Analysis of Bias in Recruitment Systems Utilizing Large Language Models

Nupur Deshmukh (223201702),
Sharayu Sunil Mhaske (223201244),
Lakshmi Soujanya Chandra (223201715), and
Jayani Rachapudi (223201113)

University of Koblenz

Supervisor:
Dr. Tai Le Quy
tailequy@uni-koblenz.de

Abstract. AI-powered recruitment systems are becoming increasingly common, offering efficiency in hiring processes. However, these systems may unintentionally perpetuate biases, leading to unfair hiring practices. This study examines gender, racial, and age biases in recruitment systems using Large Language Models (LLMs) like BERT, GPT-2, and GPT-Neo. The research analyzes hiring decision datasets using machine learning classification techniques and the Word Embedding Association Test (WEAT) to assess bias in AI-driven recruitment models. Findings reveal varying levels of bias across LLMs, with GPT-2 showing the highest bias scores in gender and age-related associations. BERT demonstrates more balance but still exhibits moderate biases, while GPT-Neo reflects significant racial bias.
The findings reveal significant biases in gender and racial associations, highlighting potential disparities in AI-driven hiring recommendations. This research underscores the need for bias mitigation strategies in AI-based recruitment tools, advocating for transparent and equitable hiring practices.

1 Introduction

Artificial intelligence (AI) is becoming increasingly prevalent in recruitment processes, offering efficiency and scalability in hiring decisions. However, the use of AI-driven hiring algorithms raises serious ethical concerns, particularly regarding bias and equity [1, 2, 25]. The integration of AI in hiring processes has shown promise in reducing human error, processing large volumes of applications quickly, and potentially mitigating unconscious biases. Nevertheless, this potential is threatened by the reality of biased AI systems, which can stem from historical data used to train AI models, perpetuating societal stereotypes and leading to unfair hiring practices [2, 26].

The implications of biased AI recruitment systems are profound. Qualified candidates from underrepresented groups may be systematically overlooked, perpetuating a lack of diversity within organizations and hindering overall innovation [1, 27]. These biases can reinforce existing inequalities in the job market and potentially violate anti-discrimination laws [27]. The increasing reliance on AI in recruitment demands immediate attention, as unchecked biases can exacerbate existing societal inequalities and undermine the principles of fair employment [2, 27].

Understanding and mitigating these biases is crucial for ensuring fair and equitable hiring practices in the age of AI. Recent studies have shown that AI systems can exhibit gender and racial biases, leading to discriminatory outcomes in hiring processes [26, 27]. For instance, research has demonstrated that some AI-powered resume screening tools may favor candidates with names typically associated with certain ethnic groups, potentially disadvantaging equally qualified applicants from other backgrounds [1, 2].

To address these challenges, this paper will explore several key aspects of bias in AI-driven recruitment models. We will examine the extent of bias in models using Large Language Models (LLMs) such as BERT, GPT-2, and GPT-Neo, which have shown promising results in natural language processing tasks but may also inherit biases present in their training data. The study will apply machine learning classification techniques and the Word Embedding Association Test (WEAT) to assess bias levels, building on previous work that has used these methods to quantify bias in AI systems.

Furthermore, we will investigate the impact of pre-processing techniques, including dataset balancing, on reducing bias in AI recruitment models. This approach has shown the potential to mitigate some forms of algorithmic bias in other domains. Finally, we will explore strategies for ethical AI adoption in hiring, including debiasing word embeddings, fairness-aware model training, and real-time bias monitoring.

By examining these aspects, we aim to contribute to the development of more equitable AI-driven recruitment practices. Our research emphasizes the importance of incorporating ethical guidelines to promote transparency and fairness in hiring processes, aligning with recent calls for responsible AI development in human resources. Through this comprehensive analysis, we hope to provide insights that will help organizations leverage the benefits of AI in recruitment while minimizing the risk of perpetuating or exacerbating existing biases.

2 Methodology

2.1 Dataset Overview

The dataset utilized for this study is obtained from Kaggle and consists of 10,000 job applicants, each represented by nine key features that encapsulate demographic information, job-related details, and textual descriptions [23]. These features enable an in-depth analysis of recruitment bias using Large Language Models (LLMs). The dataset is structured with both categorical and unstructured

data, including demographic attributes such as age, gender, race, and ethnicity, along with job-specific attributes like the job role applied for. Additionally, it contains textual data, including resumes and job descriptions, as well as an outcome variable labeled as Best Match, which indicates whether the candidate's resume aligns with the job description (0 = No Match, 1 = Match).

The age distribution of applicants ranges from 25 to 55 years, with an average age of approximately 40 years and a standard deviation of approximately 8.95 years. The dataset maintains a balanced gender distribution, consisting of 5,059 males (50.6 %) and 4,941 females (49.4%). The racial composition is evenly distributed, with approximately 33.6 % Mongoloid/Asian, 33.2% White/Caucasian, and 33.2 % Negroid/Black applicants.

The dataset spans 51 unique job roles, with notable concentrations in technical, managerial, and healthcare professions. The ten most frequent roles include Personal Trainer, Urban Planner, Biomedical Engineer, Construction Manager, Mechanical Engineer, Robotics Engineer, Operations Manager, Pilot, Machine Learning Engineer, and Web Developer. The distribution of roles suggests a diverse job applicant pool, ranging from engineering and AI positions to managerial and healthcare-related roles.

2.2 Data Preprocessing

Data preprocessing refers to the series of steps applied to raw data in order to clean and prepare it for analysis or modeling. It is an essential part of the machine learning pipeline because real-world data is often messy (e.g., missing values, duplicate entries), incomplete, or inconsistent, which can severely affect the performance of machine learning algorithms. The objective of data preprocessing is to transform the raw data into a form suitable for modeling, enabling algorithms to learn effectively [4].

In the context of our research on bias in recruitment systems, data preprocessing is essential to mitigate potential biases in the raw data. Recruitment datasets often contain biased information due to historical inequalities, underrepresentation of certain groups, or other structural biases that have been encoded into the data over time. Without proper preprocessing, these biases can propagate through machine learning models and lead to unfair outcomes, such as discrimination against specific groups (e.g., women, racial minorities, etc.).

2.3 Encoding

In the context of our research on bias in recruitment systems, data preprocessing is essential to mitigate potential biases in the raw data. Recruitment datasets often contain biased information due to historical inequalities, underrepresentation of certain groups, or other structural biases that have been encoded into the data over time. Without proper preprocessing, these biases can propagate through machine learning models and lead to unfair outcomes, such as discrimination against specific groups (e.g., women, racial minorities, etc.).

One-Hot Encoding One-hot encoding is a method that is used to convert categorical variables into a numerical form that is useful for machine learning models. This method will create binary columns for each unique category in the feature. Each of these binary column represent one category, meaning that a 1 in a binary column represents the presence of that category while 0 shows that it is not present [5]. For the purpose of this study, One-hot Encoding was applied to the Gender feature. Gender is considered nominal variables, or not in any order, so by applying One-hot encoding, it confirms that the machine learning model does not assume any kind of ranking or order of the categories of the gender. For example, The dataset contains the categories Male and Female then One-hot encoding will create separate binary columns, each of the columns is independent column.

Label Encoding Label Encoding is used to convert the categorical variables into numerical values in which each unique category is assigned a unique integer. This encoding method is useful for ordinal data or categorical variables where numerical relationships are meaningful or where the machine learning model can work with numerical labels [5]. In this study, Label Encoding was applied to the Race feature. Since the race categories in the dataset are nominal (without a meaningful order), the use of Label Encoding assigns an integer label to each race category. However, it is important to make a note that this does not imply any ordinal relationship between the categories. The "Race" feature was transformed into unique integer values such as 0, 1, 2, etc., each corresponding to unique racial category.

Frequency Encoding Frequency encoding is a technique for converting categorical features to numeric values by occurrence counts, or how often a given category appeared in the data. It corresponds frequency counts with a numeric value that indicates the frequency of a category, and allows for the detection of patterns based on category frequency [6]. Frequency encoding was employed in the Job Applicant Name, Ethnicity, and Job Roles features for this study. For example, the Job Applicant Name feature was frequency encoded to the occurrence frequency of that applicant's name in the data accordingly. Ethnicity and Job Roles were similarly frequency encoded using the observed frequency of that category in the data, allowing the model to identify trends correlated with ethnicity and/or job roles without implying ranking or ordering of the categories. The model is now able to learn the significance of the categories through frequency counts of Ethnicity and Job Roles while the nominal character of the variables remain intact.

2.4 Skewness Transformation

Skewness occurs when numerical variables in a dataset are not symmetrically distributed which eventually leads to an imbalance in how machine learning models interpret these features. A skewness value of 0 indicates symmetrical

distribution whereas positive or negative values suggest right or left skewness respectively. Skewed data can impact model performance, particularly in algorithms that have normality assumptions, such as a linear regression model [7]. For this reason, we computed the skewness for each numerical feature in the data to identify those with a statistically significant skew. Based on this, we found two features that were skewed and required transformation: Job Applicant Name (skewness: 0.1426, positive skew) and Job Roles (skewness: -0.1150, negative skew).

To normalize the skewness, we employed the Power Transformer with the Yeo-Johnson transformation. The Yeo-Johnson transformation is a power transformation that can work with both negatives and positives while other methods (ex: Box-Cox transformation) require all values to be strictly positive. This transformation first altered the features and made the distribution's less skewed, approaching normality, which can improve the performance of models prone to non-normal distribution of the independent features [8]. Upon completing the transformation, the skewness of Job Applicant Name (skewness: 0.0360) and Job Roles (skewness: 0.042) had decreased in skewness indicating both features were better normalized.

2.5 Feature Scaling

Scaling is performed on numerical values to standardise them into a common range while making sure that the features with large magnitudes do not dominate model training. Since numerical features can vary greatly in scale so algorithms—especially those relying on gradient-based optimization—may struggle with unscaled data. To address the issue of unscaled data, we have implemented RobustScaler technnique.

In contrast to the standard scaling techniques, such as Min-Max Scaling or Standard Scaling, we have used RobustScaler that is highly effective in reducing the effects of outliers. It scales the features based on the median and interquartile range (IQR) rather than using mean and standard deviation, hence making it more robust in cases with extreme values [9]. By implementing the RobustScaler each of the numerical features was transformed ensuring consistency across the variables to successfully improve model stability and reliability of predictions.

2.6 Supervised Learning Models

This study employed supervised classification to explore recruitment bias and predict job-resume matches. Supervised learning refers to training a model on labeled data. Labeled data means that the dataset contains labeled input-output pairs. The main goal of supervised learning is for the model to learn the input-output relationship and predict correct outputs for unseen data [10].

In this study, the Logistic Regression and Random Forest Classifier are the common machine learning algorithms that were used to predict job-resume

matches. The dataset has several features that represent the job applicant characteristics (e.g., age, gender, job role). The target variable, "Best Match," is a binary classification label indicating if a resume matches a job posting.

Logistic Regression This is a linear model that is classically used for binary classification problems, making it a good fit to predict job-resume matches. Logistic Regression gives us interpretability because it associates coefficients to each variable which allow researchers to learn how each attribute of the applicants (age, gender, etc.) is contributing to the decision of best match. The interpretability is useful when we need to explain about model predictions and/or when we want to know how the decision is being formed [11].

Random Forest Classifier It is an ensemble learning method which constructs multiple decision trees and combines their predictions to improve accuracy and robustness. Based on this, Random Forest can capture non-linear relationships in the features (compared with Logistic Regression), allowing it to detect non-linear patterns in the data. In addition, Random Forest gives us insight into the importance of each feature related to the classification decision and how big of an influence it has on the job-resume match decision [12].

2.7 Data Visualization and Bias Analysis

Data Visualization is important for spotting patterns, relationships, and distributions in your data, as well as checking how well the model is performing. In this study, we employed a correlation heatmap to examine relationships between numerical features and categorical variables after encoding [13].

A correlation heatmap is a graphical representation of the correlation matrix which shows the relationships between multiple variables in a dataset. It uses different colors to represent the values in the matrix which will make easier to identify patterns and correlations among variables. The heatmap is mainly useful when we are dealing with a large number of features as it allows a quick visualization of how each feature is correlated with others [15].

The heatmap was created using Seaborn.heatmap() and results in highlighting the correlations among features such as Age, Gender, Race, Ethnicity, and Job Roles with our target variable Best Match [14]. This information allowed us to examine potential multicollinearity and biases affecting performance.

2.8 Word Embedding Association Test (WEAT)

Word embeddings are vector representations of words in multi-dimensional space. Words with similar meanings are located closer to each other in this space. Consider the words 'king' and 'man'. In a word embedding space, these words would be represented as vectors (lists of numbers). Because 'king' and 'man' have related meanings (both are male humans), their vectors would be located relatively

close to each other in that space. On the other hand, the words 'king' and 'apple' have very different meanings. Therefore, their vectors would be located far apart in the embedding space. This spatial distance reflects the difference in their semantic meaning [16].

Word Embedding Association Test (WEAT) is a method used to measure bias in word embeddings. WEAT is used to measure potential biases in three different language models: BERT, GPT-2, and GPT-Neo [17, 18] This analysis aimed to examine whether these models exhibit biases related to gender, race, and age when processing job-related text, such as resumes and job descriptions. The WEAT method works by computing the cosine similarity between different sets of words. Cosine similarity measures how closely related two vectors are in the embedding space. If two words have high cosine similarity, they are semantically related in the model's understanding [19]. The key idea is to compare the similarity of target words with attribute words. This is done by implementing the WEAT Effect size.

WEAT Effect Size (d) :
Consider

- Target Sets (X,Y): The concepts you're testing for association.
  Example: X =e.g., programmer, engineer, scientist, ... Y = nurse, teacher, librarian, ...,
- Attribute Sets (A, B): The groups you're comparing.
  Example: A = Male-associated words, B = Female-associated words.

s(w, A, B) is the association score of a word w with the attribute sets A and B. We take the average of these scores for all words in X and subtract the average for all words in Y. The standard deviation measures how much the association scores vary across all words in X and Y. This ensures that the effect size is standardized, so it's not influenced by the scale of the data [24].

$$d = \frac{\text{mean}_{x \in X}\, s(x, A, B) - \text{mean}_{y \in Y}\, s(y, A, B)}{\text{std dev}_{w \in X \cup Y}\, s(w, A, B)} \tag{1}$$

Here, d $> 0 \rightarrow$ More association between X and A.
If, d $< 0 \rightarrow$ More association between X and B.
The range of this lies from -2 to 2
Here, the Association score for a word: Measures how much a word w is associated with A vs B.

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

2.9 WEAT Methodology

Before we start the bias detection it is important to establish: two sets of target words and two sets of attribute words. From the inspection of the dataset, we

decided to extract the 2 targets STEM and NON-STEM as they better represent the clear difference in the unique job roles we have. The implementation involved multiple steps, including data preprocessing, keyword extraction, word embedding extraction, and bias quantification using WEAT effect sizes. The same process is applied across multiple transformer models, with embeddings being generated separately for each model while keeping the rest of the workflow consistent. BERT, GPT-2, and GPT-Neo [17, 18] are used for embedding generation. Below is a detailed breakdown of the methodology and results.

2.10 Data Acquisition and Preprocessing

The dataset containing resumes and job descriptions is loaded using pandas [21]. The text data underwent several preprocessing steps to ensure consistency and reduce noise so that it can be further used for the generation of separate sets of words.

- All text was converted to lowercase to standardize the data.
- Regular Expression Cleaning: Non-alphabetic characters and special symbols were removed using regular expressions, retaining only alphabetical characters and spaces.
- Tokenization and Lemmatization: The text was tokenized into individual words and stop words (common English words) were removed using NLTK's stopwords corpus [20].
- Lemmatization was performed to reduce words to their base form using NLTK's WordNetLemmatizer [20].
- Combined Text Data: The preprocessed resumes and job descriptions were combined into a single list for subsequent feature extraction.

2.11 Keyword Extraction

To identify relevant keywords for our target sets, TF-IDF (Term Frequency-Inverse Document Frequency) was initially used to extract a list of top keywords from the preprocessed text [22]. These keywords were further refined using cosine similarity with pre-defined STEM and non-STEM reference word lists. TF-IDF was applied to the combined text data to identify the most important words. A stop list was used to filter out common words that do not provide meaningful classification.

2.12 Keyword Classification Using Transformer-Based Embeddings and Cosine Similarity

The extracted keywords are classified into STEM and non-STEM categories based on their semantic similarity to predefined reference word lists.
stem_reference = ["algorithm", "neural", "processor", "dataset", "AI", "ML", "deep", "network", "programming", "software", "hardware", "databases", "cybersecurity", "engineering", "science", "physics", "mathematics", "robotics",

"biotechnology", "statistics", "machine learning", "data science", "cloud computing", "computer vision", "natural language processing", "big data", "automation", "genomics", "bioinformatics", "quantum computing", "electrical engineering"]
non_stem_reference = ["management", "leadership", "marketing", "sales", "communication", "teamwork", "customer", "business", "strategy", "collaboration", "writing", "counseling", "creativity", "organization", "planning", "support", "human resources", "social work", "public relations", "event planning", "journalism", "advertising", "psychology", "education", "training", "hospitality", "tourism", "real estate"]

2.13 Model-Specific Embeddings

For each model (BERT, GPT-Neo, GPT-2), the corresponding tokenizer and model are loaded. Embeddings for the keywords are generated using the model.[17, 18] This project utilizes the Hugging Face Transformers library for this task. The library enables dynamic model loading, tokenization, and embedding generation, ensuring a streamlined and scalable workflow.[22]

2.14 Cosine Similarity Measurement

The cosine similarity between the model-specific embeddings of the extracted keywords and the model-specific embeddings of the reference words was calculated.

$$\text{similarity} = 1 - \text{cosine\_distance}(\text{word\_embedding}, \text{reference\_embedding})$$

2.15 Category Assignment

Keywords were assigned to the category STEM or NON-STEM with the highest average cosine similarity, above a similarity threshold (0.3). The similarity threshold of 0.3 was chosen for better filtering of words and to avoid noise. Each word is then assigned to either STEM or non-STEM based on the higher similarity score, forming two final lists: stem_keywords and non_stem_keywords. Because the Embeddings are model specific, the STEM and non-STEM keyword lists will vary from model to model. These sets were preserved for subsequent bias detection using WEAT. To optimize computational efficiency, similarity calculations and classification are performed in batches.Further the embeddings are generated for the respect constraints chosen for quantifying the bias in the chosen Large Language Models.

3 Results and Discussions

3.1 Machine Learning Model Performance

Logistic Regression The Logistic Regression model achieved an accuracy of 62.40% in predicting job-resume matches, indicating moderate performance.

However, it struggles to capture complex relationships between features, which is a significant limitation.

The model's classification performance is balanced but suboptimal, with similar precision and recall values for both match and non-match classes. This suggests that while the model is fair, it doesn't excel at identifying either matches or non-matches accurately.

Overall, while Logistic Regression provides a useful baseline and interpretability, more advanced models may be needed to improve predictive performance and capture the complexity of job-resume matching.

Random Forest Classifier The Random Forest Classifier demonstrated superior performance in job-resume matching, achieving an accuracy of 70.93%. This represents a significant improvement over the Logistic Regression model. The Random Forest model showed an enhanced ability to distinguish job matches, with improved recall and precision. Its capacity to capture complex relationships between features likely contributed to this performance boost. Overall, the Random Forest Classifier proved to be a more effective tool for predicting job-resume matches in this context.

3.2 Correlation Analysis

Understanding the relationships between applicant features and the Best Match outcome is crucial for identifying potential biases in hiring. A correlation heatmap (Figure 1) was generated to analyze how demographic and identifying factors influence selection decisions. This analysis helps uncover systematic preferences or biases that may affect fair hiring practices.

The correlation heatmap (Figure 1) reveals notable patterns that suggest potential biases in the job selection process. There is a clear correlation between gender and the Best Match outcome, with a negative correlation for females (-0.26) and a positive correlation for males (0.26). This pattern indicates a tendency to favor male candidates in the selection process, warranting further investigation. Additionally, a strong negative correlation (-0.56) between Job Applicant Name and Race suggests the possibility of racial bias, potentially arising from name-based assumptions influencing hiring decisions.

Meanwhile, Age and Ethnicity exhibit weak correlations with the Best Match outcome, implying that they may not play a significant role in selection. However, the presence of more subtle or indirect biases cannot be ruled out. These findings highlight the need for further investigation and potential interventions to ensure fairness and eliminate discriminatory recruitment practices.

3.3 WEAT Analysis Results

To assess bias in different models, the Word Embedding Association Test (WEAT) was conducted, focusing on gender, racial, and age biases within the context of STEM and non-STEM fields. Each bias category is analyzed across three models, with individual visualizations illustrating the results for each model.

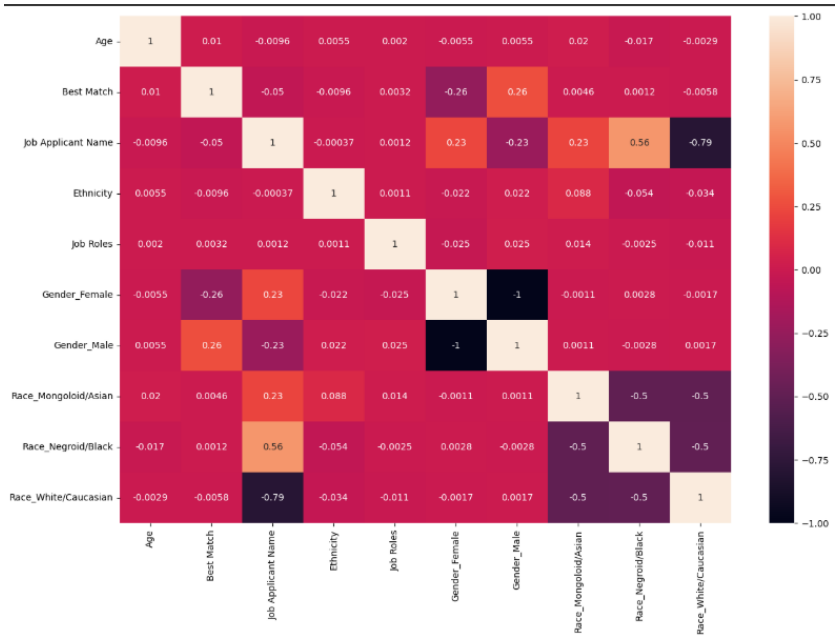

Fig. 1. Correlation HeatMap

Gender Bias The visualizations in Figures 2–5 illustrate gender bias detected using the WEAT analysis across three language models: BERT, GPT-2, and GPT-Neo. The WEAT effect sizes for each model (Table 1) indicate how strongly STEM and non-STEM fields are associated with male and female names.

A negative WEAT effect size for "STEM vs. Male" suggests that STEM fields are less associated with male names, while a positive effect size for "STEM vs. Female" indicates that STEM fields are more associated with female names. Interestingly, all three models exhibit this reversed bias, contradicting the common societal stereotype that STEM is more male-associated.

- BERT (Figure 2) demonstrates the strongest gender bias, associating STEM more with female names and non-STEM with male names.
- GPT-2 (Figure 3) and GPT-Neo (Figure 4) show similar trends, but GPT-Neo has the weakest bias among the three models.
- Figure 5 provides a comparative visualization of all three models, highlighting variations in gender bias across them.

For non-STEM fields, the pattern is reversed: non-STEM fields are less associated with male names and more associated with female names. These findings confirm that while biases exist, they do not align with traditional societal stereotypes, warranting further analysis of how these models internalize and reflect gender associations.

The following table shows WEAT effect sizes for gender bias:

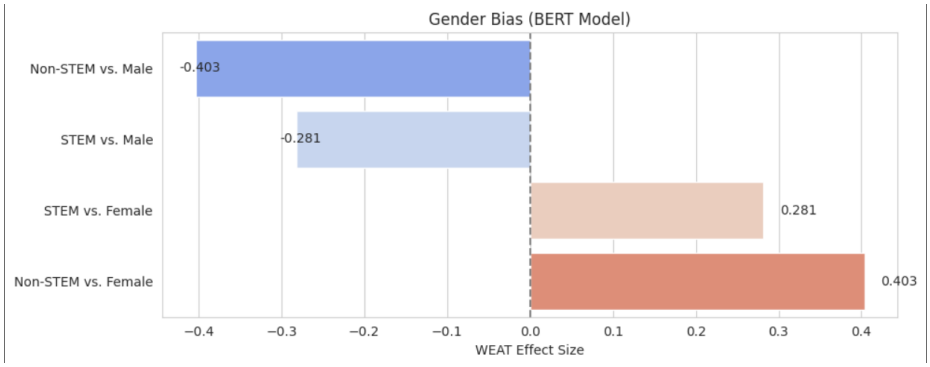

Fig. 2. Gender Bias BERT Model

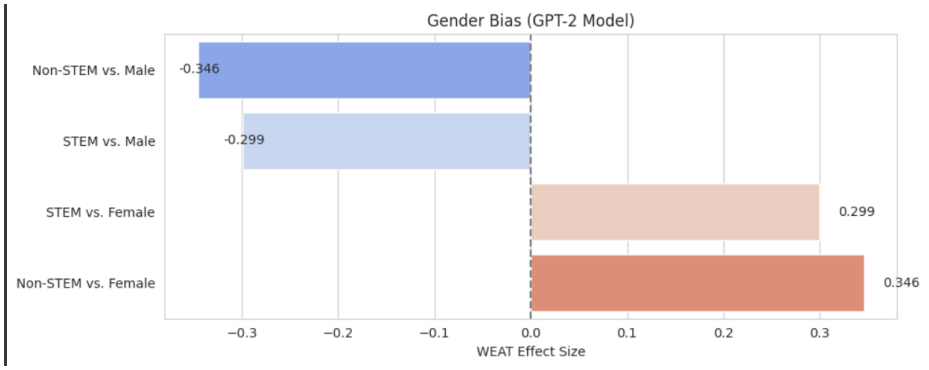

Fig. 3. Gender Bias GPT-2 Model

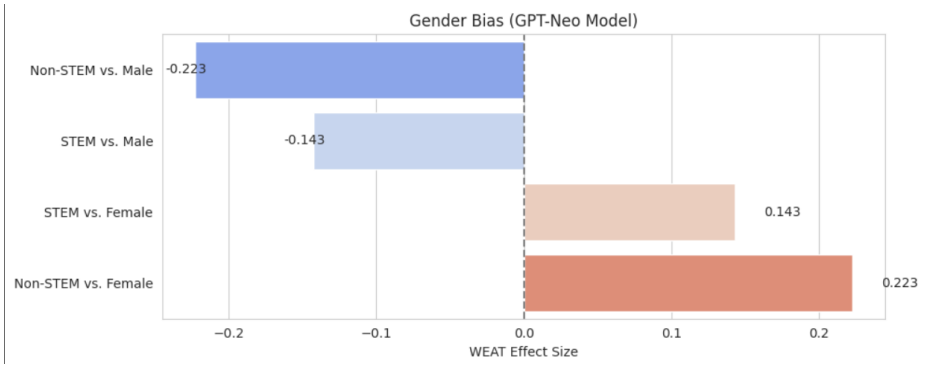

Fig. 4. Gender Bias GPT-NEO Model

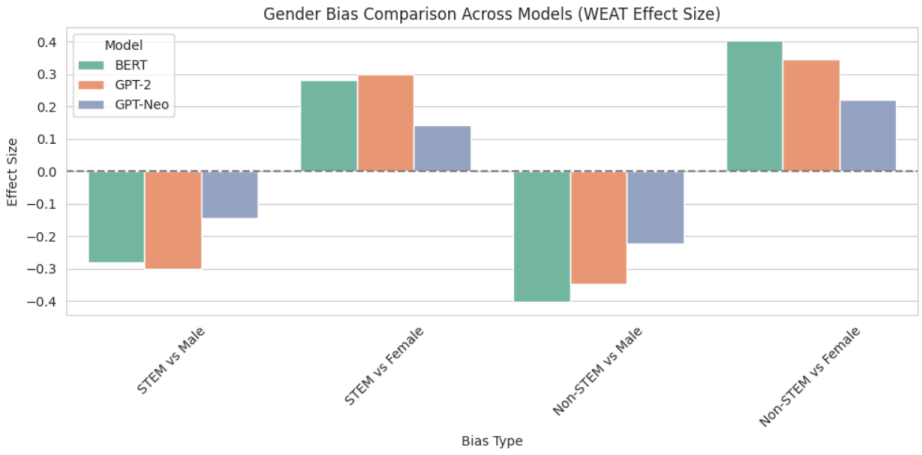

Fig. 5. Gender Bias Comparison Across Models

Model	STEM vs Male	STEM vs Female	Non-STEM vs Male	Non-STEM vs Female
BERT	0.2813	0.2813	-0.4034	0.4034
GPT NEO	-0.1427	0.1427	-0.2225	0.2225
GPT 2	-0.2995	0.2995	-0.3457	0.3457

Table 1. Model Comparison of Gender Bias

First Name Bias (Race Associations) The visualizations in Figures 6–9 illustrate the first name bias detected across the BERT, GPT-2, and GPT-Neo models using WEAT analysis. This analysis examines how strongly different models associate White, Black, and Asian names with the general concept of a first name. The corresponding WEAT effect sizes for each model are presented in Table 2.

- BERT (Figure 6) exhibits a strong positive association between White names and general first names, indicating a bias where White names are perceived as more prototypical first names compared to Black or Asian names.
- GPT-2 (Figure 8) and GPT-Neo (Figure 7) display a negative association between White names and general names, suggesting a bias where Black and Asian names are more strongly associated with the concept of a first name.
- Across all models, the bias between Black and Asian names is relatively low, meaning these two groups are treated more similarly compared to their association with White names.
- Figure 9 provides a comparative visualization of all three models, highlighting the differences in how they encode name-related biases.

The WEAT effect sizes in Table 2 further support these observations, quantifying the extent of bias in each model. These findings suggest that language models encode varying biases in name associations, which may have significant impli-

cations for applications involving name-based predictions, identity recognition, and automated decision-making systems.

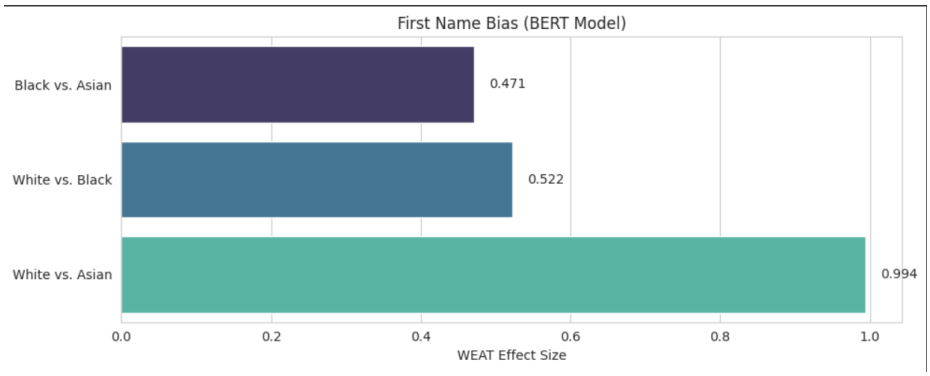

Fig. 6. First Name Bias BERT Model

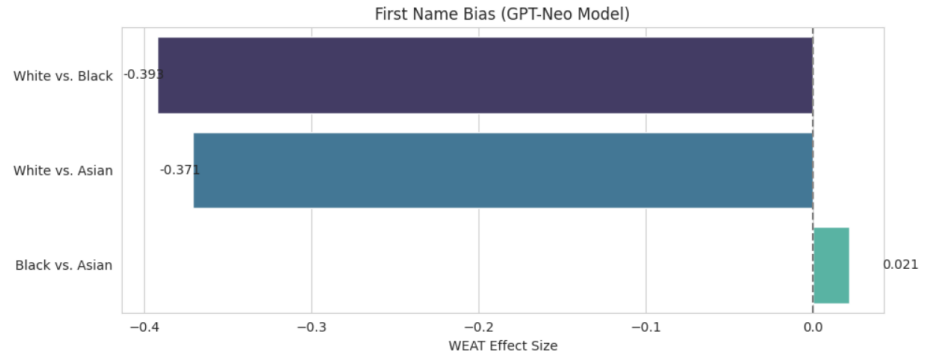

Fig. 7. First Name Bias GPT-Neo Model

Model	(White vs. Black)	(White vs. Asian)	(Black vs. Asian)
BERT	0.5225	0.9938	0.4713
GPT NEO	-0.3927	-0.3713	0.0214
GPT 2	-0.2905	-0.3111	-0.0205

Table 2. First Name Bias Comparison

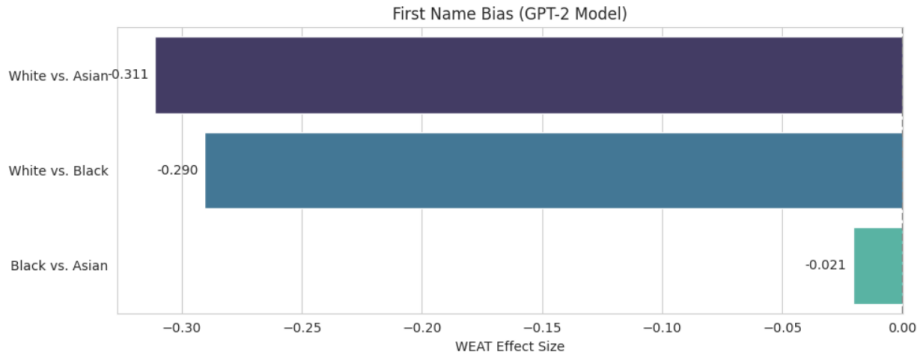

Fig. 8. First Name Bias GPT-2 Model

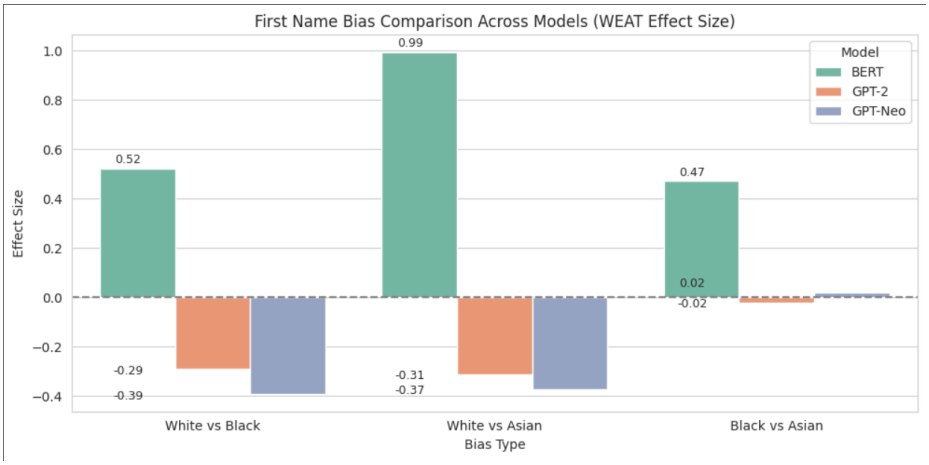

Fig. 9. First Name Bias Comparison Across Models

Race Bias: STEM/Non-STEM Associations The visualizations in Figures 10–13 illustrate the racial biases in STEM associations detected across BERT, GPT-2, and GPT-Neo using WEAT analysis. This analysis measures how strongly each model links White, Black, and Asian names with STEM-related terms. The corresponding WEAT effect sizes for each model are presented in Table 3.

- BERT (Figure 10) exhibits a strong positive association between White names and STEM, reinforcing the stereotype that White individuals are more closely linked to STEM careers compared to Black or Asian individuals.
- GPT-2 (Figure 12) and GPT-Neo (Figure 11) display a negative association between White names and STEM, instead showing a bias where Black and Asian names are more strongly associated with STEM fields—though the effect sizes are smaller than BERT's White-STEM bias.
- Across all models, the bias between Black and Asian names in STEM associations is relatively low, indicating that these groups are treated more similarly compared to their association with White names.
- Figure 13 provides a comparative visualization of all three models, highlighting the stark contrast in how autoencoding (BERT) and autoregressive (GPT-2, GPT-Neo) architectures encode racial biases in STEM associations.

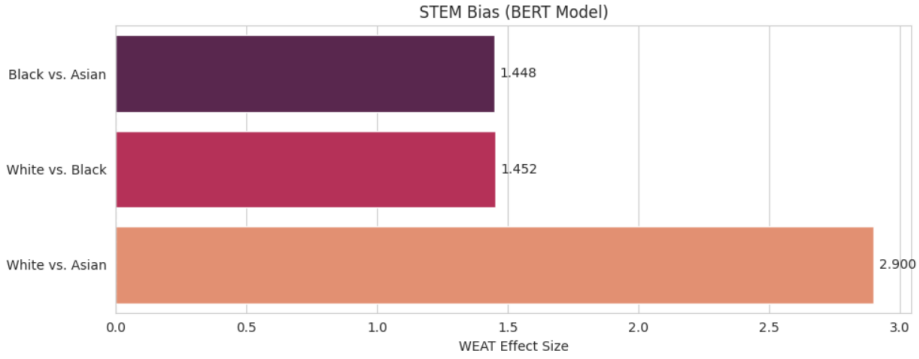

Fig. 10. STEM Bias BERT Model

Best Match Bias (STEM vs. Non-STEM Within Groups) The visualizations in Figures 14–17 present the "best match" bias analysis, examining how strongly BERT, GPT-2, and GPT-Neo associate White, Black, and Asian names with being the "best match" for STEM fields compared to non-STEM fields. The corresponding WEAT effect sizes for each model are reported in Table 4.

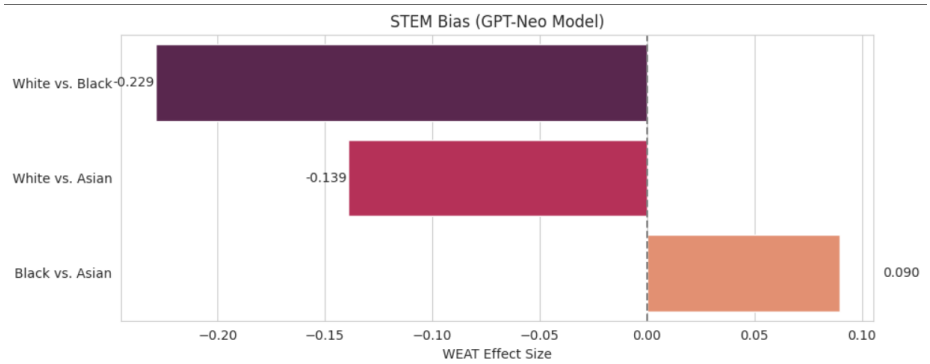

Fig. 11. STEM Bias GPT-Neo Model

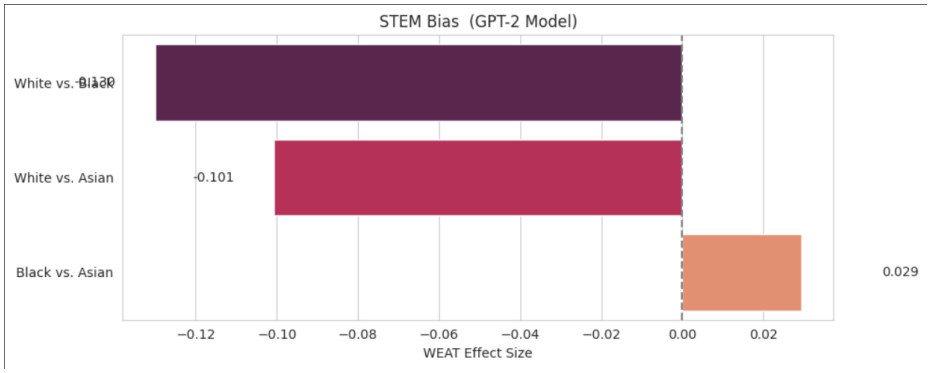

Fig. 12. STEM Bias GPT-2 Model

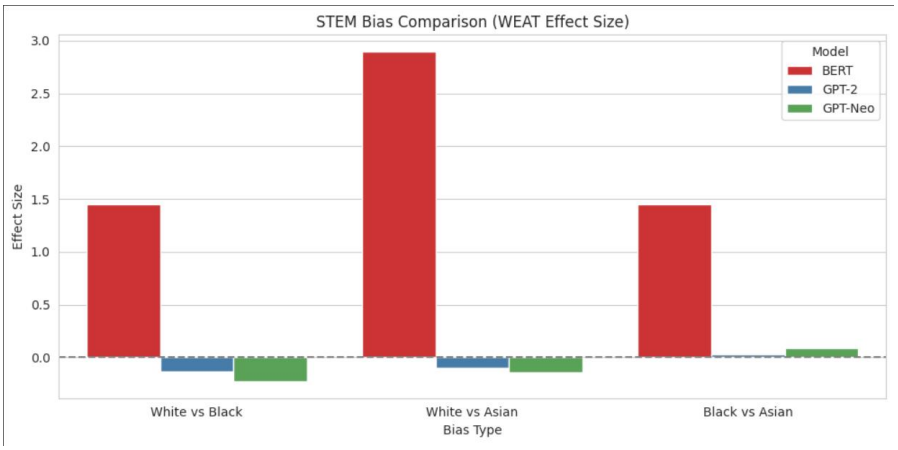

Fig. 13. STEM Bias Comparison

Model	(White vs. Black)	(White vs. Asian)	(Black vs. Asian)
BERT	1.4517	2.9000	1.4483
GPT NEO	-0.2288	-0.1390	0.0897
GPT 2	-0.1300	-0.1007	0.0293

Table 3. STEM Bias Across Different Ethnic Groups

- BERT (Figure 14) shows a strong positive bias, associating White individuals as the "best match" for STEM careers, reinforcing a pro-White STEM preference.
- GPT-2 (Figure 16) and GPT-Neo (Figure 15) exhibit a negative bias, instead suggesting that Black and Asian individuals are more strongly linked as the "best match" for STEM—though with smaller effect sizes than BERT's White-STEM association.
- The bias strength varies significantly between models, with BERT encoding the strongest racial bias, while GPT-2 and GPT-Neo show weaker but inverse tendencies.
- Figure 17 provides a comparative analysis, illustrating how the three models differ in their "best match" STEM associations across racial groups.

The WEAT scores in Table 4 demonstrate BERT's statistically significant, large-magnitude pro-White STEM bias, while GPT-2 and GPT-Neo exhibit measur-

able but weaker anti-White (pro-Black/Asian) biases, consistent with their overall tendency toward less pronounced racial bias encoding.

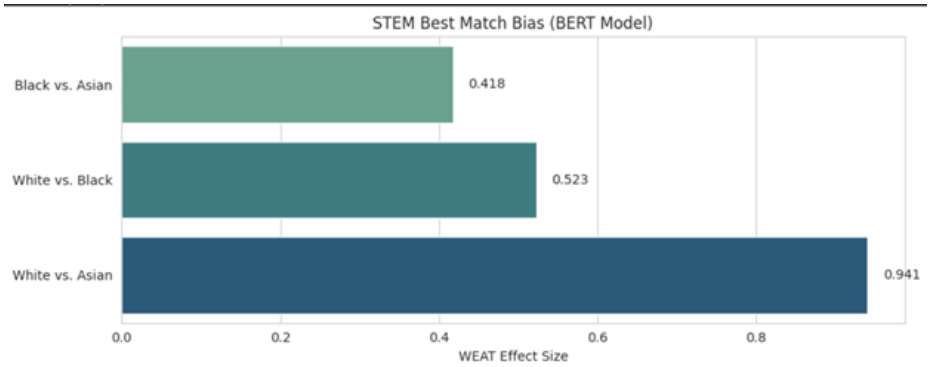

Fig. 14. STEM Best Match Bias BERT Model

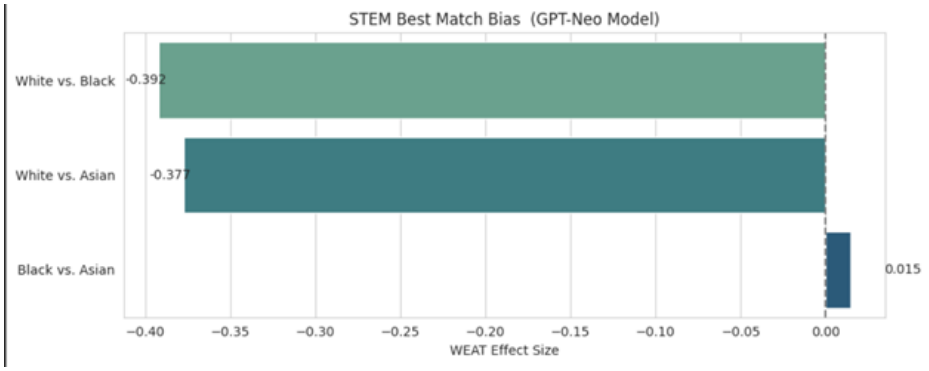

Fig. 15. STEM Best Match Bias GPT-Neo Model

Model	(White vs. Black)	(White vs. Asian)	(Black vs. Asian)
BERT	0.5230	0.9406	0.4176
GPT NEO	-0.3918	-0.3771	0.0146
GPT 2	-0.2406	-0.2518	-0.0112

Table 4. STEM Best Match Bias Across Different Ethnic Groups

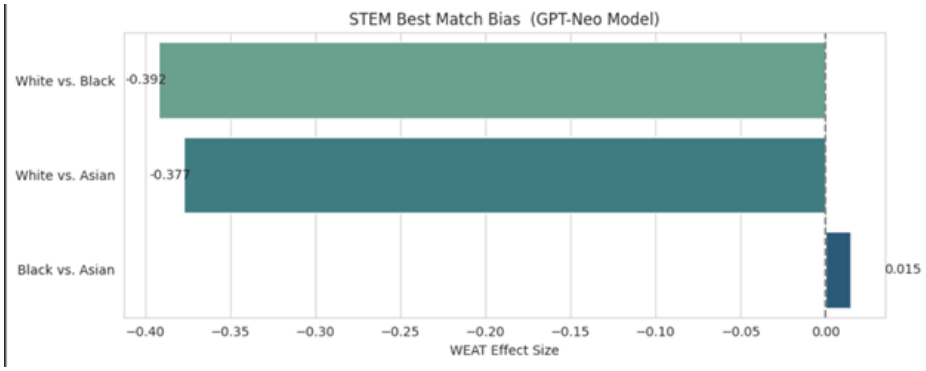

Fig. 16. STEM Best Match Bias GPT-2 Model

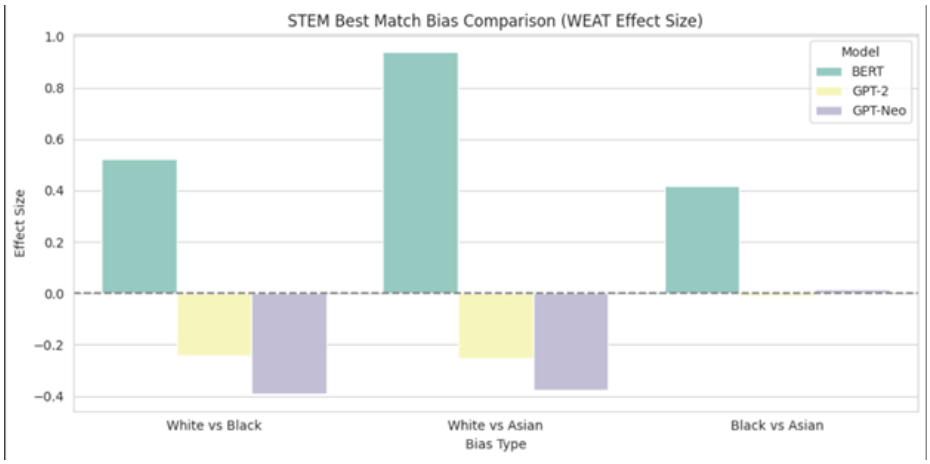

Fig. 17. STEM Best Match Bias Comparison

Age Bias: STEM/Non-STEM Associations Our analysis of age bias in STEM/non-STEM associations reveals minimal systematic bias across all models, as shown in Figure 18. The WEAT effect sizes (Table 5) are consistently small, indicating that:

- None of the models exhibit strong age-related biases in STEM associations.
- The observed effects are substantially weaker than the gender and racial biases identified in previous analyses.
- The patterns remain consistent across BERT, GPT-2, and GPT-Neo architectures.

Figure 18 visually confirms this absence of strong age bias, with all models showing near-neutral associations between age groups and STEM fields. The WEAT scores in Table 5 quantitatively support this finding, with effect sizes an order of magnitude smaller than those observed for racial and gender biases.

This suggests that age stereotypes regarding STEM/non-STEM careers are either:

- Less prominent in the training data.
- Not as strongly learned by the models.
- More evenly distributed across age categories.

Compared to other forms of demographic bias. The consistency across architectures implies this may be a general characteristic of current language models rather than an architecture-specific phenomenon.

Model	Age Bias (STEM vs. Non-STEM)
BERT	0.0102
GPT NEO	0.0507
GPT 2	-0.0244

Table 5. Age Bias in STEM vs. Non-STEM Fields

3.4 Discussion

Comparing the three models, BERT exhibits the most substantial biases across most categories, particularly in racial biases related to name associations and STEM fields, displaying a prominent pro-White tendency. This is likely due to BERT's training on a vast corpus of historical data, which inherently reflects and amplifies existing societal biases prevalent during those periods.

In contrast, GPT NEO and GPT 2 generally show weaker biases and, in some cases, reverse the direction of biases observed in BERT, indicating a pro-Black/Asian association. This reversal could stem from several factors: potentially, these autoregressive Analysis of Bias in Recruitment Systems Utilizing Large Language Models 21 models might have been trained on more recent,

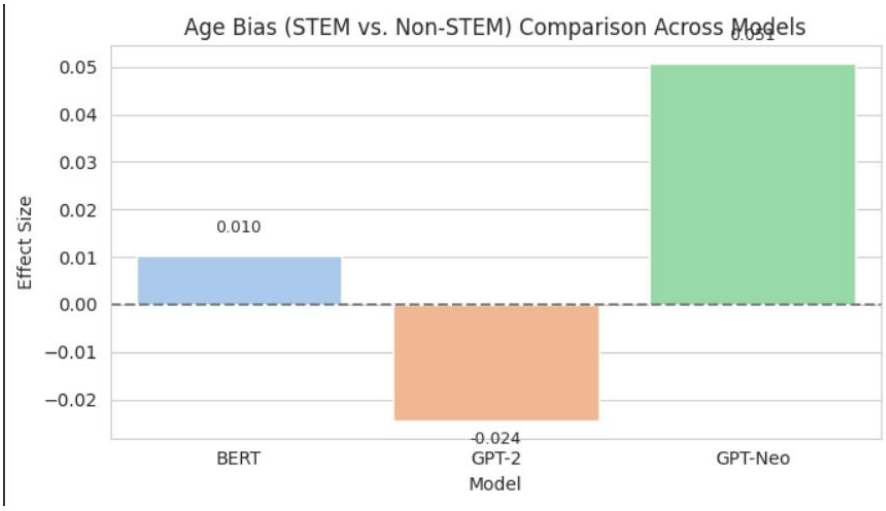

Fig. 18. Age Bias STEM vs Non-STEM comparison across all models

diverse datasets that reflect evolving societal norms, or they might inherently process and contextualize information differently, leading to varied bias representations.

All three models show a reverse gender bias, associating STEM with female names, which is a surprising result that warrants further investigation. Additionally, all models demonstrate minimal age bias. These variations highlight the differences in how these models learn and represent societal biases from their training data, emphasizing the necessity for careful examination and mitigation of biases in large language models.

4 Conclusion

This study has investigated bias in recruitment systems utilizing large language models, focusing on BERT, GPT-2, and GPT-Neo. Our findings reveal several important conclusions:

Persistent Biases: All models exhibit some form of bias related to gender and race, although the direction and magnitude vary across models. This confirms that bias remains an intrinsic challenge in AI-driven recruitment systems.

Architectural Differences: The distinct bias patterns observed between BERT and GPT models suggest that model architecture influences how biases are encoded and expressed. BERT consistently shows stronger biases, particularly racial biases with a pro-White tendency

Reversed Gender Stereotypes: The unexpected association of STEM fields with female names across all models challenges traditional narratives about gender bias in AI systems but still represents a form of bias that could influence recruitment decisions

Minimal Age Bias: All models demonstrate negligible age-related biases in STEM/non-STEM associations, suggesting that age discrimination may manifest differently or be less encoded in language patterns.

Correlation with Traditional ML The biases identified through WEAT align with patterns observed in traditional machine learning models, confirming that bias exists at multiple levels in recruitment systems.

These findings underscore the need for comprehensive bias detection and mitigation strategies in AI-driven recruitment tools. While LLMs offer powerful capabilities for processing and analyzing job applications, their deployment must be accompanied by rigorous fairness assessments and interventions to ensure equitable outcomes for all demographic groups.

5 Future Scope

Building on this research, several promising directions for future work emerge:

Intersectional Bias Analysis Investigate how biases interact across multiple demographic dimensions (e.g., race and gender combined) to produce compounded effects for certain groups.

Bias Mitigation Techniques : Develop and evaluate methods for reducing bias in recruitment LLMs, such as adversarial debiasing, counterfactual data augmentation, or fine-tuning with balanced datasets.

Longitudinal Studies : Track bias in recruitment systems over time as models evolve and training data changes, providing insights into the trajectory of AI fairness.

Contextual Analysis: Explore how different job domains and industries may exhibit varying patterns of bias, moving beyond the STEM/non-STEM dichotomy.

Human-AI Collaboration: Study how human recruiters interact with AI recommendations and whether they amplify or mitigate algorithmic biases through their decision-making.

Regulatory Frameworks: Develop guidelines and standards for assessing and certifying the fairness of AI recruitment tools, contributing to responsible AI deployment.

5.1 Alternative Fairness Metrics:

Explore metrics beyond WEAT for quantifying bias in recruitment systems, including outcome-based measures and causal frameworks.

Explainable AI for Bias Detection: Create tools that make bias in recruitment systems transparent and interpretable to non-technical stakeholders, fostering accountability

By pursuing these research directions, we can work toward recruitment systems that harness the power of large language models while promoting fairness, diversity, and inclusion in hiring practices.

References

1. Vultus Inc. (2023, December 8). Ethical considerations of AI-driven recruitment: A detailed guide. LinkedIn. https://www.linkedin.com/pulse/ethical-considerations-ai-driven-recruitment-detailed-guide-q56ze/
2. Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanities and Social Sciences Communications, 10(1).
3. Deady, D. (2024, July 29). The Ethical Considerations of AI in Hiring. SocialTalent. https://www.socialtalent.com/blog/recruiting/the-ethical-considerations-of-ai-in-hiring
4. Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
5. Brownlee, J. (2020, August 17). Ordinal and one-hot encodings for categorical data. Machine Learning Mastery. Retrieved March 20, 2025, from https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/
6. Neural Ninja. (2023, June 12). Frequency encoding: Counting categories for representation. Let's Data Science. Retrieved March 20, 2025, from https://letsdatascience.com/frequency-encoding/
7. Tuychiev, B. (2023, December 6). Understanding skewness and kurtosis and how to plot them. DataCamp. Retrieved March 20, 2025, from https://www.datacamp.com/tutorial/understanding-skewness-and-kurtosis
8. Kim, E. (2019, October 7). Box-Cox vs Yeo-Johnson. Cross Validated. Retrieved March 20, 2025, from https://stats.stackexchange.com/questions/430419/box-cox-vs-yeo-johnson

9. Scikit-Learn Developers. (n.d.). sklearn. preprocessing.RobustScaler. Scikit-Learn. Retrieved March 20, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html
10. Mohanty, Ananya, et al. "AI in Recruitment: A Review of Supervised Machine Learning Techniques." Journal of Artificial Intelligence Research, vol. 70, 2021, pp. 1-25.
11. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
12. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
13. Tufte, E. R. (2001). The Visual Display of Quantitative Information (2nd ed.). Graphics Press.
14. Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
15. Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. The American Statistician, 56(4), 316-324.
16. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.
17. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
19. Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35-43.
20. Loper, E., & Bird, S. (n.d.). NLTK: Natural Language Toolkit. Retrieved March 20, 2025, from https://www.nltk.org/
21. The Pandas Development Team. (n.d.). Pandas documentation. Retrieved March 20, 2025, from https://pandas.pydata.org/
22. Wolf, T., & others. (n.d.). Transformers: State-of-the-art Natural Language Processing. Hugging Face. Retrieved March 20, 2025, from https://huggingface.co/transformers/
23. Surendra, S. (2020). Recruitment dataset. Kaggle. https://www.kaggle.com/datasets/surendra365/recruitement-dataset
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need (arXiv:1706.03762). arXiv. Retrieved March 20, 2025, from ttps://arxiv.org/pdf/1608.07187
25. Beatty, D., Masanthia, K., Kaphol, T., & Sethi, N. (2024). Revealing Hidden Bias in AI: Lessons from Large Language Models. arXiv preprint arXiv:2410.16927.
26. Payne, B. (2024, January 24). Breaking barriers: Tackling bias in AI recruitment for fair and inclusive hiring. Gem. https://www.gem.com/blog/ai-recruitment-bias
27. Seppälä, P., & Małecka, M. (2024). AI and discriminative decisions in recruitment: Challenging the core assumptions. Big Data & Society, 11(1). https://doi.org/10.1177/20539517241235872 (Original work published 2024)