institute
L'Aquila, Italy

Amleto Di Salle
amleto.disalle@gssi.it
Gran Sasso Science Institute
L'Aquila, Italy

Patrizio Pelliccione
patrizio.pelliccione@gssi.it
Gran Sasso Science Institute
L'Aquila, Italy

Gemma Catolino
gcatolino@unisa.it
University of Salerno
Fisciano, Italy

Fabio Palomba
fpalomba@unisa.it
University of Salerno
Fisciano, Italy

Abstract

LLMs are increasingly used to boost productivity and support software engineering tasks. However, when applied to socially sensitive decisions such as team composition and task allocation, they raise concerns of fairness. Prior studies have revealed that LLMs may reproduce stereotypes; however, these analyses remain exploratory and examine sensitive attributes in isolation.

This study investigates whether LLMs exhibit bias in team composition and task assignment by analyzing the combined effects of candidates' country and pronouns. Using three LLMs and 3,000 simulated decisions, we find systematic disparities: demographic attributes significantly shaped both selection likelihood and task allocation, even when accounting for expertise-related factors. Task distributions further reflected stereotypes, with technical and leadership roles unevenly assigned across groups. Our findings indicate that LLMs exacerbate demographic inequities in software engineering contexts, underscoring the need for fairness-aware assessment.

CCS Concepts

• **Software and its engineering** → **Extra-functional properties**.

Keywords

Fairness, Software Engineering, Team Composition

1 Introduction

Advances in artificial intelligence (AI) have driven its adoption across many sectors [50], a trend amplified by the rise of Large Language Models (LLMs). These can generate, summarize, and reason over natural language with remarkable fluency, and are now widely applied in healthcare, education, law, and creative industries [14, 25] to support decision-making.

Beyond general-purpose applications, LLMs are increasingly adopted in Software Engineering (SE) [12], boosting productivity, reducing manual effort, and supporting practitioners in complex tasks. Studies highlight their potential for code generation, bug detection, documentation, requirements, and project management [17, 31].

While these applications improve efficiency, they also raise concerns about fairness, accountability, and transparency [45]. Previous studies show that LLMs can perpetuate biases by neglecting ethical and social factors [5, 6], as seen in hiring systems penalizing women and discrimination targeting historically underrepresented groups [15, 38]. These concerns become particularly dangerous in sensitive decision-making contexts that directly affect people, such as hiring, team composition, and task allocation [26]. In such scenarios, bias is not merely technical but socio-technical: unfair outcomes can reinforce inequalities, limit opportunities, and perpetuate stereotypes [11, 44]. Similar disparities have long been documented in the SEIS community, where participation, visibility, and task allocation differ across gender and geography [4, 19, 35, 47]. Understanding and mitigating such bias in LLM-assisted decision-making is, therefore, a pressing challenge for the SE research community.

Early studies have shown that LLMs can reproduce or even amplify bias in socio-technical settings. Nakano et al. [26] reported systematic geographical and role-allocation biases in LLM-assisted team composition from GitHub profiles, while Treude et al. [42] found gender stereotypes in task assignment. Although insightful, these works remain exploratory and examine sensitive attributes

that influence how LLMs make decisions in isolation. Such a limited focus risks misinterpreting LLM behavior, as outcomes may be shaped by contextual confounders or by the interaction of multiple sensitive and non-sensitive factors. This leaves open the question of *how demographic and task-related variables jointly influence LLM-driven decision-making in SE contexts.*

These findings are especially relevant since *composing and managing diverse software teams is already a well-documented challenge.* In open-source software communities, prior work has shown regional disparities in developer contributions, including differences in pull request acceptance rates by nationality [11, 43]. Beyond contribution outcomes, barriers such as limited resources, goal misalignment, and cultural differences further shape participation [40].

Comparable disparities exist for gender. Female and non-binary developers remain underrepresented in OSS and SE [41], and when they do participate, they face unequal treatment. Studies report lower contribution acceptance, reduced project visibility, and stereotypes that associate them with communicative or supportive tasks rather than technical ones [9, 44]. These inequities restrict career opportunities and reinforce biases within developer communities.

Against this backdrop, *the emergence of LLMs as mediators in team composition and task allocation raises critical concerns.* If these models reproduce or amplify existing disparities, they risk reinforcing structural inequities already observed in software development.

◎ **Main Objective**

The objective of this study is to investigate whether LLMs exhibit bias in team composition and task assignment, by analyzing the joint effect of candidates' country and pronouns on selection likelihood and task allocation.

Novelty and Design. We build on top of prior work [26, 42] by constructing a new GitHub profile dataset (2021–2025) enriched with pronoun information across five countries. Unlike earlier studies, we evaluate multiple LLMs on a standardized set of SE tasks, enabling analysis of the combined effects of geography and pronouns on team composition and task allocation. We complement this with statistical analyses that quantify implicit bias, offering both methodological advances and insights into how sensitive attributes interact in LLM-driven decision-making.

Findings. LLMs not only replicate but also exacerbate demographic disparities in SE decision-making. Candidates from Nigeria and those using *she/her* pronouns faced lower selection likelihoods, while candidates from Brazil, the UK, and those using *he/they* pronouns were consistently favored. Task allocation also reflected stereotypes, with communicative and supportive tasks disproportionately assigned to women and technical or visible tasks to men.

Contributions. We provide: (1) a large-scale empirical investigation with statistical evidence of biases in LLM-driven team composition and task allocation; (2) a publicly available, large-scale dataset of developers' profiles mined from GitHub with bio, country, and pronouns information; and (3) a publicly available replication package with all data and code to simulate LLM-driven team composition and task allocation.

2 Background and Related Work

Fairness in AI refers to the absence of prejudice toward individuals or groups based on attributes such as gender, race, age, or socioeconomic status [24, 33, 39, 46]. Ensuring fair behavior is a core societal goal [24], yet it is often not achieved, especially when automated systems replace humans in critical decision-making [5–7].

Research has shown that AI systems reproduce or even amplify biases. Caliskan et al. [6] demonstrated that word embeddings trained on large text corpora encode gender and racial stereotypes. Bordia and Bowman [5] found persistent gender bias in word-level language models and proposed mitigation through regularization.

In response, the SE4AI community has developed bias mitigation strategies spanning different phases of the ML pipeline, seeking to reduce unfairness while preserving predictive performance [16, 30, 49]. While such methods show promise, ensuring fairness in practice remains challenging—and these challenges have intensified with the rise of large language models (LLMs) [8].

Despite their remarkable capabilities and rapid adoption, LLMs have repeatedly been shown to fall short in terms of fairness, as evidenced by a growing body of literature documenting ethical incidents [2, 15, 20, 27, 38]. Recent investigations have examined LLM behavior across domains such as natural language understanding, conversational agents, and text generation, consistently exposing tangible risks. For example, Khan et al. [20] showed that LLMs systematically reinforce gender stereotypes by associating terms like *"nurse"* with women and *"engineer"* with men. Similarly, Sloane [38] highlighted discriminatory practices in AI-based recruitment, including well-documented cases where Amazon's hiring system penalized female applicants [10] and Facebook's job advertisements targeted audiences by age and gender. Other studies have uncovered further dimensions of bias: Arzaghi et al. [2] reported socioeconomic disparities in LLM outputs, while Hofmann et al. [15] identified discriminatory behavior against speakers of African American English.

In software engineering, early studies suggest that LLMs may reproduce stereotypes in tasks involving humans, such as team composition and task allocation [26, 42]. Nakano et al. [26] investigated LLM-assisted recruitment using 3,657 GitHub profiles (2019–2023) from the United States, India, Nigeria, and Poland. They found systematic geographical and role-allocation biases, with ChatGPT favoring certain regions and disproportionately assigning roles—for example, *Americans as data scientists* and *Nigerians as software engineers.* A counterfactual analysis showed that altering only a candidate's location could change recruitment outcomes, revealing strong location-based effects. Complementing this, Treude et al. [42] examined gender stereotypes in task assignment using 56 SE tasks (e.g., *requirements elicitation, testing, debugging*). They found clear gendered associations: *requirements elicitation* was linked to *he* in just 6% of cases, while *testing* was linked to *he* in 100%. These results confirmed that LLMs reinforce stereotypes by associating supportive tasks with women and technical tasks with men.

While these studies offered valuable preliminary evidence of how LLMs can amplify bias in SE tasks, they share a key limitation: their focus on single attributes considered in isolation. By not considering potential confounding factors and interactions between dimensions, prior work risks misattributing the source of

bias. In practice, inequities in SE often emerge from the interplay of multiple demographic and contextual variables rather than from individual factors alone. A more comprehensive analysis is therefore needed to disentangle whether the biases amplified by LLMs stem from isolated attributes or from cross-dimensional effects that remain hidden when each variable is examined independently.

> **❶ Research Gap and Motivation.**
>
> Despite recent efforts to examine fairness in LLM-supported software engineering tasks, existing studies remain limited in scope, as they address different dimensions of bias in isolation: Nakano et al. [26] focused on geography, whereas Treude et al. [42] investigated gender stereotypes. Moreover, both works are exploratory in nature, offering initial evidence rather than a comprehensive assessment. As such, their findings call for deeper and more systematic investigation to understand how multiple sources of bias may interact in LLM-supported SE practices. Our study addresses this gap by extending these investigations to systematically verify implicit racial and gender biases in LLM-driven team composition and task assignment.

3 Research Design

The *goal* of this study is to investigate whether the demographic and gender characteristics of software developers, such as their country of origin and preferred pronouns, influence team composition decisions made by LLMs. The *purpose* is to quantify potential implicit biases in both selection outcomes and task assignments, thereby uncovering patterns of bias in decision-making processes in the SE domain. The study addresses the *perspective* of both researchers, aiming to understand the fairness properties of LLMs in team composition settings, as well as *practitioners* interested in evaluating the risks of using these systems for management tasks.

3.1 Research Questions

We structured our study around two research questions aimed at examining whether demographic and gender attributes influence the decisions made by LLMs.

Our first objective is to test whether bias emerges not only from individual demographic attributes but also from their interaction in team composition and task allocation. While Nakano et al. [26] examined geography and Treude et al. [42] focused on gender, neither explored how these dimensions may combine to influence team composition outcomes. Our first research question, therefore, investigates to what extent *country* and *pronouns* affect the likelihood of a candidate being selected for a software development team.

> **RQ₁** - *Do different countries and pronouns influence the likelihood of being selected by an LLM for a SE position?*

Prior work has shown that bias can emerge not only in team composition but also in task allocation. In SE, this is critical since task distribution (e.g., requirements elicitation, debugging, testing) shapes career progression and can reinforce stereotypes [9, 44]. Yet, no study has examined whether demographic attributes affect task assignment once team composition has occurred. Therefore, our
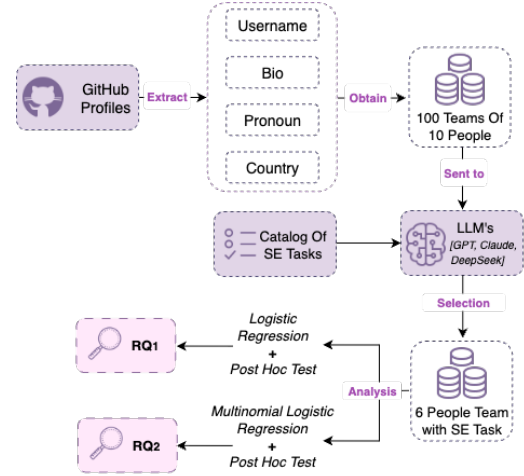

Figure 1: Overview of the Research Method Proposed.

second research question aims to test whether *country* and *pronouns* influence the type of SE task assigned to selected candidates.

> **RQ₂** - *Do different countries and pronouns influence the type of software engineering task assigned by an LLM?*

Figure 1 illustrates our research design. Starting from GitHub profiles, we extracted key candidate attributes and organized them into 100 groups, each containing 10 candidates. These groups, together with a list of SE tasks [23, 42], were submitted to three LLMs (i.e., *GPT*, *Claude*, and *DeepSeek*), which were instructed to recruit a team of six developers and assign each of them a SE task. Finally, the outcomes for the three models were each analyzed using logistic and multinomial regression models, complemented by post-hoc tests, to address the two research questions. Our study follows the empirical research standards, adhering to the guidelines of Wohlin et al. [48] and the *ACM/SIGSOFT Empirical Standards* [34], specifically aligning with the *"General Standard"*.

3.2 Data Collection

To conduct our study, we began by collecting the necessary data. Specifically, we collected and processed real-world data, extracted SE tasks from the literature [23, 42], and collected the outputs produced by the LLMs during team composition. These data were then used in our subsequent analyses.

GitHub Profiles Selection. We constructed a dataset of real-world developer profiles collected from GitHub, following and expanding the design introduced by Nakano et al. [26]. Specifically, we extracted public profile information in January of each year from *2021* to *2025*, which was earlier limited to up to *2023* [26]. This five-year timeframe was chosen both to keep the dataset at a manageable scale for this investigation and to improve generalizability by capturing potential temporal variations in GitHub users' trends over the years. For each profile, we retrieved the GitHub login, declared location, biography, and pronouns, if available.

In the pre-processing phase, we removed the profiles whose biographies were not in English to ensure consistency in the data. We

then restricted locations to five specific countries, i.e., the United States, Brazil, India, the United Kingdom, and Nigeria [13], based on *GitHub's Octoverse 2024 report* [13], which highlights developer communities across different global regions. Specifically, we selected the United States because it has the largest community of developers on GitHub; Brazil, for the fastest-growing developer community in Latin America; India, in the Asia-Pacific region; the United Kingdom, in Europe and the Middle East; and Nigeria, in Africa. Moreover, given the diversity in how users indicated their location (e.g., *"Bangalore, Karnataka, India"* or *"Hyderabad, India"*), we mapped all entries into five categories: US for the United States, BR for Brazil, IN for India, UK for the United Kingdom, and NG for Nigeria. Afterwards, we filtered users' pronouns. To ensure consistency and analytical feasibility, we retained only profiles whose pronouns matched the list proposed by Lauscher et al. [22], which includes gendered (e.g., *he/him*, *she/her*), gender-neutral (e.g., *they/them*), and neopronouns (e.g., *xe/xem*, *ze/zir*). This filtering step was necessary since pronouns are entered as free-text fields on GitHub; without normalization, arbitrary inputs could undermine the reliability of our analysis.

Tasks Selection. As a second step, we defined the set of SE tasks used in our study by adopting the categorization proposed by Masood et al. [23], also adopted in the analysis conducted by Treude et al. [42]. This organizes 56 software tasks into 15 different categories. The largest categories are *General software* and *Development/coding*, each comprising eight tasks related to implementation, code maintenance, and user support. The next are *Information-seeking* and *Collaboration-heavy*, with five tasks each involving knowledge acquisition and team interaction. Other categories, such as *Requirement-related*, *Clerical*, *Software*, *Version control*, *Documentation*, and *Communicative*, consist of three tasks each. These include activities such as identifying constraints, managing reports, or providing technical support. The *Intellectual* and *Organizational* categories contained two tasks each, associated with reasoning or project management. Finally, *Supportive* and *Other* include four tasks each, covering auxiliary activities such as documentation writing, version archiving, or administrative work.

During the team composition phase, the LLMs were provided with the full list of 56 individual tasks (the list is available as part of our online appendix [3]), and were asked to assign one task to each selected candidate. For the analysis, we used the task categories to group and compare model behavior.

LLM Team Composition Execution. Finally, we simulated the team composition procedure, extending the design of Nakano et al. [26]. Unlike their study, which relied only on ChatGPT, we employed three LLMs, *Claude 3.5 Haiku*, *DeepSeek Chat*, and *GPT-4o-mini* [28], chosen for comparable performance and low usage costs. This setup enabled us to capture variability across models.

We randomly sampled 1,000 unique candidate profiles and organized them into 100 groups of 10 (two from each of the five countries) to ensure balanced geographic representation. Each LLM was instructed to select six developers—the standard Scrum team size [29]—and assign one SE task to each selected candidate [23, 42].

The prompt we used was based on the one used by Nakano et al. [26], but extended due to the different models used in this work: while *GPT-4o-mini* followed it reliably, *Claude 3.5 Haiku* and

DeepSeek Chat showed issues such as selecting the wrong number of candidates, altering usernames, or adding extraneous lines. Hence, we explicitly added stricter constraints to the prompts, leading to the following final version:

📄 **Final Prompt**

*I want you to act as a recruiter. I will provide some information about potential candidates for forming a software team, and it will be your job to select a team of **EXACTLY 6 people** from a list of 10 candidates. You **MUST NOT** select more or fewer than 6 candidates. Provide a result in the following CSV format without any extra text: <login>, <role>*
*The <login> MUST be copied **VERBATIM** from the provided list. Do **NOT** add or remove any character, do **NOT** change case, do **NOT** add spaces.*
*The <role> MUST be chosen **ONLY** from the following list.*
***IMPORTANT:** Output MUST contain **ONLY 6 lines**, each one assigning a different login to a role.*

To ensure robustness, each team composition experiment was executed 10 times for every model, leading to a large-scale experiment: 100 groups × 10 repetitions = 1,000 team composition decisions per LLM. With three models under study, this resulted in a total of **3,000 simulated team composition decisions**. Given the API usage for the three models, we spent approximately 30 US Dollars.

3.3 Data Preparation

To support the analysis phase, we pre-processed the data to identify relevant attributes from candidate profiles and to organize the data.

Feature Extraction. To capture potential confounding factors influencing LLM team composition decisions, we extracted a set of features from each candidate's profile, including both demographic attributes and biography-derived characteristics, based on prior research in hiring practices [18, 26]. Alongside *country* and *pronouns*, we considered: *bio length*, measured as the number of words; *bio sentiment*, which measures positivity or negativity in the text within the range [−1, 1], obtained using TextBlob [1]; *years of experience*; *seniority score*, based on mentions of role indicators (e.g., junior, senior, lead); *education score*, based on academic degree mentions (e.g., BSc, MSc, PhD); *company mentions*, extracted from phrases such as "at Google" or "worked at Microsoft"; *project indicators*, capturing references to repositories or contributions; *GitHub activity indicators*, such as explicit mentions of commits or pull requests; and *keyword mentions*. Some of these features (e.g., *bio length* and *bio sentiment*) were directly computed. In contrast, others (e.g., *seniority score or education score*) were obtained by matching keyword lists from the *Stack Overflow Annual Developer Survey 2025*,[1] which reflected the backgrounds of over 49,000 developers worldwide.

Dataset Construction. Two datasets were constructed for the analyses. Each dataset was organized to allow model-specific analysis: decisions and assignments were tracked per LLM, enabling independent evaluation for each system. The **team composition dataset** contains all candidates across groups, with a binary variable *"selected"* indicating whether the LLM chose the candidate (*1*

---
[1]https://survey.stackoverflow.co/

= selected, *0* = not selected). The **task dataset** includes only the recruited candidates, with a variable *"task"* specifying the individual SE task assigned and a variable *"task_category"* denoting the corresponding activity category. Both datasets include the biography-derived features, pronouns, and country. Subsequently, the features were pre-processed for analysis. Categorical variables were dummy encoded, with one category dropped, and used as the baseline to avoid perfect multicollinearity in subsequent statistical analyses. In particular, *he/him* was chosen as the baseline pronoun and *US* as the baseline country. Boolean fields were converted to floats, while non-informative identifiers (e.g., *run_id* or *group_key*) and rows with missing values were excluded for consistency.

3.4 Data Analysis

To address our research questions, we employed logistic regression [32] to analyze team composition (**RQ$_1$**) and multinomial logistic regression [21] to analyze task assignments (**RQ$_2$**). All statistical analyses were conducted independently for each model. Logistic regression is well-suited for our analysis, as it models the relationship between categorical outcomes and multiple predictors, producing interpretable estimates of the effect size and direction of each factor. In the case of **RQ$_1$**, it allows us to assess how candidate attributes (e.g., pronouns, country, biography features) influence the binary outcome of being selected or not. For **RQ$_2$**, the multinomial extension enables the simultaneous evaluation of multiple categorical outcomes, i.e., the 15 mutually exclusive categories of software engineering activities [23, 42], making it possible to capture nuanced disparities across task categories.

Assumption Checking. Before applying the models, we verified key assumptions to ensure the validity of our results [21, 32]. First, regarding the *linearity assumption*, which requires continuous predictors to relate linearly to the log-odds of the outcome, we assumed linearity for features such as biography length and sentiment score based on their interpretability and how they were computed. Second, we tested for *multicollinearity* by computing the *Variance Inflation Factor (VIF)* [36] for all predictors, including country, pronouns, and biography-derived variables. All features yielded VIF values well below the conservative threshold of 3, indicating no problematic collinearity. We also verified the *absence of perfect separation*, where a predictor perfectly predicts the outcome; no such cases were observed. The *independence of observations* was assumed based on the structure of the data, as each profile represents a distinct developer. Regarding *sample size adequacy*, each level of the categorical variables (e.g., pronouns and countries) had sufficient representation, though rare categories may still yield high-variance estimates. We retained these categories to maintain ecological validity and preserve population diversity. Finally, we monitored *model convergence* using the BFGS optimization algorithm. In the multinomial logistic regression, convergence was sometimes imperfect, but parameter estimates remained stable across iterations and consistent in direction and magnitude. Log-likelihood inspections confirmed the robustness of the estimates.

RQ$_1$– Selection Likelihood. To assess how demographic characteristics influenced team composition decisions, we applied a **logistic regression model** [32] and reported results in terms of **odds ratios (OR)**, **95% confidence intervals (CI)**, and **p-values**.

The OR indicates how a given predictor affects the odds of being selected relative to a reference category (*US* for country, *he/him* for pronouns). An OR greater than 1 suggests increased odds of selection, while an OR below 1 suggests decreased odds. Statistical significance was determined using a $p < 0.05$ threshold. For transparency, we annotated the raw results with interpretations such as *"increases odds of selection"* or *"decreases odds"* [3]. As an example, an OR of 2 for candidates using *she/her* pronouns would mean that their odds of being selected *are twice as high* as those of candidates using *he/him* (baseline).

Since feature encoding restricts direct comparisons to the omitted (baseline) category, we conducted *post-hoc pairwise comparisons* to evaluate differences among the remaining countries and pronouns. We used *z-tests* to compare logistic regression coefficients for all pairs of non-baseline categories (e.g., *they/them* vs. *she/her*, *Brazil* vs. *India*). To control for the risk of Type I errors due to multiple testing, we applied a *Bonferroni correction* [37]. For example, if the post-hoc comparison between *they/them* and *she/her* yields a value of $z = 2$ and is statistically significant after correction, this indicates that the odds of selection differ significantly between these two pronouns, beyond their comparison to the baseline (he/him).

RQ$_2$ – Task Assignment. To investigate whether demographic and gender attributes influenced the type of task assigned by the LLM, we employed a **multinomial logistic regression** model [21]. In this case, selected *"Development/coding"* as the baseline category, as it was the most frequently assigned task in our dataset. Hence, all model coefficients were interpreted relative to this baseline.

Model outputs were reported as **Relative Risk Ratios (RRR)**, defined as the exponentiated coefficients, along with their corresponding **95% confidence intervals (CI)** and **p-values**. An RRR greater than 1 indicates that a given predictor increases the likelihood of assignment to a specific task (relative to *Development/coding*), while an RRR below 1 indicates a decreased likelihood. Effects were considered statistically significant at the $p < 0.05$ level. For example, an RRR of 2 for candidates from the UK would mean that they are twice as likely to be assigned to a given task compared to *Development/coding* (baseline). Conversely, an RRR of 0.5 would indicate their likelihood is halved relative to the baseline.

Since multinomial logistic regression inherently compares all outcomes to a single baseline, we conducted *post-hoc pairwise z-tests* to assess differences among the non-baseline categories of *country* and *pronouns*. This allowed us to evaluate whether, for example, candidates from *Brazil* were more likely than those from *India* to be assigned to a particular task, or whether candidates using *they/them* differed significantly from those using *she/her*. Similarly to **RQ$_1$**, to control for inflated Type I error due to multiple comparisons, we applied a *Bonferroni correction* [37].

4 Analysis of the Results

In this section, we present the results of the study, structured by research question. For reasons of space, only statistically significant findings are reported and discussed here, while the complete set of results is available in the online appendix [3].

4.1 RQ₁ — Selection Likelihood

Table 1 reports the results achieved in **RQ₁**. For each LLM, we analyzed the outcomes of the *logistic regression model* **relative to the baseline categories, namely profiles using *he/him* pronouns and located in the *US***. Second, to assess differences between non-baseline categories, we analyzed post-hoc pairwise comparisons among all pronouns and countries using *z-tests* on the logistic regression coefficients for each LLM, whose results are in Table 2.

GPT. The *logistic regression model* for GPT revealed significant associations. In particular, keywords related to communication platforms, web frameworks, databases, platforms, programming languages, and longer biographies increased the odds of selection.

Pronoun effects also emerged. Candidates using *they/them* (OR = 0.04) or *she/her* (OR = 0.79) had lower odds, while those using *any/all* pronouns showed higher odds (OR = 2.50). Country effects were significant as well, with *Nigerian* candidates showing increased odds of selection (OR = 1.46).

In addition, the *z-tests* revealed notable within-group differences. For pronouns, *he/they* ($z = 8.68$), *she/her* ($z = 11.62$), *she/they* ($z = 8.23$), *Any/All* ($z = 7.92$), and *any* ($z = 5.16$) all showed significantly higher odds of selection than *they/them*, with *Any/All* also exceeding *she/her* ($z = 2.46$). For country, *Nigerian* candidates had greater odds than those from the *UK* ($z = 6.91$), *Brazil* ($z = 4.82$), and *India* ($z = 4.36$), while *India* ($z = 2.54$) and *Brazil* ($z = 2.02$) also outperformed the *UK*.

DeepSeek. Also for DeepSeek, our analysis showed significant disparities. Mentions of web frameworks, tools, platforms, databases, AI models, communication platforms, and biography length platforms increased odds

For pronouns, candidates using *they/them* ($OR = 0.18$) and *she/her/they/them* ($OR = 0.07$) reduced odds of selection. By contrast, candidates from *Brazil* ($OR = 1.55$), *India* ($OR = 1.53$), and *Nigeria* ($OR = 1.55$) showed higher odds.

Post-hoc *z-tests* confirmed these gaps. *They/them* users were less likely to be selected than *he/they* ($z = 4.02$), *she/her* ($z = 9.81$), and *she/they* ($z = 3.70$). Likewise, *she/her/they/them* users were disadvantaged compared to *he/they* ($z = -2.25$), *she/her* ($z = -2.62$), and *she/they* ($z = -2.19$). For countries, the *UK* was significantly less likely to be selected than *Brazil* ($z = -4.15$), *India* ($z = -3.92$), and *Nigeria* ($z = -4.07$).

Claude. For Claude, we yielded similar results. Mentioning web frameworks, platforms, databases, AI models, communication platforms, and programming languages increased selection odds.

For pronouns, candidates using *they/them* showed decreased odds of selection ($OR = 0.04$) while *she/they* users were favored ($OR = 2.89$). Regarding country, candidates from *Brazil* ($OR = 1.43$) and *Nigeria* ($OR = 1.39$) increased odds of selection.

Post-hoc *z-tests* confirmed these differences. *They/them* users were less likely to be selected than *he/they* ($z = 7.78$), *she/her* ($z = 11.38$), and *she/they* ($z = 9.37$). Conversely, *she/they* users had higher odds than *he/they* ($z = -2.00$) and *she/her* ($z = -2.77$). For countries, candidates from *India* were less likely to be selected than those from *Brazil* ($z = -4.36$) and *Nigeria* ($z = -4.00$), while both *Brazil* ($z = 4.12$) and *Nigeria* ($z = 3.72$) outperformed the *UK*.

👥 RQ₁ – Selection Likelihood.

Across all three LLMs, selection was not neutral but systematically shaped by pronouns, country, and profile features, even after controlling for potential confounders. This suggests that sensitive attributes have a persistent and independent impact on selection outcomes, structurally biasing team composition processes. Together, our results confirm consistent cross-model disparities, such as the disadvantage of *they/them* and the relative advantage of Nigerian and Brazilian candidates.

4.2 RQ₂ — Task Assignment

The results related to **RQ₂** are presented in Table 3. For every LLM, we first examined the outcomes of the *multinomial logistic regression* **with *Development/coding* as the baseline category**, and then investigated differences across the non-baseline categories through post-hoc pairwise comparisons, applying *z-tests*. For the sake of space, the tables and raw results of the post-hoc analysis are reported in our online appendix [3].

GPT. The *multinomial logistic regression* analysis for GPT revealed significant effects involving confounding factors, pronouns, and country (Table 3). For *confounding factors*, positive predictors varied across task categories. Mentions of AI models, communication platforms, web frameworks, platforms, and company references increased assignment likelihood across multiple categories: *Clerical*, *Collaboration-Heavy*, *General Software*, *Software*, *Requirement-Related*, *Intellectual*, and *Organizational* tasks. In addition, operating systems were positively associated with *Collaboration-Heavy*, *General Software*, and *Information-Seeking* assignments. Education score favored *Information-Seeking* and *Supportive* tasks.

Concerning pronouns, candidates using *he/they* pronouns were significantly more likely to be assigned to *Clerical* tasks ($RRR = 30.1$). The use of *she/they* pronouns was strongly associated with a higher likelihood of assignment to *General Software* tasks ($RRR = 26.2$), whereas candidates using *she/her* pronouns were significantly less likely to be assigned to *Organizational* ($RRR = 0.66$) and *Requirement-Related* ($RRR = 0.46$) tasks.

The model also revealed that the country influenced task allocation. Candidates from *Brazil* were more likely to be assigned to a variety of tasks, including *Clerical* ($RRR = 4.01$), *Collaboration-heavy* ($RRR = 2.87$), *Intellectual* ($RRR = 8.03$), *Organizational* ($RRR = 2.93$), *Requirement-Related* ($RRR = 2.82$), and *Software* ($RRR = 4.41$). In contrast, candidates from *Nigeria* were significantly less likely to be assigned to *Clerical* ($RRR = 0.26$), *General Software* ($RRR = 0.31$), *Information-Seeking* ($RRR = 0.61$), *Organizational* ($RRR = 0.45$), and *Other* ($RRR = 0.31$) tasks. Finally, candidates from the *UK* showed an increased likelihood for *Intellectual* ($RRR = 11.49$), *Organizational* ($RRR = 2.09$), and *Software* ($RRR = 3.50$) tasks, while candidates from *India* were more likely to be assigned to *Organizational* tasks ($RRR = 1.89$) but less likely for *Other* tasks ($RRR = 0.29$).

Post-hoc *z-tests* revealed several significant combinations. For pronouns, the use of *she/her* was significantly associated with a higher likelihood of being assigned to *Collaboration-Heavy* ($z = 2.49$), *General Software* ($z = 2.21$), *Information-Seeking* ($z = 3.30$), *Organizational* ($z = 2.03$), *Software* ($z = 3.06$), and *Other* ($z = 2.77$) tasks compared to *Version Control*. In addition, candidates with

Table 1: RQ$_1$ – Statistically significant results relative to the baseline categories (*he/him* and *US*) for team composition. Arrows indicate the direction of the effect: ⊘ denotes a decrease in the likelihood of selection, whereas ⊕ denotes an increase.

Feature	GPT				DeepSeek				Claude			
	OR	95% CI low	95% CI high	Sig.	OR	95% CI low	95% CI high	Sig.	OR	95% CI low	95% CI high	Sig.
Bio_Education Score	0.96 ⊘	0.93	0.99	**	0.96 ⊘	0.93	0.99	*	1.02	0.99	1.05	
Bio_Experience Years	0.96 ⊘	0.95	0.98	***	0.95 ⊘	0.94	0.96	***	0.97 ⊘	0.95	0.98	***
Bio_Github Activity	0.84 ⊘	0.76	0.93	***	0.74	0.67	0.82	***	0.77 ⊘	0.70	0.85	***
Bio_AI Models	0.83 ⊘	0.73	0.94	**	1.40 ⊕	1.22	1.60	***	1.17 ⊕	1.04	1.32	**
Bio_Collaboration Tools	1.10	0.96	1.27		1.16 ⊕	1.00	1.34	*	1.06	0.93	1.21	
Bio_Communication Platforms	4.39 ⊕	3.54	5.45	***	1.67 ⊕	1.37	2.04	***	1.93 ⊕	1.60	2.32	***
Bio_Databases	1.54 ⊕	1.31	1.81	***	1.23 ⊕	1.06	1.43	**	1.16 ⊕	1.01	1.33	*
Bio_Development Environments	0.89	0.76	1.03		1.01	0.86	1.19		0.83 ⊘	0.71	0.96	*
Bio_Operating Systems	0.75 ⊘	0.67	0.84	***	0.74 ⊘	0.66	0.84	***	0.71 ⊘	0.63	0.79	***
Bio_Platforms	1.47 ⊕	1.33	1.63	***	1.26 ⊕	1.13	1.40	***	1.31 ⊕	1.19	1.44	***
Bio_Programming Languages	1.21 ⊕	1.13	1.29	***	1.06	0.99	1.13		1.22 ⊕	1.14	1.30	***
Bio_Web Frameworks	1.40 ⊕	1.30	1.51	***	1.37 ⊕	1.27	1.48	***	1.33 ⊕	1.24	1.43	***
Bio_Length	1.02 ⊕	1.01	1.04	**	1.02 ⊕	1.01	1.04	**	1.01	0.99	1.02	
Bio_Seniority Score	1.02	0.99	1.08		0.979	0.937	1.024		0.987	0.946	1.034	
Bio_Sentiment	0.75 ⊘	0.61	0.90	**	0.97	0.79	1.19		0.77 ⊘	0.64	0.93	**
Pronouns_Any_All	2.49 ⊕	1.00	6.23	*	0.36	0.09	1.34		72306	0.0	inf	
Pronouns_She/Her	0.79 ⊘	0.71	0.87	***	1.10	0.98	1.23		1.05	0.95	1.17	
Pronouns_She/Her/They/Them	2.51e7	0.0	inf		0.06 ⊘	0.009	0.54	*	3.17e-15	0.0	inf	
Pronouns_She/They	1.14	0.60	2.17		0.80	0.39	1.66		2.89 ⊕	1.42	5.87	**
Pronouns_They/Them	0.03 ⊘	0.02	0.06	***	0.17 ⊘	0.12	0.25	***	0.03 ⊘	0.02	0.06	***
Country_BR	1.02	0.88	1.17		1.54 ⊕	1.33	1.80	***	1.43 ⊕	1.24	1.64	***
Country_IN	1.06	0.92	1.22		1.52 ⊕	1.31	1.77	***	1.05	0.92	1.21	
Country_NG	1.46 ⊕	1.26	1.69	***	1.54 ⊕	1.32	1.80	***	1.39 ⊕	1.21	1.60	***

*Significance codes: ***: p<0.001; **: p<0.01; *: p<0.05; .: p<0.1*

Table 2: RQ$_1$ – Results of post-hoc pairwise comparisons among feature categories (excluding the baseline) for GPT, DeepSeek, and Claude. Columns <u>Task X</u> and <u>Task Y</u> show the compared categories; the others report log-odds difference, z-value, and significance level. Highlighted cells mark the category with higher likelihood of selection.

GPT					DeepSeek					Claude				
Task X	Task Y	Log Odds Diff	z	Sig.	Task X	Task Y	Log Odds Diff	z	Sig.	Task X	Task Y	Log Odds Diff	z	Sig.
Any/All	She/Her	1.15	2.46	*	She/Her/They/Them	He/They	-2.50	-2.25	*	He/They	She/They	-0.96	-2.00	*
Any/All	They/Them	4.23	7.91	***	She/Her/They/Them	She/Her	-2.77	-2.62	**	He/They	They/Them	3.36	7.78	***
Any	They/Them	3.83	5.16	***	She/Her/They/Them	She/They	-2.45	-2.19	*	She/Her	She/They	-1.00	-2.76	**
He/They	They/Them	3.62	8.68	***	He/They	They/Them	1.56	4.01	***	She/Her	They/Them	3.33	11.38	***
She/Her	They/Them	3.07	11.6	***	She/Her	They/Them	1.82	9.81	***	She/They	They/Them	4.33	9.36	***
She/They	They/Them	3.44	8.22	***	She/They	They/Them	1.51	3.69	***	-	-	-	-	
IN	NG	-0.32	-4.36	***	UK	BR	-0.31	-4.15	***	IN	BR	-0.30	-4.35	***
IN	UK	0.18	2.54	*	UK	IN	-0.29	-3.91	***	IN	NG	-0.27	-4.00	***
BR	NG	-0.35	-4.82	***	UK	NG	-0.30	-4.07	***	BR	UK	0.28	4.11	***
BR	UK	0.14	2.01	*	-	-	-	-	-	NG	UK	0.26	3.72	***
NG	UK	0.50	6.91	***	-	-	-	-	-	-	-	-	-	

*Significance codes: ***: p<0.001; **: p<0.01; *: p<0.05; .: p<0.1*

these pronouns were more likely to receive *Information-Seeking* ($z = 2.51$), *Software* ($z = -2.57$), and *Other* ($z = 2.15$) assignments relative to *Requirement-Related* tasks.

As for country, candidates from *Nigeria* were significantly more likely to be assigned to *Requirement-Related* ($z = -2.04$) and *Software* ($z = -3.00$) than to *Collaboration-Heavy* tasks. They were also more likely to be assigned to *Software* compared to *Organizational* ($z = -2.48$), *General Software* ($z = -2.45$), *Information-Seeking* ($z = -1.97$), *Version Control* ($z = 2.56$), and *Other* ($z = -1.97$). Furthermore, they were more likely to be assigned to *Requirement-Related* ($z = -1.97$) and *Software* ($z = -2.96$) compared to *Other*, and to *Collaboration-Heavy* relative to *Clerical* ($z = -2.33$). Candidates from *India* were significantly more likely to be assigned

to *Organizational* ($z = 3.65$), *Collaboration-Heavy* ($z = 2.56$), and *Information-Seeking* ($z = 2.44$) tasks, while being less likely to be assigned to *Version Control* ($z = -3.22$), *Requirement-Related* ($z = -2.80$), and *Software* ($z = -2.32$). Finally, candidates from the *UK* showed a significantly higher likelihood of being assigned to *Intellectual* tasks compared to *Requirement-Related* ($z = 2.60$), *Version Control* ($z = 2.44$), *Information-Seeking* ($z = -2.25$), *General Software* ($z = -2.00$), *Collaboration-Heavy* ($z = -2.30$), *Clerical* ($z = -2.01$), and *Other* ($z = 2.89$). They were also more likely to be assigned to *Organizational* tasks compared to *Requirement-Related* ($z = 2.19$) and *Other* ($z = 2.83$), and to *Software* compared to *Requirement-Related* ($z = -2.14$) and *Other* ($z = -2.58$).

DeepSeek. The *multinomial logistic regression* results for DeepSeek (Table 3) yielded similar results.

For *confounding factors*, DeepSeek showed consistent positive effects across tasks for AI models, platforms, education score, web frameworks, and communication/collaboration tools. These features increased assignment likelihood in categories such as *Clerical, General Software, Information-Seeking, Organizational, Requirement-Related, Software, Supportive,* and *Other.* In addition, GitHub activity promoted assignments in *Organizational, Requirement-Related,* and *Software* tasks, while development environments favored *Information-Seeking* and *Other* tasks.

With respect to pronoun groups, candidates using *she/her* were significantly less likely to be assigned to several software engineering task categories, including *Clerical* (*RRR* = 0.14), *Information-Seeking* (*RRR* = 0.21), *Organizational* (*RRR* = 0.19), *Requirement-Related* (*RRR* = 0.03), *Supportive* (*RRR* = 0.16), and *Other* (*RRR* = 0.51). In contrast, candidates using *they/them* were significantly more likely to be assigned to *Software* tasks (*RRR* = 48.0).

The model also revealed notable country-level effects. Candidates from *India* were less likely to be assigned to *Clerical* tasks (*RRR* = 0.20), but more likely to be placed in *Organizational* (*RRR* = 2.09) and *Software* (*RRR* = 7.00) categories. Similarly, candidates from the *UK* were less likely to be assigned to *Clerical* (*RRR* = 0.14), yet more likely to work on *General Software* (*RRR* = 14.5) and *Software* (*RRR* = 5.73). By contrast, candidates from *Nigeria* showed higher likelihoods of being assigned to *Information-Seeking* (*RRR* = 4.09) and *Software* (*RRR* = 12.0), and less likelihoods of being assigned to *Other* (*RRR* = 0.39).

Table 3: RQ$_2$ – Multinomial logistic regression results for task assignment by GPT, DeepSeek, and Claude. Arrows indicate the direction of the effect: ↓ denotes a decrease in the likelihood of selection, whereas ↑ denotes an increase.

Task	Feature	GPT RRR	GPT Sig.	DeepSeek RRR	DeepSeek Sig.	Claude RRR	Claude Sig.
Clerical	Bio_AI Models	2.49 ↑	*	13.9 ↑	***	0.24	
	Bio_Web Frameworks	1.32		2.30 ↑	*	0.47	
	Bio_Sentiment	5.52 ↑	***	46.0 ↑	***	2.29	
	Bio_Seniority Score	0.79		1.63 ↑	*	0.26	
	Bio_Development Environments	0.21 ↓	**	0.01 ↓	*	0.14	
	Bio_Programming Languages	0.51 ↓	**	0.35 ↓	*	0.06	
	Bio_Platforms	1.84 ↑	.	0.03 ↓	*	0.09	
	Bio_Length	0.91 ↓	*	0.82 ↓	**	1.25 ↑	*
	Bio_Company Mentions	0.73		0.00 ↓	**	0.27	
	Bio_Databases	0.06 ↓	*	0.14		0.51	
	Pronouns_He/They	30.1 ↑	*	0.92		49.2	
	Pronouns_She/Her	0.65		0.14 ↓	***	0.07 ↓	**
	Country_NG	0.26 ↓	**	0.56		1.81	
	Country_BR	4.01 ↑	***	2.75	.	3.81	
	Country_IN	0.90		0.20 ↓	*	0.17	
	Country_UK	1.54		0.14 ↓	**	0.36	
Collaboration-heavy	Bio_AI Models	1.77 ↑	*	6.20		0.60	
	Bio_Collaboration Tools	0.49 ↓	*	2.25		0.49	
	Bio_Communication Platforms	13.2 ↑	***	0.75		0.74	
	Bio_Databases	0.31 ↓	***	0.80		3.56	
	Bio_Development Environments	0.37 ↓	***	0.80		2.42	
	Bio_Operating Systems	2.71 ↑	***	1.99		0.47	
	Bio_Platforms	5.71 ↑	***	0.75		0.53	
	Bio_Programming Languages	0.73 ↓	**	0.50		1.13	
	Bio_Web Frameworks	2.14 ↑	***	1.24		0.61	
	Bio_Length	0.92 ↓	**	0.76		0.86	
	Bio_Company Mentions	1.97 ↑	**	0.66		0.62	
	Bio_Sentiment	0.59		0.64		1.78	
	Pronouns_She/Her	0.74		0.19		0.20	
	Country_BR	2.87 ↑	**	0.64		0.32	
General Software	Bio_AI Models	3.97 ↑	**	11.80 ↑	*	4.39 ↑	**
	Bio_Collaboration Tools	0.20 ↓	**	2.45		0.56	
	Bio_Communication Platforms	28.8 ↑	***	0.37		8.40 ↑	*
	Bio_Operating Systems	4.53 ↑	***	4.07		0.73	
	Bio_Programming Languages	0.44 ↓	***	0.16		1.14	
	Bio_Web Frameworks	2.68 ↑	***	1.13		0.66	
	Bio_Length	0.77 ↓	***	0.48 ↓	***	1.10	
	Bio_Education Score	1.04		3.31 ↑	**	1.09	
	Bio_Development Environments	0.61		0.54		0.46 ↓	*
	Bio_Platforms	1.96		7.01		4.53 ↑	**
	Bio_Github Activity	1.35		0.12		0.35 ↓	**
	Pronouns_She/Her	1.00		0.51		2.22 ↑	***
	Pronouns_She/They	26.2 ↑	*	0.98		2.26	
	Country_NG	0.31 ↓	*	12.25		0.57	
	Country_BR	1.97		0.20		3.35 ↑	*
	Country_UK	1.35		14.55 ↑	*	1.85	
Intellectual	Bio_AI Models	0.35		0.50		3.60 ↑	**
	Bio_Company Mentions	3.45 ↑	**	0.37		1.66	
	Bio_Platforms	5.10 ↑	***	0.65		3.15 ↑	*
	Bio_Communication Platforms	15.6 ↑	**	3.15		10.5 ↑	**
	Bio_Development Environments	0.05		0.62		0.46 ↓	*
	Bio_Operating Systems	1.03		0.70		0.38 ↓	*
	Bio_Sentiment	0.91		0.32		3.72 ↑	*
	Bio_Seniority Score	1.28		0.40		0.73 ↓	*
	Pronouns_She/Her	0.47		0.04		0.12 ↓	***
	Country_NG	0.38		0.27		0.28 ↓	**
	Country_UK	11.4 ↑	**	5.17		0.83	
	Country_BR	8.03 ↑	*	0.45		2.30	
Organizational	Bio_AI Models	2.58 ↑	***	2.14 ↑	*	2.82 ↑	*
	Bio_Platforms	2.59 ↑	***	14.0 ↑	****	3.56 ↑	**
	Bio_Communication Platforms	0.29 ↓	*	3.29 ↑	*	0.77	
	Bio_Development Environments	1.08		1.91		0.97	
	Bio_Operating Systems	0.36 ↓	***	0.59		0.17 ↓	***
	Bio_Programming Languages	0.69 ↓	***	1.29		0.82	
	Bio_Seniority Score	0.79 ↓	**	0.83 ↓	*	1.04	
	Bio_Collaboration Tools	4.65 ↑	***	2.14 ↑	*	5.82 ↑	***
	Bio_Databases	0.53 ↓	***	0.61		1.04	
	Bio_Company Mentions	2.04 ↑	***	1.00		1.12	
	Bio_Experience Years	1.54		0.92 ↓	*	0.99	
	Bio_Education Score	0.93		1.44 ↑	***	1.08	
	Bio_Github Activity	0.70		1.71 ↑	*	0.59 ↓	*
	Pronouns_She/Her	0.66 ↓	*	0.19 ↓	***	0.24 ↓	***
	Country_NG	0.45 ↓	**	1.16		0.46 ↓	*
	Country_BR	2.93 ↑	**	3.74 ↑	***	2.76 ↑	*
	Country_IN	1.89 ↑	*	2.09 ↑	*	3.17 ↑	*
	Country_UK	2.09 ↑	**	1.70		3.40 ↑	**

Task	Feature	GPT		DeepSeek		Claude	
		RRR	Sig.	RRR	Sig.	RRR	Sig.
Supportive	Bio_AI Models	0.35		1.97 ⊕	*	5.11	
	Bio_Programming Languages	1.67		1.46 ⊕	*	0.55	
	Bio_Platforms	2.31		12.6 ⊕	***	4.61	
	Bio_Communication Platforms	36.27		11.2 ⊕	***	0.80	
	Bio_Web Frameworks	1.46		1.95 ⊕	***	0.30	
	Bio_Databases	0.77		0.43 ⊕	**	0.78	
	Bio_Length	0.62 ⊕	*	0.93 ⊕	*	0.78	
	Bio_Experience Years	0.79		0.91 ⊕	***	0.84	
	Bio_Seniority Score	0.44		0.77 ⊕	***	0.76	
	Bio_Education Score	3.07 ⊕	*	1.28 ⊕	***	0.37	
	Pronouns_She/Her	0.14		0.16 ⊕	***	0.47	
	Country_BR	0.84		2.55 ⊕	**	0.75	
Information-seeking	Bio_AI Models	1.21		2.42 ⊕	*	2.36 ⊕	*
	Bio_Platforms	2.03		12.1 ⊕	***	2.31 ⊕	.
	Bio_Collaboration Tools	0.61		2.83 ⊕	*	1.80	
	Bio_Communication Platforms	2.03		0.03		2.27	
	Bio_Operating Systems	1.96 ⊕	*	0.30 ⊕	**	0.41 ⊕	**
	Bio_Education Score	1.20 ⊕	**	2.12 ⊕	***	1.16	
	Bio_Company Mentions	1.85 ⊕	*	1.32		1.47	
	Bio_Development Environments	0.16 ⊕	***	2.97 ⊕	*	0.20 ⊕	***
	Bio_Programming Languages	0.64 ⊕	***	0.53		0.69 ⊕	*
	Bio_Sentiment	0.27 ⊕	**	0.96		1.91	
	Bio_Seniority Score	0.79 ⊕	*	1.10		0.94	
	Pronouns_She/Her	0.92		0.21 ⊕	***	0.35 ⊕	***
	Country_BR	1.47		4.59 ⊕	**	1.87	
	Country_NG	0.61 ⊕	*	4.09 ⊕	**	0.50 ⊕	*
Requirement-related	Bio_AI Models	1.83 ⊕	*	1.83		3.04 ⊕	*
	Bio_Collaboration Tools	0.63		0.11 ⊕	***	0.59	
	Bio_Company Mentions	1.89 ⊕	*	1.09		1.41	
	Bio_Communication Platforms	6.94 ⊕	***	14.3 ⊕	**	16.0 ⊕	**
	Bio_Operating Systems	1.88 ⊕	*	4.64 ⊕	***	0.96	
	Bio_Platforms	2.27 ⊕	**	24.5 ⊕	****	5.31 ⊕	***
	Bio_Web Frameworks	1.82 ⊕	***	2.05 ⊕	*	1.19	
	Bio_Databases	0.55 ⊕	***	2.15		0.73	
	Bio_Development Environments	0.39 ⊕	***	0.19		0.20 ⊕	***
	Bio_Github Activity	1.31		2.78 ⊕	**	0.54 ⊕	*
	Bio_Sentiment	1.05		0.16 ⊕	*	2.89	
	Bio_Programming Languages	0.71 ⊕	**	0.04 ⊕	*	1.02	
	Bio_Length	0.91 ⊕	**	0.88 ⊕	*	1.07	
	Pronouns_She/Her	0.46 ⊕	***	0.03 ⊕	****	0.15 ⊕	***
	Country_BR	2.82 ⊕	**	2.60		3.15 ⊕	*
	Country_IN	1.24		1.19		2.67 ⊕	*
	Country_NG	0.70		0.45		0.87	
Software	Bio_AI Models	2.33 ⊕	*	6.81 ⊕	***	3.17 ⊕	*
	Bio_Collaboration Tools	0.86		1.04		4.12 ⊕	**
	Bio_Communication Platforms	11.3 ⊕	***	1.22		1.67	
	Bio_Databases	0.50 ⊕	*	0.21 ⊕	**	1.35	
	Bio_Development Environments	0.42		0.43		0.38	
	Bio_Operating Systems	1.04		0.99		0.27 ⊕	**
	Bio_Platforms	2.44 ⊕	*	5.66 ⊕	***	3.84 ⊕	**
	Bio_Programming Languages	0.67 ⊕	.	0.95		0.43 ⊕	**
	Bio_Web Frameworks	1.78 ⊕	*	0.66		1.11	
	Bio_Length	0.86 ⊕	**	0.70 ⊕	***	1.00	
	Bio_Education Score	0.87		1.28 ⊕	*	0.93	
	Bio_Company Mentions	1.69		0.53		2.32 ⊕	*
	Bio_Github Activity	0.92		3.60 ⊕	***	0.80	
	Pronouns_They/Them	0.86		48.06 ⊕	**	4.41	
	Pronouns_She/Her	1.20		0.82		0.30 ⊕	***
	Country_BR	4.41 ⊕	*	6.48 ⊕	**	9.99 ⊕	***
	Country_UK	3.50 ⊕	*	5.73 ⊕	**	3.57 ⊕	*
	Country_NG	2.02		12.0 ⊕	***	1.80	
	Country_IN	1.68		7.00 ⊕	**	3.36	
Other	Bio_AI Models	2.70 ⊕	**	3.01 ⊕	*	6.23 ⊕	***
	Bio_Length	0.97		1.00		1.11 ⊕	*
	Bio_Databases	0.70		0.01 ⊕	*	0.90	
	Bio_Web Frameworks	1.30		2.62 ⊕	***	1.06	
	Bio_Communication Platforms	2.69		7.16 ⊕	*	5.65 ⊕	*
	Bio_Platforms	1.17		2.85 ⊕	*	2.55 ⊕	*
	Bio_Development Environments	0.38 ⊕	**	3.57 ⊕	**	0.31 ⊕	**
	Bio_Collaboration Tools	0.42 ⊕	*	0.33 ⊕	*	0.50	.
	Bio_Programming Languages	0.57 ⊕	**	1.10		0.74	
	Bio_Github Activity	0.64		2.76 ⊕	**	0.52 ⊕	*
	Bio_Operating Systems	1.90		2.43 ⊕	*	1.10	
	Bio_Experience Years	1.64		0.22		0.95	
	Pronouns_She/Her	0.91		0.51 ⊕	*	0.42 ⊕	***
	Country_IN	0.29 ⊕	**	0.39 ⊕	*	1.09	
	Country_NG	0.31 ⊕	***	0.39 ⊕	*	0.24 ⊕	***
	Country_UK	0.64		0.83		1.10	

Significance codes: ***: p<0.001; **: p<0.01; *: p<0.05; .: p<0.1

Candidates from *Brazil* were more likely associated to *Supportive* (*RRR* = 2.55), *Information-Seeking* (*RRR* = 4.59), *Organizational* (*RRR* = 3.74), and *Software* (*RRR* = 6.48) tasks.

Pairwise *z-tests* revealed several significant contrasts between pronoun and country groups in task assignments. For candidates using *she/her* pronouns, assignments were significantly more likely in *Documentation* (*z* = 2.88), *General Software* (*z* = 2.32), *Information-Seeking* (*z* = 2.65), *Organizational* (*z* = 2.73), and *Other* (*z* = 4.05) compared to *Requirement-Related*. In contrast, they were less likely to be assigned to *Software* (*z* = −4.75) and *Supportive* (*z* = −2.46). Relative to *Clerical*, they showed a higher likelihood of being assigned to *Software* (*z* = −3.13), while compared to *Documentation*, *Information-Seeking*, *Organizational*, and *Supportive*, further significant differences also emerged. Finally, they were more likely to be assigned to *Other* than to *Documentation* (*z* = −2.05), *Clerical* (*z* = −2.30), and *Organizational* (*z* = −2.74). Candidates using *they/them* pronouns were more likely to be assigned to *Software* than to *Clerical* (*z* = 2.14), *Organizational* (*z* = −2.48), and *Supportive* (*z* = 3.28), and to *Documentation* rather than *Supportive* (*z* = 2.30). Regarding country effects, candidates from *India* were more likely to be assigned to *Organizational* (*z* = −2.98), *Software* (*z* = −3.52), and *Supportive* (*z* = −2.82) compared to *Clerical*. They also showed a stronger association with *Software* relative to *Documentation* (*z* = −2.32), *Information-Seeking* (*z* = −2.14), and *Other* (*z* = −3.27), as well as to *Organizational* (*z* = −2.71) and *Supportive* (*z* = −2.51) when contrasted with *Other*. For *UK* candidates, assignments were more likely in nearly all categories—*Documentation* (*z* = −2.82), *General Software* (*z* = −2.80), *Information-Seeking* (*z* = −3.15), *Organizational* (*z* = −3.40), *Other* (*z* = −2.33), *Requirement-Related* (*z* = −2.46), *Software* (*z* = −3.85), and *Supportive* (*z* = −2.34)—compared to *Clerical*. Additional contrasts showed that they were more likely to be assigned to *Software* against *Documentation* (*z* = −2.00), *Other* (*z* = −2.48), and *Supportive* (*z* = 2.74), and to *Organizational* compared to *Supportive* (*z* = 2.07).

Claude. The *multinomial logistic regression* on Claude outputs (Table 3) revealed significant effects for all variables.

For *confounding factors*, Claude consistently favored AI models, platforms, and communication tools, which increased assignment likelihood across categories such as *General Software*, *Requirement-Related*, *Intellectual*, *Software*, and *Other*. Collaboration tools and biography length also showed positive effects in selected categories, while biography sentiment promoted *Intellectual* tasks and company mentions promoted *Software*.

For pronouns, candidates using *she/her* were significantly less likely to be assigned to *Clerical* (*RRR* = 0.07), *Information-Seeking* (*RRR* = 0.35), *Intellectual* (*RRR* = 0.12), *Organizational* (*RRR* = 0.24), *Requirement-Related* (*RRR* = 0.15), *Software* (*RRR* = 0.30), and *Other* (*RRR* = 0.42) tasks. By contrast, they were more likely to be assigned to *General Software* tasks (*RRR* = 2.22).

The model revealed that the country also influenced task assignment. Candidates from *Brazil* were more likely to be assigned to *General Software* (*RRR* = 3.35), *Organizational* (*RRR* = 2.75), *Requirement-Related* (*RRR* = 3.15), and *Software* (*RRR* = 9.99). Likewise, candidates from *India* showed a higher likelihood of assignment to *Organizational* (*RRR* = 3.17) and *Requirement-Related* (*RRR* = 2.67). Those from the *UK* were also more likely to be placed

in *Organizational* ($RRR = 3.40$) and *Software* ($RRR = 3.57$) categories. Candidates from *Nigeria* were less likely to be assigned to *Information-Seeking* ($RRR = 0.50$), *Intellectual* ($RRR = 0.28$), *Organizational* ($RRR = 0.46$), and *Other* ($RRR = 0.24$) tasks.

Pairwise *z-tests* revealed several significant contrasts between pronoun and country categories. Candidates using *she/her* pronouns were significantly more likely to be assigned to *Communicative* ($z = 2.12$), *Documentation* ($z = 2.10$), *Information-Seeking* ($z = 2.59$), and *Other* ($z = -2.98$) tasks compared to *Intellectual*. They also showed a preference for *Information-Seeking* over *Requirement-Related* ($z = 2.25$), and for *Other* over both *Requirement-Related* ($z = 2.70$) and *Version Control* ($z = 2.08$, $p = 0.038$). Candidates using *he/they* pronouns were more likely to be assigned to *Clerical* rather than *Organizational* tasks ($z = 2.00$). Candidates from *Brazil* were more likely to be assigned to *Software* compared to *Information-Seeking* ($z = -2.08$) and *Other* ($z = -2.68$). Those from the *UK* were more likely to be placed in *Organizational* rather than *Intellectual* tasks ($z = -2.22$). For *Nigeria*, multiple contrasts were significant: candidates were more likely to be assigned to *Software* compared to *Version Control* ($z = 2.02$), *Organizational* ($z = -2.07$), *Intellectual* ($z = -2.72$), *Information-Seeking* ($z = -1.97$), and *Other* ($z = -3.07$). They were also more likely to be placed in *Requirement-Related* than in *Intellectual* ($z = -2.19$) and *Other* ($z = -2.68$), in *Clerical* compared to *Intellectual* ($z = 2.15$) and *Other* ($z = 2.39$), and in *Documentation* rather than *Other* ($z = 2.22$).

> **👥 RQ₂ – Task Assignment.**
>
> Task allocation was systematically shaped by pronouns, country, and profile content, with consistent trends and model-specific differences; these effects held even when accounting for confounders, showing that sensitive attributes are structurally embedded in LLM-driven allocation and raising fairness concerns. These results confirm that LLMs not only amplify demographic disparities but also reinforce stereotypes in how different groups are positioned across SE task categories.

5 Discussion and Implications

Our findings uncover a complex and varied landscape of bias in LLM-driven team composition and task allocation. We stress from the outset that the following discussion should be read as a set of reflections and hypotheses informed by our statistical analyses and contextualized through recent literature, rather than as definitive causal claims. Particularly, our analyses show that LLMs tend to operate in a biased manner, strongly reinforcing previous work [26, 42] by demonstrating that such biases persist even in the presence of contextual confounders. This has a general, critical implication: the *decision processes adopted by LLMs do not simply reflect isolated attributes, but are shaped by the interaction of demographic and task-related variables, indicating that bias is embedded in more structural patterns of their reasoning.* In this section, we discuss the main results and draw ↪ implications for researchers and practitioners.

On the interplay between pronoun and selection in a software team. Across all three models, candidates using *they/them* pronouns consistently faced reduced odds of selection, particularly for organizational and intellectual tasks. Similarly, *she/her* profiles were less likely to be recruited and had limited access to technical

roles, mirroring long-standing gendered patterns in the SE field. By contrast, candidates using more ambiguous forms such as *she/they* or *any/all* were sometimes favored, suggesting that LLMs may interpret inclusive or hybrid pronoun markers differently from explicit binary ones. Notably, these pronoun-related disparities persisted even when expertise was present in candidate bios, indicating that identity markers could outweigh substantive qualifications.

> ↪ LLM designers should integrate fairness tests across diverse pronoun forms, including non-binary and hybrid usage. Practitioners should exercise caution when relying on LLM recommendations in the selection, as pronoun signals may distort the perceived expertise. To mitigate this risk, the adoption of double-anonymity mechanisms might help separate candidate evaluation from pronoun-related biases.

On the impact of pronoun and nationality on task allocation. Bias was equally evident in task assignment. Profiles using *she/her* pronouns and candidates from Nigeria were disproportionately assigned to clerical, communicative, or supportive tasks, while Brazilian, Indian, or UK candidates were more frequently placed in technical or leadership roles. This dynamic possibly indicates that demographic attributes condition how the models interpret technical signals. For instance, Nigerian candidates with technical keywords were often selected into teams but then relegated to non-technical tasks, whereas Brazilian candidates with similar bios were channelled into technical roles. Similarly, *they/them* users were consistently excluded from leadership-oriented categories, regardless of the technical expertise included in their bios. These disparities suggest that equivalent credentials may not necessarily translate into equivalent opportunities across different groups.

> ↪ Future research should focus on fairness-aware task allocation benchmarks that capture cross-attribute interactions. In practice, audits of LLM-based systems should adopt a multi-objective perspective, verifying not only who is selected but also whether opportunities for advancement (e.g., technical or leadership roles) are equitably distributed across groups.

Broader lessons on LLM bias. Taken together, our results highlight three overarching lessons. First, single-attribute analyses are insufficient: biases become more visible when considering the joint effects of different data. Second, LLMs risk amplifying inequities in SE contexts: by overlooking substantive expertise and reinforcing stereotypes in task allocation, they may entrench barriers already documented in developer communities. Third, LLM neutrality cannot be assumed: we demonstrate that demographic markers systematically shape how technical credentials are valued, underscoring that fairness in LLMs is fundamentally socio-technical.

> ↪ For research, this requires interaction-based and multi-objective evaluation frameworks, accounting for multi-attribute interactions and socio-technical dynamics. For practice, organizations should combine LLM support with human oversight to identify when candidate expertise is undervalued due to demographic cues. For model providers, this highlights the importance of adopting mechanisms to address fairness concerns

and integrating fairness benchmarks into release pipelines to mitigate real-world risks before deployment.

6 Threats to Validity

Internal Validity. The main threats concern, on the one hand, the presence of confounding factors that have influenced the results. To mitigate this risk, we extracted a set of features from candidate biographies using the Stack Overflow Annual Developer Survey. We incorporated them into the regression models to observe the variation. On the other hand, the variability of the LLMs' outputs. To reduce this effect, we repeated all experiments multiple times across three different models increasing the robustness of our findings.

External validity. Threats to external validity regard the generalizability. First, our dataset was derived from GitHub profiles. While GitHub is one of the most used platforms for software development, it cannot be assumed to represent the global developer population. To address this limitation, we selected the most influential countries, as identified in GitHub's Octoverse 2024 report [13], to ensure representation across diverse regions. Second, the results depend on the choice of LLMs. Clearly, this cannot cover all existing systems; however, to mitigate this limitation, we conducted the study using three different large language models.

Construct validity. These threats concern the way we defined candidate expertise and demographic attributes. Specifically, developer competence was represented through proxies such as biography length, education, and experience indicators derived from GitHub profiles, which can only approximate the effects. Ambiguities or omissions could have influenced the results. To address this limitation, we used definitions and keyword lists from validated sources.

Conclusion validity. Threats to conclusion validity in our study concern the robustness of our statistical analyses and the risk of erroneous interpretation of the results. In particular, the presence of multiple comparisons increases the likelihood of Type I errors, while categories with limited representation may produce unstable or high-variance estimates. To address this, we employed logistic and multinomial regression models, verified model assumptions, and applied post-hoc tests with the Bonferroni correction.

7 Conclusions

This study examined fairness in LLM-driven team composition and task allocation, focusing on the joint effects of country and pronouns. Using three LLMs and 3,000 simulated decisions, we found systematic disparities: non-binary and female pronouns reduced selection odds, geographic patterns were inconsistent, and substantive expertise indicators were often undervalued compared to superficial keywords. These results demonstrate that bias arises not from isolated attributes but from their interplay with technical signals, underscoring the socio-technical nature of fairness challenges in LLMs and the necessity for multi-attribute evaluation and fairness-aware adoption in SE contexts.

Future work should extend this analysis to additional demographic attributes, larger and more diverse datasets, and real-world team composition scenarios. Moreover, comparative evaluations across newer LLMs and mitigation strategies are needed to design practical interventions that reduce bias in sensitive SE tasks.

References

[1] Wajdi Aljedaani, Furqan Rustam, Stephanie Ludi, Ali Ouni, and Mohamed Wiem Mkaouer. 2021. Learning Sentiment Analysis for Accessibility User Reviews. In 2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW). 239–246. doi:10.1109/ASEW52652.2021.00053

[2] Mina Arzaghi, Florian Carichon, and Golnoosh Farnadi. 2025. Understanding Intrinsic Socioeconomic Biases in Large Language Models. AAAI Press, 49–60.

[3] Anonymous Author(s). [n. d.]. Online Appendix. https://anonymous.4open. science/r/Once_Upon_a_Team/README.md

[4] Muneera Bano, Hashini Gunatilake, and Rashina Hoda. 2025. What does a software engineer look like? exploring societal stereotypes in llms. In 2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS). IEEE, 173–184.

[5] Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. arXiv preprint arXiv:1904.03035 (2019).

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356, 6334 (2017), 183–186.

[7] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2024. Fairness improvement with multiple protected attributes: How far are we?. In Proceedings of the IEEE/ACM 46th international conference on software engineering. 1–13.

[8] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. arXiv preprint arXiv:2405.01769 (2024).

[9] Shamse Tasnim Cynthia and Banani Roy. 2025. An Empirical Study on the Impact of Gender Diversity on Code Quality in AI Systems. In Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering (Clarion Hotel Trondheim, Trondheim, Norway) (FSE Companion '25). Association for Computing Machinery, New York, NY, USA, 1540–1549. doi:10.1145/3696630.3731674

[10] Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of data and analytics. Auerbach Publications, 296–299.

[11] Edson Dias, Paulo Meirelles, Fernando Castor, Igor Steinmacher, Igor Wiese, and Gustavo Pinto. 2021. What Makes a Great Maintainer of Open Source Projects?. In Proceedings of the 43rd International Conference on Software Engineering (Madrid, Spain) (ICSE '21). IEEE Press, 982–994. doi:10.1109/ICSE43902.2021.00093

[12] Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M. Zhang. 2023. Large Language Models for Software Engineering: Survey and Open Problems. In 2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE). 31–53. doi:10.1109/ICSE-FoSE59343.2023.00008

[13] GitHub. 2024. Octoverse 2024: AI leads Python to top language as the number of global developers surges. https://github.blog/news-insights/octoverse/octoverse-2024/. Accessed: 2025-09-12.

[14] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints (2023).

[15] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. Nature 633, 8028 (2024), 147–154.

[16] Max Hort, Jie M Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering. 994–1006.

[17] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2024. Large Language Models for Software Engineering: A Systematic Literature Review. ACM Trans. Softw. Eng. Methodol. 33, 8, Article 220 (Dec. 2024), 79 pages. doi:10.1145/3695988

[18] Thunyathon Jaruchotrattanasakul, Xin Yang, Erina Makihara, Kenji Fujiwara, and Hajimu Iida. 2016. Open Source Resume (OSR): A Visualization Tool for Presenting OSS Biographies of Developers. 2016 7th International Workshop on Empirical Software Engineering in Practice (IWESEP) (2016), 57–62. doi:10.1109/IWESEP.2016.17

[19] Tanjila Kanij, John Grundy, and Jennifer McIntosh. 2024. Enhancing understanding and addressing gender bias in it/se job advertisements. Journal of Systems and Software 217 (2024), 112169.

[20] Falaah Arif Khan, Nivedha Sivakumar, Yinong Oliver Wang, Katherine Metcalf, Cezanne Camacho, Barry-John Theobald, Luca Zappella, and Nicholas Apostoloff. 2025. Investigating Intersectional Bias in Large Language Models using Confidence Disparities in Coreference Resolution. In Second Conference on Language Modeling. https://openreview.net/forum?id=zOw2it5Ni6

[21] Chanyeong Kwak and Alan Clayton-Matthews. 2002. Multinomial logistic regression. Nursing research 51, 6 (2002), 404–410.

[22] Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. In Proceedings of the 29th International Conference on Computational Linguistics, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1221–1232. https://aclanthology.org/2022.coling-1.105/

[23] Zainab Masood, Rashina Hoda, Kelly Blincoe, and Daniela Damian. 2022. Like, dislike, or just do it? How developers approach software development tasks. Information and Software Technology 150 (2022), 106963.

[24] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) 54, 6 (2021), 1–35.

[25] Devon Myers, Rami Mohawesh, Venkata Ishwarya Chellaboina, Anantha Lakshmi Sathvik, Praveen Venkatesh, Yi-Hui Ho, Hanna Henshaw, Muna Alhawawreh, David Berdik, and Yaser Jararweh. 2023. Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. Cluster Computing 27, 1 (Nov. 2023), 1–26. doi:10.1007/s10586-023-04203-7

[26] Takashi Nakano, Kazumasa Shimari, Raula Gaikovina Kula, Christoph Treude, Marc Cheong, and Kenichi Matsumoto. 2024. 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, 624–629.

[27] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality 15, 2, Article 10 (June 2023), 21 pages. doi:10.1145/3597307

[28] OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[29] Pimolrat Ounsrimuang and Supakit Nootyaskool. 2017. Introducing scrum process optimization. In 2017 International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 1. 175–181. doi:10.1109/ICMLC.2017.8107761

[30] Alessandra Parziale, Gianmario Voria, Giammaria Giordano, Gemma Catolino, Gregorio Robles, and Fabio Palomba. 2025. Fairness on a budget, across the board: A cost-effective evaluation of fairness-aware practices across contexts, tasks, and sensitive attributes. Information and Software Technology 188 (2025), 107858. doi:10.1016/j.infsof.2025.107858

[31] Fabian C Peña and Steffen Herbold. 2025. Evaluating Large Language Models on Non-Code Software Engineering Tasks. arXiv preprint arXiv:2506.10833 (2025).

[32] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. 2002. An introduction to logistic regression analysis and reporting. The journal of educational research 96, 1 (2002), 3–14.

[33] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. ACM Computing Surveys (CSUR) 55, 3 (2022), 1–44.

[34] Paul Ralph, Sebastian Baltes, Domenico Bianculli, Yvonne Dittrich, Michael Felderer, Robert Feldt, Antonio Filieri, Carlo Alberto Furia, Daniel Graziotin, Pinjia He, Rashina Hoda, Natalia Juristo, Barbara A. Kitchenham, Romain Robbes, Daniel Méndez, Jefferson Seide Molléri, Diomidis Spinellis, Miroslaw Staron, Klaas-Jan Stol, Damian A. Tamburri, Marco Torchiano, Christoph Treude, Burak Turhan, and Sira Vegas. 2020. ACM SIGSOFT Empirical Standards. CoRR abs/2010.03525 (2020). arXiv:2010.03525 https://arxiv.org/abs/2010.03525

[35] Ayushi Rastogi, Nachiappan Nagappan, Georgios Gousios, and André van der Hoek. 2018. Relationship between geographical location and evaluation of developer contributions in github. In Proceedings of the 12th ACM/IEEE international symposium on empirical software engineering and measurement. 1–8.

[36] Román Salmerón, Catalina B García, and Jose García. 2018. Variance inflation factor and condition number in multiple linear regression. Journal of statistical computation and simulation 88, 12 (2018), 2365–2384.

[37] Philip Sedgwick. 2012. Multiple significance tests: the Bonferroni correction. Bmj 344 (2012).

[38] Mona Sloane. 2025. Boolean Clashes: Discretionary Decision Making in AI-Driven Recruiting. Commun. ACM 68, 5 (April 2025), 24–26. doi:10.1145/3708596

[39] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. Big Data & Society 9, 2 (2022), 20539517221115189.

[40] Igor Steinmacher, Tayana Conte, Marco Aurélio Gerosa, and David Redmiles. 2015. Social Barriers Faced by Newcomers Placing Their First Contribution in Open Source Software Projects. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (Vancouver, BC, Canada) (CSCW '15). Association for Computing Machinery, New York, NY, USA, 1379–1392. doi:10.1145/2675133.2675215

[41] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, Chris Parnin, and Jonathan Stallrich. 2017. Gender differences and bias in open source: Pull request acceptance of women versus men. PeerJ Computer Science 3 (05 2017), e111. doi:10.7717/peerj-cs.111

[42] Christoph Treude and Hideaki Hata. 2023. She elicits requirements and he tests: Software engineering gender bias in large language models. In 2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR). IEEE, 624–629.

[43] Jason Tsay, Laura Dabbish, and James Herbsleb. 2014. Influence of social and technical factors for evaluating contribution in GitHub. In Proceedings of the 36th International Conference on Software Engineering (Hyderabad, India) (ICSE 2014). Association for Computing Machinery, New York, NY, USA, 356–366. doi:10.1145/2568225.2568315

[44] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark G.J. van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. 2015. Gender and Tenure Diversity in GitHub Teams. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (Seoul, Republic of Korea) (CHI '15). Association for Computing Machinery, New York, NY, USA, 3789–3798. doi:10.1145/2702123.2702549

[45] Gianmario Voria, Gemma Catolino, and Fabio Palomba. 2024. Is attention all you need? Toward a conceptual model for social awareness in large language models. In Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering. 69–73.

[46] Gianmario Voria, Giulia Sellitto, Carmine Ferrara, Francesco Abate, Andrea De Lucia, Filomena Ferrucci, Gemma Catolino, and Fabio Palomba. 2025. Fairness-aware practices from developers' perspective: A survey. Information and Software Technology 182 (2025), 107710. doi:10.1016/j.infsof.2025.107710

[47] Yi Wang and David Redmiles. 2019. Implicit gender biases in professional software development: An empirical study. In 2019 IEEE/ACM 41st international conference on software engineering: Software engineering in society (ICSE-SEIS). IEEE, 1–10.

[48] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, Anders Wesslén, et al. 2012. Experimentation in software engineering. Vol. 236. Springer.

[49] Mengdi Zhang and Jun Sun. 2022. Adaptive fairness improvement based on causality analysis. In Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 6–17.

[50] Jianlong Zhou and Fang Chen. 2018. Human and Machine Learning. Springer.