MSc Artificial Intelligence
Master Thesis

AI-Based Hiring and the Appeal of Novelty: Do LLMs Solve or Exacerbate the Problem of Discrimination?

by
Alexia Mureșan
12871192

June 30, 2024

48 EC
1/11/2023 - 30/06/2024

Supervisor:
Leonard Bereska, MSc

Examiner:
Prof. Efstratios Gavves

Second reader:
Leonard Bereska, MSc

Universiteit van Amsterdam

Contents

Abstract

As Artificial Intelligence (AI) becomes increasingly integrated in Human Resource processes such as hiring, it is essential to consider the social and ethical ramifications of automating impactful decisions. While efforts are being made in this direction, the AI landscape is evolving with the expansion of Large Language Models (LLMs), raising new questions and inciting the need to rethink how we address bias and discrimination in AI. While research on this topic is steadily increasing, a side-by-side comparison of LLMs with more traditional AI models is lacking. This thesis aims to address that knowledge gap to determine whether the 'trendy' transition to LLMs in AI-based hiring is worth it in terms of fairness and resources. To this end, four traditional models and two LLMs are evaluated for discriminatory bias in three scenarios, which differ in the data used to train and test the models. The three datasets included in the scenarios are synthetically generated to respond to the needs of this study. They will be made publicly available for future research in this domain. The discriminatory biases evaluated here are based on two different sensitive attributes: gender and ethnicity. To make this comparison more comprehensive, five well-known fairness metrics are used. The results show that the models do not possess any inherent bias, emphasizing the importance of training and deploying models responsibly. Furthermore, the strong impact of biased training data highlights the need for more balanced datasets, along with bias avoidance and mitigation methods. Overall, the expected contrast between LLMs and more traditional models is not observed. However, GPT-3.5 Turbo distinguishes itself through its high robustness to bias, proving itself to be a viable improvement to the process of AI-based hiring in terms of fairness, but also performance. The relevant code for these experiments, along with the synthetic datasets, can be found at https://github.com/alexiamuresan/AI-Master-Thesis-Fairness-in-AI-Based-Hiring.

0.1 Foreword: Ethics of AI: A Complex Interdisciplinary Matter

Like in other scientific fields, innovation in AI can have a significant social impact, which is simultaneously promising and worrying. The field of AI is so vast and fast-paced that this effect is amplified. This explains why AI is studied and discussed from all perspectives; beyond sciences, it poses an interest to the fields of law, ethics, philosophy, psychology, cognitive neuroscience, and more.

The application of AI in hiring merges scientific, social, legal, and ethical ramifications. Therefore, bias and fairness in this context cannot be studied through the lens of one single discipline, and attempting to do so would be limiting and fruitless. Such complex social notions cannot be represented numerically, through a formula or metric, in a way that accounts for all its subtleties, and oversimplification can lead to scientifically interesting but socially irrelevant results. This is an AI master thesis and therefore focuses on the AI viewpoint of this wider problem. However, it attempts to provide sufficient interdisciplinary context to put the results and limitations of this study into perspective, connecting the scientific findings with conclusions relevant to the 'real world' regarding AI-based hiring.

Chapter 1

Introduction

As Artificial Intelligence (AI) is increasingly dominant and relied upon in most professional fields, it is crucial to understand and control its impact and potential social ramifications. One prime example is the already vast use of AI in automated hiring practices. It has been found that over 66% of companies rely on automated recruitment methods [107]. This number is higher for Fortune 500 companies, of which 97% use AI-based systems to manage various parts of the applicant selection and hiring process [84]. This highlights that AI is entrusted with high-stakes decisions that can significantly impact people's professional careers and, as a result, their livelihoods. Therefore, ensuring that these systems are fair and free of discrimination is crucial, which, unfortunately, is not the status quo.

One of the first widely publicized cases of AI-related discrimination in hiring was Amazon's failed attempt at using an autonomous, AI-based hiring system for new recruits [26, 115]. The system was found to exhibit significant gender bias in its predictions, automatically downgrading any applications that included the word 'female' in them [110]. This was due to their training data, which consisted of resumes of Amazon's employees, mostly white men [72]. This brings forward the importance of the data used to train these models, which is often a root cause of discrimination through AI [80]. In the aftermath of the Amazon scandal, other similar situations followed, showing that AI systems learn and perpetuate social biases (gender bias, religious bias, racial bias...) that are prevalent in our society [80]. It is now common knowledge that AI can discriminate and that there is reason to be skeptical of the flashy and novel AI hiring techniques, no matter how efficient. This questions the trustworthiness of those methods and gives rise to ethical, legal, and social dilemmas.

As a result, various ways to evaluate fairness and diminish discrimination in automated hiring have been established, ranging from dataset debiasing methods [5] and model interpretability techniques [62], to creating regulations for prevention and accountability purposes. One example is the General Data Protection Regulation (GDPR) [43], in effect since 2016 and applicable to the European Union (EU). The more recent (2024) AI Act [44], applicable within the same jurisdiction, is more relevant to this thesis since it specifically targets the ethical and socially responsible use of AI. So, although measures are taken against discrimination through AI, the problem persists and existing solutions are limited in terms of their realistic applicability.

In parallel, Large Language Models (LLMs) are increasing in popularity and performance. They are starting to be used for various personal and professional tasks, including as part of Human Resource (HR) processes such as resume screening, interview analysis, and even cognitive assessments [28, 24]. While the appeal of LLMs to employers is understandable, given

their superior performance on multiple and diverse tasks [121], it also gives rise to new questions and problems, considering the technical complexity and novelty associated with LLMs. While research on bias in LLMs is steadily increasing, more knowledge is needed about fairness-related implications of replacing more 'traditional' AI models (such as SVMs or Gradient Boosting) with LLMs, which would require a side-by-side comparison of the two approaches. Furthermore, studies on bias learnt from data usually rely on real datasets, meaning that the extent and nature of bias potentially present in the data is uncertain, thus limiting the reliability of the results.

This research aims to address these limitations by providing a necessary comparison of traditional ML models and LLMs in the context of potentially discriminatory AI-based hiring practices. The study is based on synthetic data, which makes it possible to control the amount of bias the models are exposed to, and helps address the following research questions:

- **How do traditional ML models and LLMs compare in terms of bias and fairness when applied to hiring decisions?**
- **To what extent are these models robust to biased training data in various hiring scenarios?**
- **Do these models contain any inherent bias, unrelated to the data they are exposed to?**

I hypothesize that LLMs, despite their advanced capabilities, may exhibit higher levels of inherent bias than traditional models as they are pre-trained on large amounts of data that inevitably contain some level of bias. However, LLMs might be more robust than traditional models to biased training data, given that the LLMs used in this study are alleged to possess safeguards against bias [54, 87].

This study brings three main contributions to the field. First of all, **an overview of AI-based hiring practices is provided**, underlining the main applications, benefits, and risks and the impact of the increasing prevalence of LLMs in this field. Secondly, **three synthetic resume datasets are generated and made available for future research** in this direction. They contain various relevant features, two sensitive attributes, and labels reflecting the appropriateness of each candidate for a given position. Unlike existing datasets based on real data, the amount of bias is pre-defined and easily controllable, resulting in a first dataset that does not represent one gender or ethnicity as more qualified than the others. This dataset can also serve as a baseline for future studies. The other two datasets are intentionally biased in order to create an imbalance that makes certain demographic groups appear more qualified than others. Most importantly, **this study delivers a comparison of traditional models and LLMs in the context of hiring, focused on resume classification tasks**. The two groups of models are compared through the amounts of inherent bias they possess, and through their robustness to biased data in various contexts, all relevant to resume screening.

The findings suggest that the comparison of traditional models and LLMs is not as binary as expected, given the tremendous differences observed between BERT and GPT-3.5. The optimal choice of model for a hiring task ultimately depends on various factors. Still, this study shows that GPT-3.5 Turbo can be an asset for HR departments in terms of performance and fairness. The patterns observed in this study emphasize the fact that there are ways to avoid discrimination in AI-based hiring. Therefore, it is essential to raise awareness of this issue, to make solutions more accessible, and if need be, to strengthen regulations demanding fairness in automated decision-making.

Chapter 2

Definitions

2.1 Automated Decision-Making

Automated decision-making (ADM) is the process of making decisions using algorithmic systems, with minimal or no human intervention [41]. It can include complex decisions such as medical diagnoses, legal verdicts, or hiring choices.

While ADM systems can potentially offer more efficiency and impartiality than humans, they come with significant ethical challenges. A primary concern is the use of *black-box* models for these tasks, which are inherently opaque and uninterpretable [60]. Relying on opaque algorithms for high-stakes decisions is dangerous, as they can perpetuate bias, discrimination, and injustice [86]. Using such models raises questions about transparency, accountability, and fairness. These concerns are particularly salient when decisions significantly impact an individual's life and potentially infringe on their fundamental human rights [44].

2.2 Applicant Tracking Systems and AI-based Hiring

An *Applicant Tracking System (ATS)* is a software application designed to streamline the recruitment process by facilitating the management of applicants for a given job and automating various steps within the hiring process [100]. AI is a key component in some steps, such as *resume screening*, which refers to the classification, selection, or ranking of resumes through AI practices [37]. In this study, we define *AI-based hiring* as the collection of processes that rely heavily on AI to automate parts of the hiring process.

2.3 Sensitive Attribute

Sensitive attributes refer to characteristics of individuals that can pose privacy, safety, or discrimination concerns if not handled appropriately. The use of sensitive attributes, including but not limited to race, gender, marital status, etc., is strictly regulated and limited by law [30]. These regulations aim to prevent discrimination and protect individual privacy rights. In automated decision-making, the explicit inclusion of sensitive attributes in individual-linked data is generally prohibited under data protection regulations such as the GDPR [42].

However, completely excluding sensitive information is challenging. Sensitive attributes often correlate with seemingly neutral variables, known as *proxy variables* [120]. While these proxy

variables may be vital to the performance of automated decision-making systems, their use can inadvertently reintroduce bias, compromising the system's fairness. To address the ethical concerns surrounding the use of sensitive attributes in AI systems, [122] propose four key criteria for fairness:

- Protected attributes are not explicitly used in decision-making.
- Measures of predictive performance are equal across the groups defined by the protected attributes.
- Outcomes do not depend on protected attributes.
- Similar individuals are treated similarly.

While many sensitive attributes exist, this study focuses on gender and ethnicity. These attributes are chosen due to their well-documented potential for discrimination in AI-based hiring processes [26, 35]. Moreover, gender and ethnicity can often be inferred from other information in applicants' resumes, even when not explicitly stated, making them particularly relevant for investigating bias.

2.4 Fairness

Despite extensive research, there remains a need for more consensus on a single, optimal definition of *fairness* for use across all scientific contexts. From a technical perspective, a fair algorithm is optimised "without altering or manipulating [it] for purposes unrelated to the users' interest" [113]. Fairness is also defined as "the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making" [80]. This can be considered from three different perspectives [80]:

- **Individual fairness** claims that similar individuals should be treated similarly by the system and, therefore, obtain similar predictions.
- **Group fairness** is based on the idea that different groups (like minorities versus majority) should receive the same treatment overall.
- **Subgroup fairness** claims that group fairness constraints should hold over most subgroups.

This study focuses on group fairness, and therefore, when a model's fairness is discussed here, it refers to the extent to which it provides similar treatment to all groups within a sensitive attribute. This implies that all candidates are treated in the same way, regardless of their gender and ethnicity.

2.5 Discrimination

Discrimination is legally defined as "the unfair or unequal treatment of an individual (or group) based on certain characteristics such as income, education, gender or ethnicity" [48]. When bias in an AI system causes discrimination, it negatively impacts the system's fairness.

2.6 Bias

From a neutral standpoint, *bias* refers to a deviation from the standard, which can help identify statistical patterns in data [48]. However, it is a multifaceted term used in many contexts, with differing connotations, making it challenging to separate harmful unwanted social bias from a neutral statistical phenomenon [33]. Therefore, bias is not inherently harmful, but it definitely can be in instances where it treats individuals differently based on attributes such as gender, skin colour, religious beliefs, etc. This is not only highly unethical but also illegal, as it infringes on fundamental human rights [45]. So in this study, the term bias specifically refers to *discriminatory bias:* the problematic and systematic disparity between groups within a sensitive attribute, which can lead to discrimination and unfairness.

This type of harmful bias can be traced back to and observed in three different stages of the Machine Learning (ML) process [48]:

- **Bias in modeling** arises from how data is processed within the algorithm, often linked to the technical approaches or parameters chosen, which are usually only approximations of real-world phenomena that tend to be more complex than can be modeled.
- **Bias in training** originates from biased data and how it is pre-processed.
- **Bias in usage** arises when the algorithm is used in a context which it is not intended for and not appropriate for.

This study primarily indicates bias in training through the variation of the training data used but also reflects bias in usage through the different test datasets.

2.7 Privilege

The term *privilege* is used here to refer to instances where a model systematically favours one gender or ethnicity over another. That favoured group is then said to be privileged compared to the other group(s).

2.8 Traditional Model

In this study, *traditional models* are models that have become standard practice in AI-based hiring, specifically in resume classification. They are relatively simple ML models that do not rely on Deep Learning (DL) methods. The traditional models studied here are the Support Vector Classifier (SVC), Logistic Regression (LR), Random Forest Classifier (RF), and Gradient Boosting Classifier (GB). They have been in use for well over a decade and abundantly studied. In this context, they are being compared to *Large Language Models (LLMs)*, which are more novel, complex, and demanding in terms of resources. LLMs have a much higher number of parameters (hundreds of millions to billions) and are based on DL methods that are more difficult to interpret and understand. However, this complexity often allows for higher performance on more varied tasks, many of which cannot be completed with simpler models.

Chapter 3

Related Work

3.1 Hiring with AI

AI is becoming increasingly widespread and trusted in most (if not all) professional fields. It has already revolutionised healthcare [13], finance [23], marketing [58], and more. AI is now also commonly used in various HR tasks, including hiring-related tasks like resume screening.

3.1.1 Tasks

AI, with its efficiency and accuracy, can significantly enhance various stages and tasks within the hiring process, from job promotion to final decisions. It contributes to making the process more efficient and effective.

AI can be used as early as during the promotion of an available job or in the options presented to potential candidates as they search for a job [28]. Before or after the application process, AI can be instrumentalised for social media screening [70] to find suitable candidates, or to begin assessing applicants' suitability for a given position. This is often conducted in conjunction with keyword detection [70] when recruiters are looking for specific skills or experience.

AI can also be used in many ways with respect to resumes. Especially when the number of applications is unmanageable, AI can assist in filtering them (for example, with ATS) [28, 70] to remove unlikely candidates and decrease the number of resumes the HR department has to assess. Oftentimes, resume data is also automatically sorted or labeled through Named-Entity Recognition (NER) [25] to obtain more uniform data in a format that allows for further AI-based assessment. For instance, various ML models can be used to classify resumes into categories of the employer's choice or to rank those resumes in order of suitability for a job based on specific criteria extracted from applications.

Beyond the classical written job application, AI is now also being used in ulterior steps in the candidate selection process, with gamified tasks [28], often pertaining to positions that require intuition, fast thinking, or good reflexes. Conversations with chatbots are also beginning to appear as another step in the application process [28]. Recently, AI is even being used in holding or analysing candidate interviews [24, 73, 28], going as far as conducting AI-based social profiling to assess whether a candidate's personality is a good fit for the given job. This approach is starting to be used by large companies such as Unilever, which recruited half of its new hires in 2022 using a video-based AI assessment system [3]. Suffice it to say that AI has

its place in numerous aspects of the hiring process.

3.1.2 Models

The methods employed are numerous and diverse, but at this time, the most common approaches are rooted in relatively simple machine learning and statistical models [98], such as logistic regression, gradient boosting, support vector machines, decision trees [104] or neural networks [7]. They generally predict the quality of a candidate based on the various attributes available on their resumes or applications. These attributes are usually selected and classified using NER [71] to create a more straightforward format before prediction.

3.1.3 Benefits

Of course, this extensive use of AI during the hiring process yields considerable benefits.

Many large companies strongly focus on time, money, and workforce efficiency. AI can help automate processes that otherwise require the intervention of numerous workers [108]. This helps accelerate the application and hiring process while reducing and simplifying recruiters' workload [98, 81]. In the long run, the instrumentalisation of AI in hiring provides a more cost and time-efficient alternative to a fully human hiring process [116].

It is, however, essential to ensure that hiring with AI does not entail a decreased performance for companies not to have compromise on the quality of their new hires. Indeed, through the HIRE framework [116], designed to compare human and AI-assisted hiring, AI is not only suitable, but at least as performant and consistent as human hiring, if not more so.

Lastly, the use of AI in hiring is also based on the belief that it yields fairer outcomes and is less plagued by human bias and subjectivity. These beliefs are at least somewhat correct, as AI can help circumvent common human biases when evaluating applicants [96]. Using AI for this task can also improve diversity among the selected applicants [116].

3.2 Ethical Concerns in Hiring with AI

Hiring is a relatively high-stakes process with significant implications on individuals' professional lives and, therefore, their livelihood and well-being. Entrusting AI with such a task is risky and poses numerous ethical concerns, primarily related to the lack of human control and understanding of the ADM process.

AI systems can make mistakes that could be prevented or corrected with appropriate human oversight. Those mistakes, along with the lack of transparency of these models, can raise questions of responsibility, accountability, and liability [106]. Beyond this lack of transparency and human oversight, there is also a risk of problematic machine behaviour [14], such as distortion of reality or exploitation of individuals and their data. And of course, a central issue to AI-based hiring is bias and discrimination, which this thesis addresses. Lastly, AI lacks emotional intelligence [112], which is needed to evaluate humans on ambiguous criteria, such as personality, motivation, and compatibility with company culture.

These factors contribute to a general skepticism about AI-based hiring systems [89, 10], which can, in turn, affect the trust and respect people have for the companies making use of these systems and their decisions. Consequently, it is generally agreed that preserving a 'human

touch' in the hiring process is necessary, and that fully automated hiring is not something to aspire to [112].

3.2.1 Bias and Discrimination

Although a common justification for the increasing use of AI in hiring is the claim that it is objective and devoid of human biases [39], this is proving to be far from the truth. An overwhelming number of studies [39, 72, 62, 93, 69], along with real-world cases [115] show that AI-based hiring is far from objective and often just as biased as humans are, if not more so.

A well-known case of bias in an AI-based hiring system is Amazon's failed attempt at using an ATS to select top candidates in 2014. Their system was found to consistently downgrade female applicants because the training data used consisted mainly of resumes of white male employees [72], suggesting the importance of diverse and balanced training data for obtaining a fair AI hiring system. Ultimately, this led to the demise of Amazon's system as they could not fix the discrimination issue [72].

Multiple studies have shown AI-based hiring to exhibit the same social biases observed in society, sometimes even accentuating them. Within the existing literature, there is considerable proof of gender bias [90, 72] and racial bias [77, 72] in AI-based hiring practices. There are, however, other minorities that are less represented in this research, such as individuals with disabilities [85], yet are still very much affected by AI discrimination in hiring. This suggests that even when relying upon AI, the hiring process is at risk of the same social biases that prevail in society. Special attention should be given to potentially discriminating factors that there is not yet much research on in the case of AI-based hiring, such as appearance, age, etc. This also suggests the need for increased attention to intersectional biases [56].

There are numerous causes and potential explanations for bias in AI-based hiring. They can be split up into three categories:

Data

Data-induced bias stems from incomplete, unrepresentative, or biased data used for training the model(s) at the core of the AI hiring system. This is the most studied and prevalent cause of bias. It can be split into two phases: bias transferred from people and society to the data, and from the data to the algorithm [80].

From the examples of **bias originating from society** provided in [80], here are some that apply to the case of automated hiring:

- **Historical bias:** exists in the sampled population, and is therefore represented in the data originating from this population. As a result, the model learns these patterns of bias from the data.
- **Population bias:** stems from the use of data that is not representative of the population the AI system is meant to be applied to.
- **Temporal bias:** originates from changes in the population over time, making the data obsolete and no longer representative of the population.
- **Self-selection bias:** is caused by the participants of a study selecting themselves or opting in. For example, people choose to share their resumes for inclusion in a dataset.

- **Behavioural bias:** stems from differences in user behaviour, which is reflected in the content they generate.

Secondly, from the examples of **bias that arises during the data-to-model phase** provided in [80], some that apply to automated hiring are:

- **Measurement bias:** bias from the way various characteristics or features of the data are selected, measured, or used.
- **Omitted variable bias:** bias from one or several variables essential to a model's predictions being omitted from the data or the model.
- **Aggregation bias:** bias from incorrectly drawing conclusions about a data point based on generalisations.

Model Design and Architecture-Related Decisions

Bias can originate from the model design and the decisions surrounding its setup, such as:

- Bias from using a biased estimator, often unavoidable for performance reasons [33].
- Bias from parameter value choices like smoothing or regularisation: overall parameters that allow for the learning of a pattern, leading to assumptions that may lead to unjust decisions [33].
- Bias from the individuals making the design choices [68].
- Bias from reliance on protected attributes: to what extent the models use protected attributes (or proxies for those attributes) in making a prediction [68].

Application

This bias originates from how a finalised model is implemented and how decisions are drawn from it. For instance:

- Bias from application to inappropriate tasks or populations [33]; for example, training a model in data from one country and deploying the model in a culturally different country.
- Bias from over-reliance on the model in decision-making [68]: relying mindlessly on the model's output instead of making an informed decision based on the recommendation of a model with known risks and limits.
- Incorrect interpretation of results based on misinformation or lack of transparency about the models outputs [68, 33]; for instance, a model that outputs the five most qualified candidates in no particular order, but the user interprets it as a ranking.

3.2.2 Consequences

Hiring and recruitment are crucial processes in people's professional lives. The problem of bias in AI hiring can significantly impact applicants' careers, potentially affecting their livelihood and that of their families. This ripple effect can influence opportunities, education, health, and more. If pre-existing social biases are perpetuated and reinforced through the hiring selection process, society is at risk of increasing social determinism through AI, a tool marketed as fair and mathematically objective.

Furthermore, it is also in the best interest of employers to have a fair and unbiased hiring process. This allows them to find the most qualified and appropriate candidate for the job, thus contributing to the organization's productivity, philosophy, and overall success. Additionally, given the increasing regulations and requirements centred around AI fairness, employers risk litigation or a bad reputation if they do not comply with those new standards.

3.3 Bias Detection and Mitigation Strategies

3.3.1 Quantifying and Evaluating Bias

Quantifying bias is necessary to evaluate or mitigate bias computationally. This is a difficult task because bias is a social concept with various and sometimes conflicting definitions and interpretations. There have, however, been numerous attempts to align mathematical formulations of bias with its nuanced social definition [47, 95, 33, 48]. Beyond the theoretical discussion, some metrics, such as Disparate Impact [95] or Equal Opportunity Difference [59], have become standard in fairness and bias evaluation. Multiple toolkits [15, 102, 17] now offer out-of-the-box, easy to implement bias metrics. While such metrics are not synonymous with bias, they reflect an aspect of it and pave the way towards bias mitigation.

3.3.2 Bias Mitigation

This is an active area of research, with dozens of studies aiming to find solutions to the problem of bias in AI and its applications, including hiring [38, 62, 93, 35, 52, 120, 47, 5]. So, by now, solutions aiming to mitigate social bias in general-purpose AI systems do exist [6], and new ones continue to be suggested.

The most apparent solution is removing sensitive attributes from the data, which is also required by law [57]. However, the assumption that eliminating sensitive attributes circumvents bias is naive and simplistic, as they are not characteristics that can be isolated but a reflection of power dynamics that are ingrained in several aspects of AI-based hiring [39]. So removing the 'sensitive attribute columns' from a dataset can promote fairness but is not sufficient due to other information in the data acting as proxies [120] for sensitive attributes. Removing these proxies as well can lead to an excessive loss of information, which could significantly impact model performance. Furthermore, it needs to be clarified that including sensitive attributes in the training data really does lead to unfairness; in fact, some studies show the contrary [57].

A range of technical solutions have been proposed for bias mitigation. They can be split into three categories [99]:

- **Pre-processing solutions:** methods acting upon the dataset or the feature selection process, such as reweighing protected attributes, denoising the data, etc.

- **In-processing solutions:** methods targeting the model and its inner workings. Examples include an adversarial method to decrease bias [61] and a form of regularisation geared towards fairness [91].
- **Post-processing solutions:** techniques applied to the model outcomes, such as re-ranking the outputs after including individuals from protected classes, which had been filtered out by the model [22].

Beyond these suggestions for targeted fixes, there is a high demand for more transparency in AI-based hiring, given that the opacity of these models interferes with applicants' rights to non-discrimination [69]. Transparency is highly recommended regarding the dataset and metrics used, as well as the intent of the model and its applications [72]. However, there are trade secret concerns and limitations [72] to be taken into consideration.

There is also demand for new or modified regulation on this front [95], to ensure that employers abide by principles of fair use of AI and to increase public trust in these systems. The importance of diversity in AI is also emphasised as a means to include various perspectives in the model design process and limit blind spots when it comes to algorithmic bias [95].

Lastly, even with a highly regulated model, with safeguards and incorporated bias mitigation methods, some human control and verifiability are recommended [28]. This implies that in any high-stakes automated decision-making process, including hiring, AI shouldn't have the final say but merely serve as a tool providing nuanced recommendations to competent professionals who can evaluate them critically before making final decisions.

However, even if in theory those solutions exist and can be implemented, in practice, the issue of bias in AI hiring endures. The broader issue of (un)fairness in AI is complex, multi-faceted, and context-dependent [15], making it difficult to tackle by experts and even more so by companies/employers who use an ATS but do not have a deep understanding of AI and its fairness-related complexities. The gap between academia and industry could also make the available solutions inaccessible to those who should be implementing them. There are, however, increasing efforts to bridge this gap, rendering debiasing methods understandable and accessible to the general public to facilitate the introduction of these debiasing initiatives in the professional world. This can be observed through multiple bias auditing and correction toolkits, such as AIF360 [15] or FairLearn [17], which aim to make debiasing approachable and free.

3.4 Emergence of LLMs: Implications for Fairness in Hiring

3.4.1 Emergence of LLMs

However, on the technical side, AI is going through a revolution with the increasing power and use of Large Language Models (LLMs) [83, 27] in numerous domains. Fields related to language and communication have significantly benefited from this phenomenon, given the proficiency of LLMs in tasks like summarization or translation [21]. The medical, educational, and finance fields have also been transformed with these models [83] that have only begun being vastly used a few years ago.

3.4.2 The Potential of LLMs in Hiring-Related Tasks

Naturally, given the commonality of text data in hiring and the importance of 'understanding' context, subtleties, and queries, it is reasonable to expect LLMs to start being used in hiring as well. Sure enough, research on this is already appearing, looking into how various parts of the hiring process can be transformed to rely fully or partially on LLMs.

An interesting initiative is using LLM-based chatbots to assist HR departments during the hiring process [105]. These chatbots can help accelerate the assessment of candidates by finding relevant and targeted questions to ask, as well as by being able to analyse video and audio content from interviews, using tone of voice and facial expressions. In addition to being cost and time-efficient, this can allegedly also help decrease interviewer bias [105]. Another novel LLM-based hiring practice is the use of GANs to infer implicit information about candidates from the self-descriptions they provide [40]. This information can help improve the job recommendations suggested to the candidates [40]. Comparisons of traditional, supervised models and LLMs are also emerging. One study [29] specifically makes this comparison in the context of a binary (text) classification task. The findings show that the GPT-3.5 Turbo model outperforms all other models, including the supervised SVM baseline, when the prompt is well-selected. The authors also highlight a necessary paradigm change driven by the transition to GPT-based solutions; performance improvements are less dependent on how the model is parameterised, but rather on the quality of the prompt engineering [29].

Most relevant to this study are the developments related to the use of LLMs in resume screening, which encompasses filtering/shortlisting resumes that are most promising, ranking resumes, as well as classifying them. A recent study [92] compared LLMs with more traditional, supervised approaches to resume classification. The findings showed that LLMs, specifically GPT-based architectures, significantly outperformed other models based on common performance metrics such as accuracy, precision, and recall. Another study [109] investigated the use of Gemini (Google's multimodal LLM) for resume screening as well, finding that it was more efficient and more objective than supervised models due to its ability to better understand semantics and overall context through an optimised mix of ML and NLP techniques. There is also abundant literature [55, 34, 2] on the benefits of using BERT-based models for ranking or shortlisting the most appropriate candidates based on their resumes. These studies unanimously conclude that BERT-based models by far outperform more traditional methods, which often primarily rely on keywords, on this task in terms of accuracy and efficiency. These models are also less prone to unwanted bias [55]. So, it is clear that automated hiring processes can benefit from the increasing capacities of LLMs. However, these studies are primarily focused on performance and efficiency. While some of them do address bias in the use of LLMs in hiring, it is mostly viewed as a positive byproduct of the increased performance [105, 55], not as a risk that increases along with the trust we have in LLMs and the responsibilities allocated to them. This highlights the need for careful and critical consideration of these 'improvements' from the perspective of fairness and bias.

3.4.3 Implications for Bias

Bias in BERT and GPT

Multiple studies on bias and discrimination in LLM outputs and applications have emerged in the past year, pertaining to various tasks. Gender bias has been observed in BERT model embeddings [16], influencing its predictions in various tasks. One study [66] tested for bias

after varying different elements of the BERT training pipeline, concluding that although bias levels do vary with the pipeline modifications, the bias remains significant in all scenarios and is therefore related to the architecture of BERT rather than to task-specific data. Although most bias research in BERT is centred around gender bias, there have been studies also investigating and observing racial bias in BERT, for instance, in a resume retrieval system [118]. GPT-based models are not immune to bias either. Various gender stereotypes were found in GPT-generated content [78], and both racial and gender bias were observed in instances of GPT used in the medical field [117]. A more comprehensive study [97] grouped the 23 types of biases and 22 limitations observed or suspected in GPT models. Cultural, gender, and racial bias are at the top of the list, which interestingly also includes many biases characteristic of human reasoning, such as cognitive bias and confirmation bias [97].

In the context of hiring, a study on GPT-3.5, Llama-2-70b, and Mistral 7b has revealed ethnic bias in generating response letters (positive or negative) to various applicants, with white candidates being favoured over Hispanic ones. However, it is essential to note that this study has some limitations, including its inclusion of only three ethnicities, and the models' strong sensitivity to the prompts/templates used. In resume generation using GPT-3.5, gender and ethnic bias against Hispanic and Asian applicants were also observed. Another study, which applied GPT-3.5 to resume classification and summarization tasks, found no bias related to race and gender, but did observe small amounts of bias related to maternity employment gaps, pregnancy status, and political affiliation. However, these amounts of bias were found to be less significant than those observed in other models.

While research on bias in BERT models in hiring practices is currently limited, both GPT and BERT have been extensively studied in terms of their potential for hiring-related applications. This underscores the urgent need for further research on bias in these models, particularly in the context of specific hiring tasks such as resume classification, which is the focus of this study.

Bias Detection in LLMs

Bias in LLMs is a well-known problem, and some research on debiasing methods already exists. The bias evaluation and mitigation methods proposed differ by the size of the model [75], so it is clear that there is no universal solution to biased LLMs.

Given the complexity of these models, the first challenge is detecting and identifying biases. Beyond common bias/fairness metrics [15, 17], LLM-specific bias detection mechanisms are emerging. For transformer-based models such as BERT and GPT, a gender bias detection method based on the attention mechanism has been suggested [74]. It leverages attention scores to highlight the correlation between gender indicators (names and pronouns) and professional occupations (such as nurse or pilot). Other studies focus on methods able to disclose implicit biases [38, 12] in LLMs to account for the fact that they can 'pass' tests designed to detect bias while still relying on biases in predicting outcomes. These methods include an indirect probing framework [38], a prompt-based method[12] targeting implicit biases, as well as a way to detect more subtle forms of discrimination present in specific decision-making tasks [12]. Another study aims to address the low interpretability of most bias detection methods by creating the GPTBIAS framework [119], which outputs not only a numerical indication of bias but also identifies the types of bias and affected groups, and provides suggestions for bias mitigation. Lastly, there have also been efforts to find ways to detect more complex intersectional biases [111], which are often overlooked. While bias detection is the first step to fairer LLMs, mitigating those biases is just as big of a challenge.

Bias Avoidance and Mitigation Techniques

Some novel ways to debias LMMs are Hyperparameter Tuning, Instruction Guiding, and Debias Tuning [38]. Still, more information is needed on their applicability and efficiency in models based on BERT and GPT architectures.

Several bias mitigation methods have recently been proposed for BERT-based architectures, many acting upon the loss function or word embeddings. One suggestion is to define the loss term as the correlation between the embeddings and a 'gender bias subspace' [118] to optimise training in a way that limits semantic bias. Applied in the context of resume ranking, this method achieves a decrease of over 60% of the ranking gap between genders [118]. Another proposed way to decrease gender bias in BERT is removing gender-specific information from every layer of the model, which has proven successful in decreasing gender bias in downstream tasks [16]. Another study focused on mitigating racial bias suggests a reweighing technique for the intermediate values of the fine-tuning process that are then used within the loss function [82]. Another interesting way to address ethnic bias in a multilingual context is using various somewhat different BERT-based models. This can be done either with a multilingual instance of BERT, which is trained on data in different languages, and thus with different disadvantaged groups, which allows for a compensation of the different biases, decreasing the extent of those biases in the target model [4]. Alternatively, two monolingual BERT models can be used to alleviate ethnic bias by aligning the contextual word embedding of the target model with the embeddings of a less biased model [4].

Research on debiasing GPT models is less abundant, even though the occurrence and impact of social biases in GPT are thoroughly researched. This could indicate that GPT is more difficult to debias than other models, as suggested by a study [79] that proposes a novel debiasing method that is successful for all models included in the study except for GPT-2. This could be partially explained by the high number of parameters in GPT models, and hence the time and resource limitations involved in debiasing such a model without compromising performance. This could dissuade employers/companies from debiasing their models. However, there have also been successful attempts at debiasing GPT models. One example is based on a reinforcement learning approach, which successfully decreases political bias with minimal use of resources and without compromising the quality of the generated content [76]. Another efficient approach to debiasing is fine-tuning about 1% of the parameters, selected strategically, and freezing the rest [53] to reduce gender bias at a low cost. Furthermore, a zero-shot self-debiasing method was also proposed and tested on GPT-3.5 Turbo [50] to make the debiasing process more efficient. Lastly, data augmentation and saliency methods could also help debias LLMs, including GPT-based models, while preserving performance [67].

While debiasing methods for LLMs exist, they are difficult to employ in non-academic contexts because of actual or alleged drawbacks, such as trade-offs with performance or difficulty of implementation [47]. Indeed, many of these methods (mentioned above) are proposed academically and are not as ideal in the professional world since they are often very specific (thus relevant scientifically but not so applicable) or difficult to implement for those who do not have sufficient technical knowledge. Therefore, discrimination in AI-based hiring persists.

Chapter 4

Method

4.1 Dataset

4.1.1 Dataset Generation

In the context of this research proposal, many publicly available datasets (from sources like Kaggle or Census data) related to HR tasks were considered. While there is no shortage of such datasets, none were adequate in terms of the specific requirements of this study. The main problems observed in the existing datasets include the absence of sensitive attributes, vital to evaluating bias, and the absence of labels, which are necessary in training models for classification tasks. Furthermore, existing datasets exhibit improper formatting and insufficient data, making them impractical without time-consuming alterations.

A notable attempt to overcome these shortcomings was made in a recent publication [51], where the authors augmented an existing dataset by labeling it using GPT, making it ideal for this study. This dataset has not been made publicly available, but the approach of [51] inspired the dataset generation strategy employed in this thesis.

Given the lack of appropriate and accessible resume datasets for bias evaluation, this research includes the development of three custom-made datasets, which will also be made publicly available to facilitate future research. To this end, GPT-3.5 Turbo was used through the OpenAI API to generate synthetic resumes, including sensitive attributes and a label that indicates each candidate's suitability for a given profession. A list of names was generated beforehand by the same means to ensure better name diversity, and then one resume was generated for each name. The choice of the GPT-3.5 Turbo model was based on its state-of-the-art performance and relatively simple implementation. The Turbo version of the model was selected to make the data generation faster and less computationally expensive.

Prompt for dataset generation with GPT-3.5 Turbo: *"Generate 10 resumes for a hairdresser job corresponding to the names ..., ..., ..., ..., ..., ..., ..., ..., ... and .... These should reflect 10 excellent / slightly above average / average / below average / completely inadequate candidates for the job. Each resume should include a name, gender associated with the name, email address, education, GPA, work experience, skills, and any awards received. Each resume must contain at least 200 words. Ensure that the resumes are detailed and unique, showcasing a range of qualifications and experiences, and a wide variety of backgrounds. Do not use placeholders such as 'John Doe' or 'Cityville'; names and locations should be real."*

4.1.2 Synthetic Datasets

Three datasets were generated. These datasets, presented in tabular format (CSV), contain a diverse range of personal and professional information typically found in resumes. The information is organized under pre-defined headings, providing a comprehensive picture of each candidate's profile. A small sample of the dataset is shown in Appendix C.

7 Features: Name, Email, Education, GPA, Professional Experience, Awards, Job (position the candidate is applying for, six options)

5 Labels: Very bad, Bad, Average, Good, Very Good. They indicate the candidate's suitability for a given position (stated in the 'Job' column). The five classes are present in the same amount throughout all datasets.

6 Professions: Secretary, Engineer, Hairdresser, Surgeon, Nurse, Construction Worker. Although in this study, the amount of bias is not evaluated based on these categories, they are selected in a way to allow others to evaluate profession-specific bias when the datasets are made publicly available (such as 'female candidates are more likely to be nurses, and male candidates more likely to be surgeons'). The different jobs are equally present in all three datasets. Within the datasets, there are 48 resumes is every label-job combination (so 48 'Very Bad' engineers, 48 'Good' nurses, etc.).

2 Sensitive Attributes: The sensitive attributes selected for this study are gender and ethnicity. This choice was based on the abundant research showing that those attributes are discriminatory factors in AI-based hiring [26, 39, 93]. Consequently, the forms of discrimination addressed here represent a real-world problem. Of course, other biases exist and are worth studying and confronting. However, due to time and resource constraints, a choice had to be made, and studying well-known biases facilitates the comparison of results with previous research.

- **Gender:** This sensitive attribute is binary, so each candidate is either male (M) or female (F).
- **Ethnicity:** This attribute can take one of 6 different values: White American (WA), Black American / African American (BA), East Asian (EA), White European (WE), Hispanic (H) and African (AF). These ethnicities were selected by taking the six most prevalent ethnicities in the USA labour market [65]. The datasets were modeled after the USA population because it is larger and more diverse than most other countries. Furthermore, the ethnic contrast between individuals is easier to represent in a resume dataset than, for example, the contrast within a European country, which is less diverse and primarily a mix of various European nationalities [64], often without clear name differences, and no significant discrimination to speak of.

Balanced Dataset

The first dataset aims to be devoid of discriminatory bias by being completely balanced regarding gender and ethnicity. This means that there is the same number of male and female resumes, and the same number of resumes corresponding to candidates of each of the six ethnicities included in the data. Furthermore, the data is also balanced in terms of quality; no gender or ethnicity has 'better' candidates than the others, which is reflected in the content of the resume data, as well as in their associated labels. This dataset serves as a baseline, reflecting

the 'ideal' scenario in which models are trained on fair, balanced data. This particular choice of dataset also allows for an evaluation of the inherent bias in models, which does not originate from biased data.

Gender-Biased Dataset

A second dataset is derived from the balanced dataset through the voluntary introduction of gender bias. As can be seen in figure 4.1, gender bias is created by randomly sampling 50% of the female candidates from each class (except 'Very Bad') and downgrading them by one level. For instance, a 'Very Good' resume would be relabeled as 'Good'. The opposite process is conducted for the male candidates to keep the number of resumes under each level constant. So 50% of male candidates in each class (except 'Very Good') are upgraded by one level. Modifying the labels for 50% of the dataset is somewhat arbitrary. Still, it aims to reflect a moderate amount of bias, which translates to half the female candidates being slightly disadvantaged and half the male candidates being somewhat privileged. The resulting distribution is shown in table 4.2. Gender is a sensitive attribute and, therefore, not explicitly included as a feature during the training process. However, it can be inferred from the candidates' names and the pronouns used in the text data.

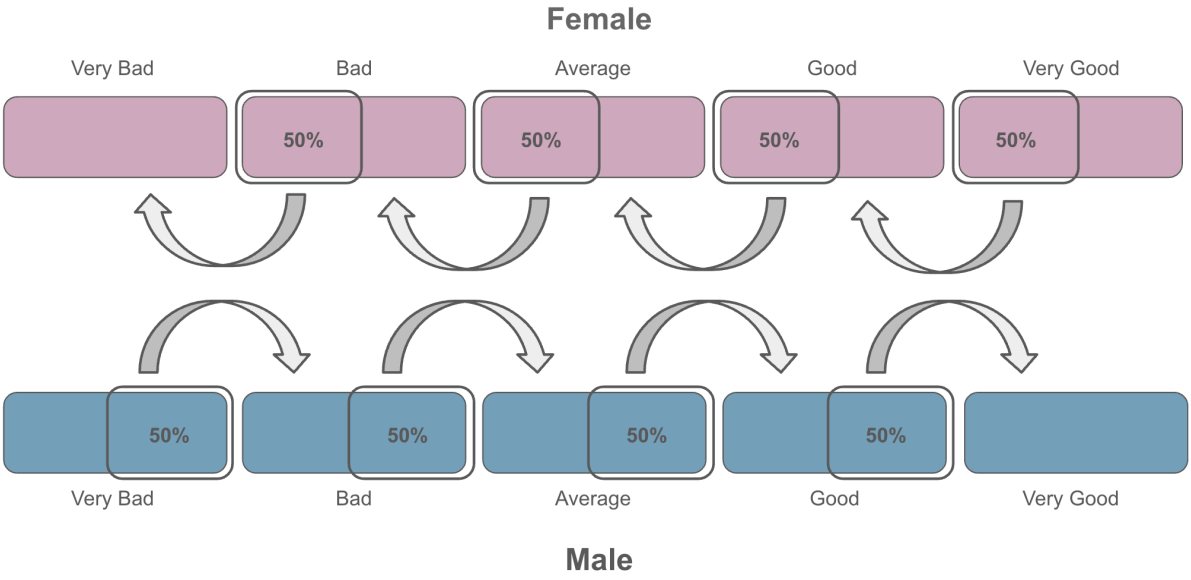

Figure 4.1: Diagram of the Data Biasing Process

Ethnicity-Biased Dataset

To create the third dataset, ethnic bias is introduced in the balanced dataset in a similar manner and in an amount comparable to the gender bias. The six ethnicities are split into two groups, based on which ethnicities commonly put candidates at an advantage/disadvantage in the USA labour market. Research shows that White Americans (WA) and Europeans (WE), as well as East Asian (EA) candidates, tend to be overestimated or privileged in terms of hiring [88, 101]. Meanwhile, Black American (BA), Latino (L), and African (AF) ethnicities tend to face discrimination [88]. Therefore, the expected privileged group is defined as (WA, WE, EA), and the expected underprivileged group is (BA, H, AF). The biasing process is the same as in figure 4.1, with 50% of the candidates of expected underprivileged ethnicities being downgraded (like the female candidates), while 50% of the candidates of expected privileged ethnicities are upgraded (like the male candidates). Similarly to the gender attribute, modifying the labels for

50% of this dataset is somewhat arbitrary but aims to reflect a realistic scenario of moderate yet significant bias. This is done separately for each ethnicity, resulting in the distribution displayed in table 4.3. The ethnicity attribute is also not directly used as a feature during the training process but can be inferred from the candidates' names and, in some cases, from mentions of locations (such as schools attended and previous workplaces).

	Female	Male
Very Bad	216	72
Bad	144	144
Average	144	144
Good	144	144
Very Good	72	216

Figure 4.2: Gender-Biased Data
(in the balanced data, all entries are 144)

	WA	H	BA	EA	WE	AF
Very Bad	24	72	72	24	24	72
Bad	48	48	48	48	48	48
Average	48	48	48	48	48	48
Good	48	48	48	48	48	48
Very Good	72	24	24	72	72	24

Figure 4.3: Ethnicity-Biased Data
(in the balanced data, all entries are 48)

4.1.3 Dataset Verification

After the datasets were generated, some verifications were required to confirm that the assumptions made about the dataset were correct. First, the various proportions of candidates in each subgroup were verified. This confirmed that the balanced dataset was completely balanced regarding gender and ethnicity across job titles and labels. The same verification was conducted to confirm that the biased datasets also contained the expected gender and ethnicity ratios for each subgroup.

Furthermore, the three datasets were fully checked by hand, leading to several modifications:

- Format inconsistencies linked to the use of GPT for data generation were fixed.
- The ethnicity and gender corresponding to the name of each candidate were added after the data generation to limit the influence of sensitive attributes on the quality of the resumes.
- Minor adjustments were made to the number of items in each subgroup when the proportions were not exactly as planned.
- Most gender-neutral names were replaced by more gender-specific ones to avoid ambiguity over gender in some subgroups compared to others.
- Some neutral or mixed names in terms of ethnicity were also replaced for the same reason.

4.1.4 Data Within the Experimental Setup

The training and testing data used in this study includes all the features of the dataset, with the exception of the email (which seemed irrelevant) and the GPA, which was too strongly correlated to the labels and hence prevented learning based on the other features. The sensitive attribute values and the labels are, of course, also excluded. For all models and all scenarios, the train/test split is 80%-20% (randomised).

4.2 Models

4.2.1 Traditional Models

This study compares the extent of bias in models traditionally used for resume classification with the extent of bias in LLMs. The traditional models included here are selected to reflect common and well-established hiring practices. Those models are abundantly studied and often come up in research about AI-based hiring, specifically resume classification [104]. For practical purposes, another criterion in selecting those models was their fast training time and similar output format so that the same metrics can be used for evaluating all models.

The four traditional models used in this study are from the Scikit Learn library [1], version 1.4.1.post1. Bootstrap resampling with replacement is used for the four models, reaching 10 runs per model. The resulting fairness metrics and accuracy are computed as the mean of the ten values obtained. The margin of error is also based on those ten values, and Bessel's correction is used. During the pre-processing phase, TF-IDF vectorization from the sklearn.feature_extraction.text library is applied to the data. The output of each model has three components: a list of the model's predictions on the test set, a list of the true labels corresponding to those items, and a list of the relevant sensitive attribute values.

1. Support Vector Classifier (SVC):

The SVC used here is imported from sklearn.svm. The SVC is a supervised ML model used for classification tasks. It is the classifier version of the SVM [31]. Its goal is to find the optimal hyperplane that maximizes the margin between the boundaries of the different classes.

Regarding hyperparameter settings, a linear kernel is used because it yielded the highest accuracy in this context, making it the most realistic setting (as users of hiring models would seek the highest accuracy). The regularization parameter is set to 1.7, as it also yields the highest accuracy. This indicates that less regularisation is needed here than the default. The rest of the hyperparameters are set to their default values.

2. Logistic Regression (LR):

The Logistic Regression (LR) model is imported from sklearn.linear_model. Logistic regression [32] is a classification model that predicts the probability of an outcome using the sigmoid function. It uses the probability of an input belonging to a specific class to make predictions. The model is trained by minimizing the log-loss using optimization techniques like gradient descent.

The only adjustment made to the model's default settings is the number of iterations, which is set to 1000. The rest of the hyperparameter values are left as per the default settings.

3. Gradient Boosting

The Gradient Boosting (GB) Classifier is imported from sklearn.ensemble. Gradient Boosting [49] is an ensemble ML model that builds a predefined number of decision trees sequentially, each aiming to correct the mistakes of the previous tree. The final model is a weighted sum of all the decision trees, with larger weights attributed to the more accurate trees.

The number of estimators used is 90 since it results in the highest accuracy, according to some preliminary tests. No further modifications are made, and the rest of the hyperparameters are set to their default values.

4. Random Forest (RF)

The Random Forest (RF) Classifier is imported from sklearn.ensemble. Similarly to GB, a Random Forest [19] model is an ensemble method that builds a predefined number of decision trees, each trained on a different random subset of the data and features. The final model weighs the output of each tree equally.

Identically to the GB classifier, the number of estimators used is 90 since it results in the highest accuracy according to some preliminary tests. The rest of the hyperparameters are again set to their default values.

4.2.2 LLMs

The LLMs compared with traditional models in this study are BERT [36] and GPT-3.5 Turbo, a newer and faster version of GPT-3 [20]. They were selected as they are among the most widely used LLMs and, so far, the most researched in the context of AI-based hiring. This makes them relevant models to evaluate for bias. There were also practical factors influencing this decision; BERT and GPT-3.5 Turbo were both publicly accessible without requiring a special authorisation (unlike Llama), and reasonable in terms of resource requirements, as compared to other alternatives. However, an instance of BERT in this scenario takes over two hours to train, and an instance of GPT-3.5 Turbo takes almost 15 minutes. However, due to the number of instances of each model needed for this study (8), there are still time and resource constraints to consider. Therefore, the accuracy and fairness metrics, along with their margin of error, are computed based on 5 runs per model and per scenario (unlike the preferable 10 runs per traditional model). In the margin of error computations, Bessel's correction is implemented. Both models are trained and tested on CPU.

1. BERT

The pre-trained BERT model used here is the *bert-base-uncased* from BertForSequenceClassification from the transformers library.

Data: The different categories of the text input are concatenated, resulting in one text item per candidate. Next, the data is tokenized with the BertTokenizer from the transformers library. The tokenized sequences are padded to a maximum length of 128 tokens to ensure uniform input sizes for the BERT model. The text labels from the 'Fit' column are converted to numerical labels and encoded as shown in figure 4.4. The training data is then converted to tensors and loaded as PyTorch DataLoader objects, facilitating efficient batching and shuffling during training.

Learning: The BERT model is initialised with pre-trained weights and then fine-tuned on the synthetic data for the resume classification task. The **AdamW optimiser** from the transformers library is used to adjust the weights based on the **cross-entropy loss function**. The learning rate determines the extent to which the weights are adjusted at every update. Here, the selected **learning rate is 5e-5**. The fine-tuning consists of **8 epochs** in this study, based on some preliminary tests which showed that performance increase slows down beyond the 8 epochs. This choice is also partly motivated by time and resource constraints. The **batch size is also set to 8**, which was the largest value possible here due to memory limitations.

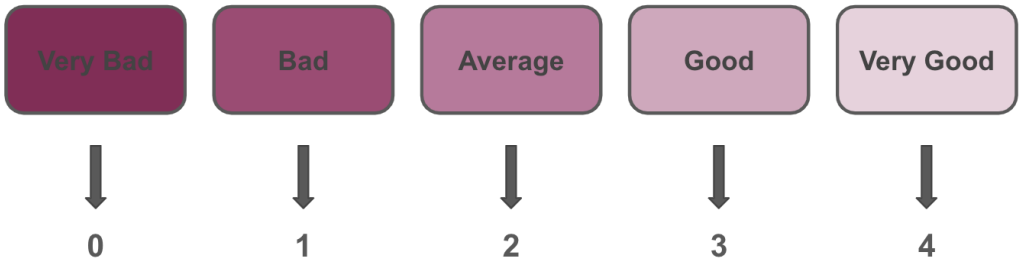

Figure 4.4: Label Encodings for BERT

Testing: Once the fine-tuning is completed, the model outputs a list of predicted labels and a list of true labels for the test set, encoded as shown in figure 4.4. A list of the value of the selected sensitive attribute (gender or ethnicity) for each item in the test set is also outputted.

2. GPT-3.5 Turbo

OpenAI's GPT-3.5 Turbo model for fine-tuning and inference is used. The first step is defining a task description along with a set of valid outputs (the labels):

Task description: *"Classify each candidate into one of the following categories, based on how good a fit they are for the given job: [very bad, bad, average, good, very good]."*

Categories: ["very bad", "bad", "average", "good", "very good"]

Data: The items in the training data are processed one by one during the fine-tuning phase, where it is concatenated into a single text string, resulting in one string per resume. During the validation phase, the test data is processed in the same manner. No further processing of the data is necessary.

Fine-tuning: During fine-tuning, each resume in the training set is prompted for classification. The prompt is iteratively updated with the next item in the training set and sent to the GPT-3.5 API to fine-tune the model. This prompt contains the predefined description, categories, and the training set item (text data and label). The **temperature hyperparameter is set to 0.7**, defining the level of randomness and diversity in the outputs. This value is relatively low given that this is a classification task with predefined classes and, therefore, benefits from consistency more than diversity. The **maximum number of tokens is set to 50**. This is not necessary here but serves as a computational precaution for the possibility of the model generating large amounts of text instead of outputting one of the predefined classes.

Testing: At inference, the resumes in the test set are prompted for classification one at a time. The outputs are in the same format as for BERT; the sensitive attribute values, the predicted labels, and the true labels are stored for the bias evaluation. The only difference is that the labels are not numerically encoded, so the outputs are simply the predefined categories.

4.3 Scenarios

This thesis studies three distinct scenarios, mirroring real-world situations. They are all conducted separately for gender and ethnic bias.

4.3.1 Scenario 1: Inherent Bias

The models are trained and tested on the balanced, unbiased dataset. This reflects an ideal scenario in which the model is provided with unbiased training data and is then applied to data that has not been corrupted by bias in a previous step of the hiring process. This scenario is rare and difficult to achieve since unbiased data is scarce (unless it is synthetic, but this approach is uncommon). Furthermore, ensuring that all previous steps of the hiring process are devoid of bias is also not a small feat. The value of this scenario is its ability to highlight the inherent bias of the models, as the bias measured in this scenario cannot originate from the data. It will also show the extent to which balanced data can prevent bias and whether investing in data debiasing methods or synthetic data generation is worth it.

4.3.2 Scenario 2: Robustness to Bias

The models are trained on the biased dataset(s) and tested on the balanced dataset. While the previous scenario was a difficultly attainable ideal, this one reflects the status quo. This scenario represents cases where an automated hiring model is trained on biased data, which is frequently the case. It is then applied to unbiased (or less biased) data. This scenario is likely since, in practice, resume classification can be the first step in the hiring process, meaning that it could not have been significantly biased beforehand. However, this claim is not always true; the set of resumes can still be biased through how the position is advertised and through any administrative hurdles (complicated online procedure, application fee, language requirements, etc.). This scenario is the leading indicator of the models' robustness to bias in the training data.

4.3.3 Scenario 3: Application to Biased Data

This scenario is designed to highlight the effect of applying a resume classification model to a dataset that has been corrupted by bias. This is common when resumes reachin this step of the hiring process have already been selected or ranked through a biased procedure, whether human or automated. This phenomenon is highly likely if the resume classification is not the first step of the hiring process, unless specific measures are taken to avoid or mitigate bias.

A. Biased Model Applied to Biased Data

The models are trained and tested on the biased dataset(s). This scenario reflects cases where the model is trained on biased data (which is usually the case) and then applied to data that suffers from the same bias as the training data.

B. Unbiased Model Applied to Biased Data

This scenario represents the unlikely event of a model being trained on balanced data and then applied to biased data. This implies that the model used by the employer/recruiter is not biased and is trained on balanced data, which is, for now, difficult to come across and seldom used. If the employer made an effort to provide balanced data, they probably also attempted to ensure fairness throughout the rest of the hiring process. Nevertheless, this scenario remains possible, and preparing for such situations is important.

4.4 Metrics

Five different fairness metrics are used here. These specific metrics were selected as they are common in bias research and included in multiple fairness toolkits [15, 17, 102], meaning that they are legitimate and representative of the bias present within a model's predictions. The chosen metrics have the added practical benefit of requiring the same input format and outputting comparable values. For all metrics, the output is between 0 and 1, 0 representing a total lack of bias (equal treatment of all demographic groups) according to the metric in question, and 1 representing the maximum amount of bias possible according to that same metric. This allows for easily comparable and reliable results.

The five metrics are representative of group fairness as they aim to expose discrepancies in the treatment of different demographic groups [80], defined by their belonging to a protected class. The metrics are all based on the simple statistical notions of False Negatives (FN), True Negatives (TN), True Positives (TP), and False Positives (FP). A visualisation of these concepts is provided in figure 4.5.

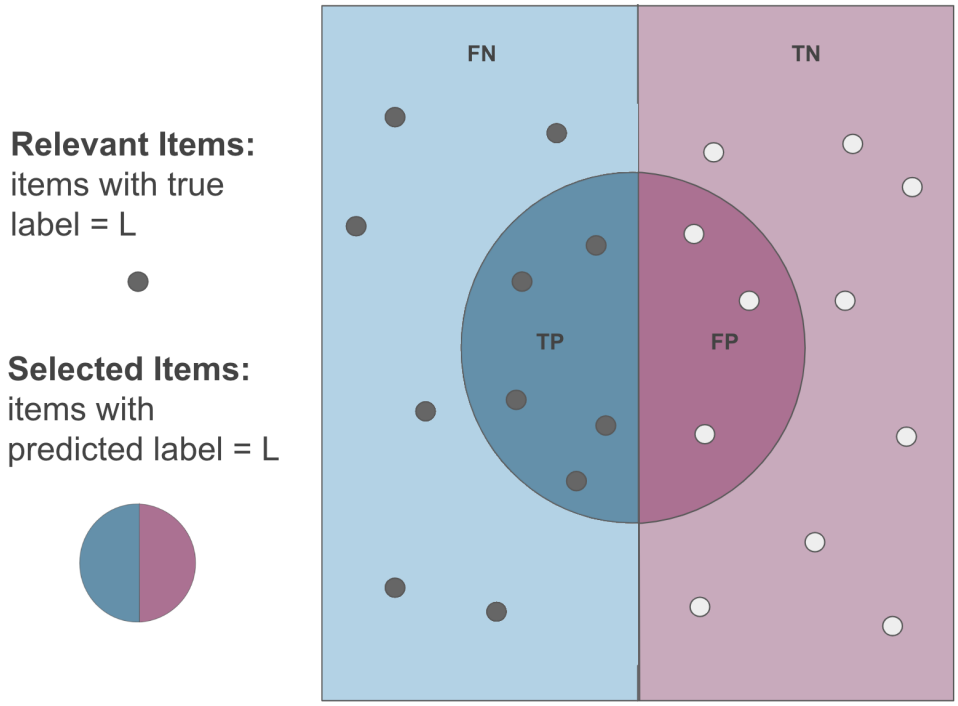

Figure 4.5: Components of the group fairness metrics used in this study

4.4.1 Demographic Parity Difference

Demographic parity difference (DPD) [46] is a simple metric that compares the rate of positive outcomes across different demographic groups. It is computed by first taking the selection rate (SR) for each label and demographic group. Then, for each label, the absolute value of the difference in selection rates between groups constitutes the DPD. Figure 4.6 provides a visualisation of this computation and a mathematical definition of DPD.

This metric is indicative of fairness since it reflects the difference of likelihood of an item from each group receiving a positive outcome for a given label. In this case, a higher selection rate for positive labels ('Good' and 'Very Good') constitutes privilege.

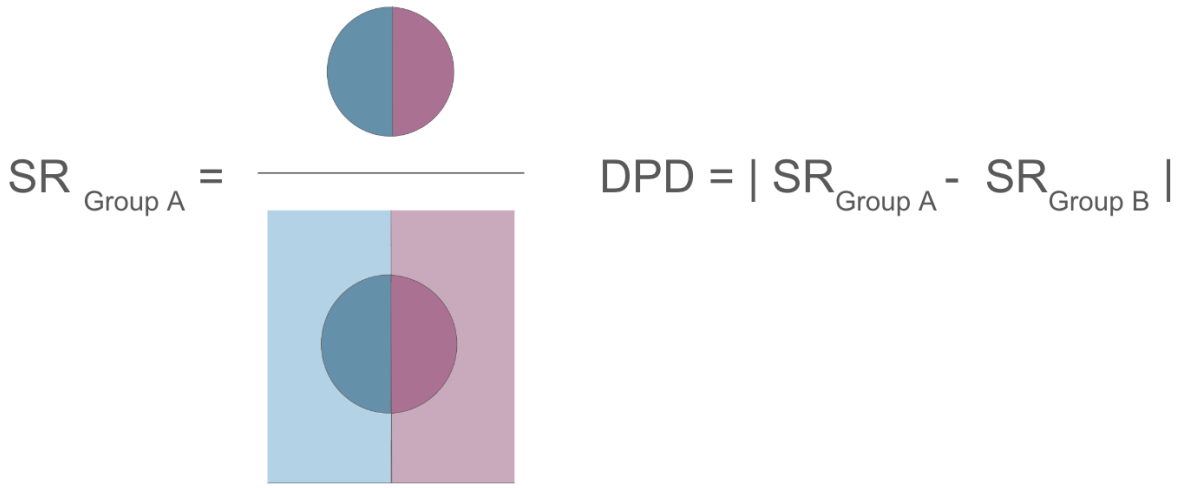

Figure 4.6: Selection Rate (SR) and Demographic Parity Difference (DPD)

4.4.2 Equal Opportunity Difference

The second metric used is Equal Opportunity Difference (EOD) [46]. It is computed as the absolute value of the sensitivity/recall difference between groups, as shown in figure 4.7.

This metric relates to fairness by showing how capable a model is of correctly identifying relevant instances in the data for different groups. Correct classifications of points with an advantageous label such as 'Good' or 'Very Good' constitute privilege. In contrast, correct classification of points with a 'Very Bad' or 'Bad' label is a disadvantage for the group.

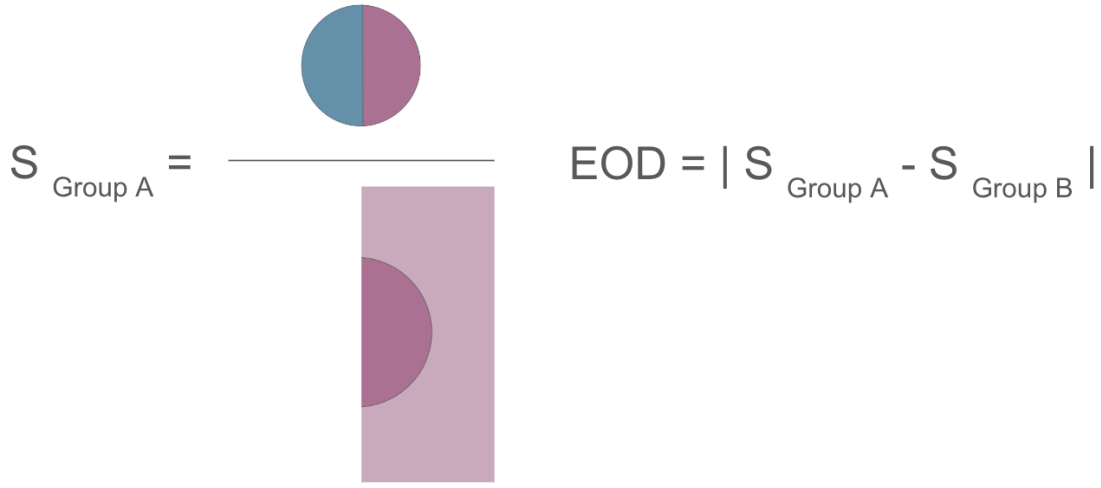

Figure 4.7: Sensitivity (S) and Equal Opportunity Difference (EOD)

4.4.3 Average Odds Difference

The Average Odds Difference (AOD) [63] computes the difference in true positive rates (TPR) and false positive rates (TPR) between demographic groups, as visualised in figures 4.8 and 4.9.

This metric reflects the likelihood of a model providing a positive outcome, considering the scenario where the positive outcome is correctly predicted, as well as when it is incorrectly predicted. In this regard, the AOD highlights the difference between groups for any given

label.

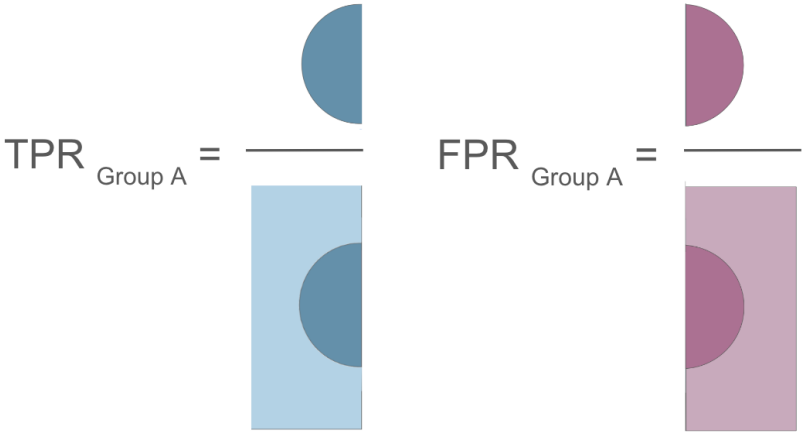

Figure 4.8: True Positive Rate (TPR) and False Positive Rate (FPR)

$$AOD = \frac{|TPR_{Group\ A} - TPR_{Group\ B}| + |FPR_{Group\ A} - FPR_{Group\ B}|}{2}$$

Figure 4.9: Average Odds Difference (AOD)

4.4.4 False Discovery Rate Difference

Next, the False Discovery Rate (FDR) [63] is the proportion of false positives in the total items classified as positive. As portrayed in figure 4.10, FDRD is the absolute difference between the FDRs for different groups and a given label.

A high FDR can be advantageous for advantageous labels, but for labels like 'Bad' or 'Very Bad', it is preferable to have a low or null FDR.

Figure 4.10: False Discovery Rate Difference (FDRD)

4.4.5 False Omission Rate Difference

Lastly, the False Omission Rate (FOR) [63] is the proportion of false negatives in the total items classified as negative. Hence, the False Omission Rate Difference (FORD) is the absolute difference of the FORs for different groups and a given label, as shown in figure 4.11.

As this is the opposite of the FDR, for advantageous labels such as 'Good', it is preferable to have a low or null FOR, in terms of privilege.

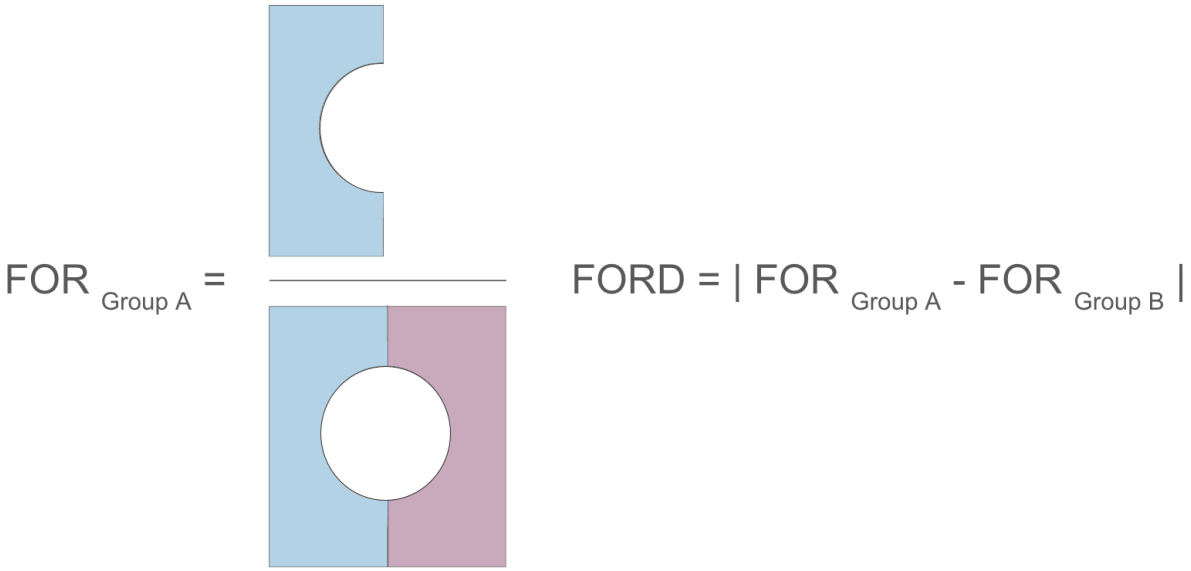

Figure 4.11: False Omission Rate Difference (FORD)

4.5 Processing of Results

Once all the metrics are computed, the results consist of one value per metric for every model, scenario, and label combination. Therefore, for the gender sensitive attribute, the results are represented by $5 * 5 * 6 * 4 = 600$ values. The same holds for the ethnicity sensitive attribute. While this is useful as it makes the study comprehensive and multi-sided, the data must be further processed to interpret it and draw conclusions from it. An overview of this process is provided in figure 4.12.

Figure 4.12: Processing of Results

The first measure taken to simplify the results is averaging the results from the five metrics for every subgroup of results. One subgroup of results refers here to the values that represent a given scenario/model/sensitive attribute/label combination. An example of such a subgroup is

the group of five values obtained for the label 'good' considering the gender attribute, and using the SVM model in the balanced-balanced scenario. **The five values, one for each metric are averaged (1)**. This is straightforward since each metric yields a value between 0 and 1; 0 representing a perfectly unbiased model, 1 representing a perfectly biased one (according to the metric in question). In this way, the amount of values is decreased by 80%, resulting in a single value, reflecting privilege, for every subgroup of results, also situated between 0 and 1. This makes the data more manageable.

Secondly, **a weighted sum of the resulting values is computed across labels (2)**, meaning that the five values (one per label) form a single value. The weights of the sum depend on the advantage each label presents; the larger the weight, the larger the privilege. The weights are shown in column two of table 4.1. The sum is computed as follows:

- Define G1 and G2 as the two groups of a sensitive attribute.
- For each label H of the 5 labels:
  - If for label L, G1 has a higher privilege value: multiply the value by the corresponding weight, and add it to the sum for G1.
  - If for label L, G2 has a higher privilege value: multiply the value obtained by the corresponding weight, and add it to to the sum for G2.
- Compute the absolute difference between the sums corresponding to G1 and G2.

Due to the choice of weights, **the resulting value is then divided by 6 to obtain a final result between 0 and 1 (3)**. This is explained through the scenario of maximum bias in figure 4.1, in which all male candidates are classified as 'Very Good' or 'Good', and all female candidates are classified as 'Bad' or 'Very Bad'. According to those values, the weighted sum for the male group is $2*1+1*1 = 3$, and for the female group it is $(-2)*1+(-1)*1 = -3$. Therefore, the absolute value of the difference between the two groups in the scenario of maximum bias is 6. The result is therefore normalised through division by 6 to obtain a more interpretable result between 0 and 1, 0 representing a total lack of bias and 1 representing 'perfect' bias. There are now only 24 values left for each sensitive attribute, and thus reduced to 6 per scenario, one corresponding to each model. In this way, comparing traditional models and LLMs in terms of bias for all scenarios becomes straightforward.

Table 4.1: Maximum Bias Scenario

Label	Weight	Value (avg. of 5 metrics)	Privileged Group
Very Bad	-2	1	M
Bad	-1	1	M
Average	0	any	any
Good	1	1	F
Very Good	2	1	F

Chapter 5

Results

5.1 Bias

5.1.1 Scenario 1: Inherent Bias

Gender:

Across all models, the amount of bias is low, though still existent, ranging from 0.001 for the SVM to 0.06 for GB and RF. This represents a range of 0.1% to 1.4% of the total theoretical possible bias. The bias in LLMs is on par with the traditional models, but the margin of error is somewhat larger. The privilege is distributed across models, four of them slightly favouring female candidates and the other two favouring male candidates. A visualisation of the results for this scenario is provided in figure 5.1. The positive values correspond to models that favour male candidates, while the negative values represent models that favour female candidates. A more comprehensive overview of the results is provided in Appendix A.

A low bias was expected as the training dataset is balanced and should, therefore, result in very little to no bias. The small amounts of bias could be attributed to accuracy-related randomness since the accuracy is not 100% and since one gender is not consistently favoured over the other.

Ethnicity:

Regarding the ethnicity sensitive attribute, the bias is also low, though significantly higher (2 to 20 times higher, depending on the model) than for the gender attribute. The bias ranges from 0.2% to 6% of the total theoretical possible bias. As figure 5.1 highlights, the BERT and GPT-3.5 Turbo models present lower bias than most traditional models. Interestingly, all traditional models favour the expected privileged group of ethnicities (WA, WE, EA), while both LLMs favour the other group.

The difference in the amount of bias for ethnicity compared to the gender attribute is surprising, as the dataset is balanced similarly for both attributes. This difference could be due to the fact that there are six ethnicities but only two genders, meaning that, per ethnicity, the models have less data to learn from. An alternative explanation could be the way that ethnicity is represented in the dataset. While both attributes are reflected first and foremost in the candidates' names, ethnicity is also somewhat connected to the candidates' educational institutions and places of work, which the models might consider as an indication of the educational prestige and quality of the candidate. This could account for the traditional models favouring the standard

'privileged' group. Another thing to address is the fact that the LLMs both favour the other group of ethnicities while having an overall much lower bias. This could be because they were pre-trained on large amounts of data, which would lessen the effect of any small bias inferred from the fine-tuning dataset. In addition, there are likely to be bias-related safeguards in these pre-trained LLMs, which could explain why the bias favours the unprivileged group; it could be a slight overcompensation of the models.

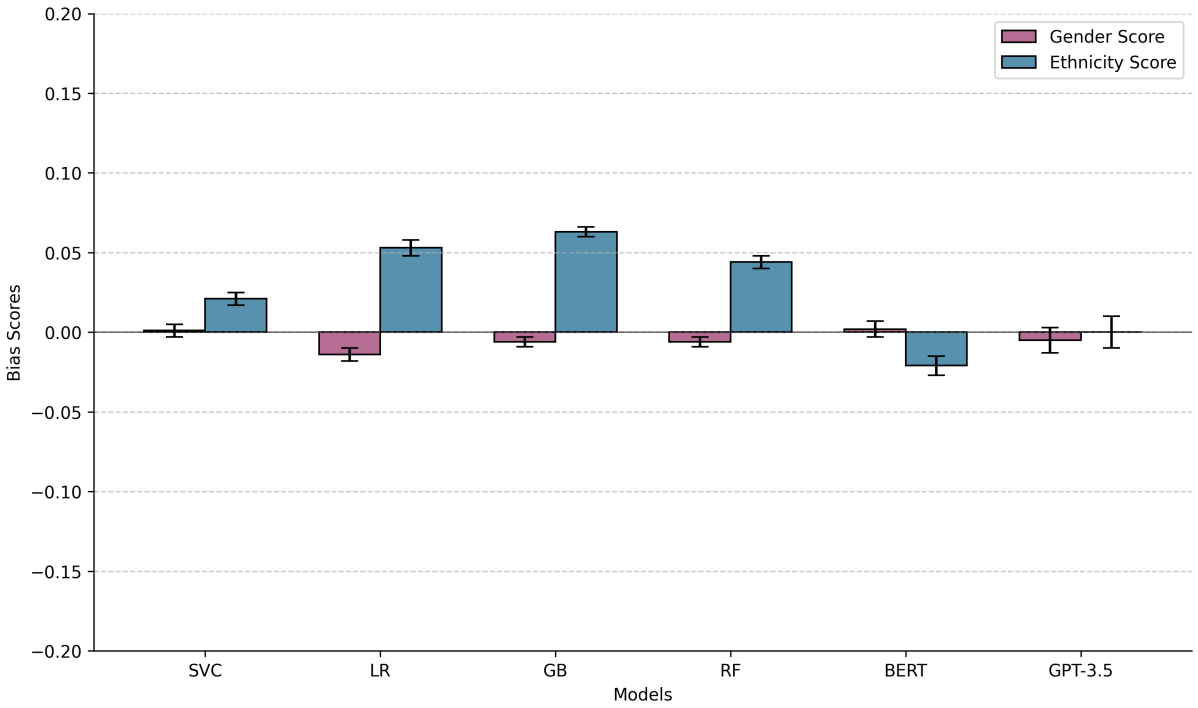

Figure 5.1: Inherent Bias Across Models

5.1.2 Scenario 2: Robustness to Bias

Gender:

This scenario induces a higher level of bias in most models. All models except for GPT-3.5 Turbo exhibit a bias of around 11 to 15% of the maximum possible bias, favouring the male candidates, as shown by the high and positive values in figure 5.2. GPT-3.5 Turbo however, has a very low bias of 0.2%

This scenario yields the expected results, showing that most models trained on a biased dataset learn that bias and transfer it to its predictions on another dataset. The fact that GPT-3.5 Turbo has a low bias, which also favours the theoretically underprivileged gender, can be accounted for through the fact that it might have stronger safeguards and compensation mechanisms for bias than BERT does. This, of course, also holds for traditional models, which do not possess any safeguard of the sort.

Ethnicity:

Similarly, the traditional models and BERT have a bias of 11 to 17% in favour of the expected privileged group (WA, WE, EA). Interestingly, the highest amount of bias is observed in BERT, whereas GPT-3.5 Turbo presents roughly 30 times less bias, which furthermore is in favour of the unprivileged ethnicities.

Similarly to the gender attribute, the expected amounts of ethnic bias are observed in most models, highlighting that models learn the bias present in the dataset, which therefore influences the predictions of the models on a balanced dataset. It should be noted that GPT-3.5 Turbo is the only model unaffected by the biased data, presenting similar amounts of bias as when trained and tested on balanced data.

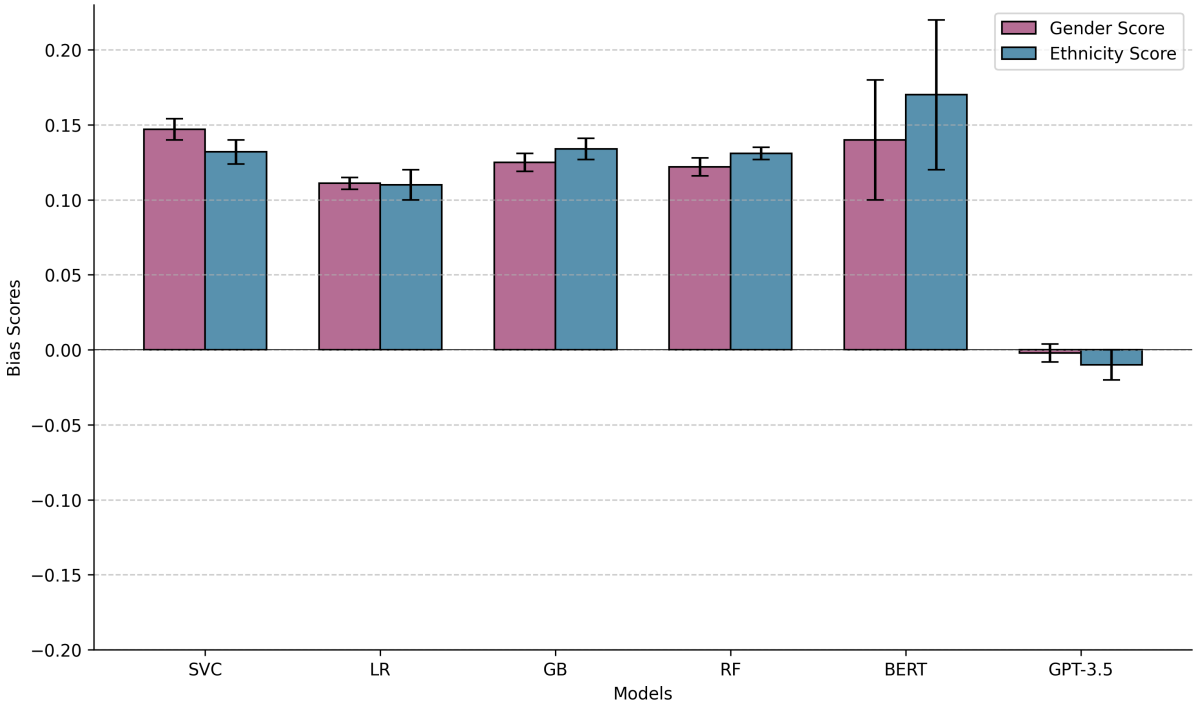

Figure 5.2: Robustness to Bias Across Models

5.1.3 Scenario 3: Application to Biased Data

A. Biased Model Applied to Biased Data

Gender:

For this scenario, the bias observed in the models ranges from 2 to 8% of the total possible bias, with BERT presenting less bias than the other models, which are all at around 7 to 8%. Interestingly, although the biased dataset is biased in favour of male candidates (as a reflection of a realistic scenario), all models favour female candidates over male ones, as highlighted in figure 5.3.

A relatively high amount of bias was expected in this scenario, but, surprisingly, this bias is opposite to the bias introduced in the dataset. The models learn a form of bias from the training data, but they don't do so perfectly, which is reflected in the overall accuracy. Therefore, the model is biased, but likely less biased than the data. During validation, the model would then slightly overestimate the underprivileged groups compared to their actual labels in the biased validation data. This could explain why the unprivileged groups come across as privileged when similarly biased data is used for training and validation.

Ethnicity:

The bias according to the ethnicity sensitive attribute in this scenario is again somewhat higher than for the gender attribute, ranging from 3% to 11%. Yet again, the group the dataset is

biased against (H, BA, AF) is, in fact, the privileged one for all models. This also applies to GPT-3.5 Turbo but not for BERT. However, both LLMs have lower bias than their traditional counterparts, though their higher margin of error should be noted.

Similarly to the gender sensitive attribute, while the bias opposite to the bias introduced in the dataset is unexpected, it can be explained by the fact that only an amount of bias is learnt by the model, leading to it overestimating the expected unprivileged group compared to the data it is tested on.

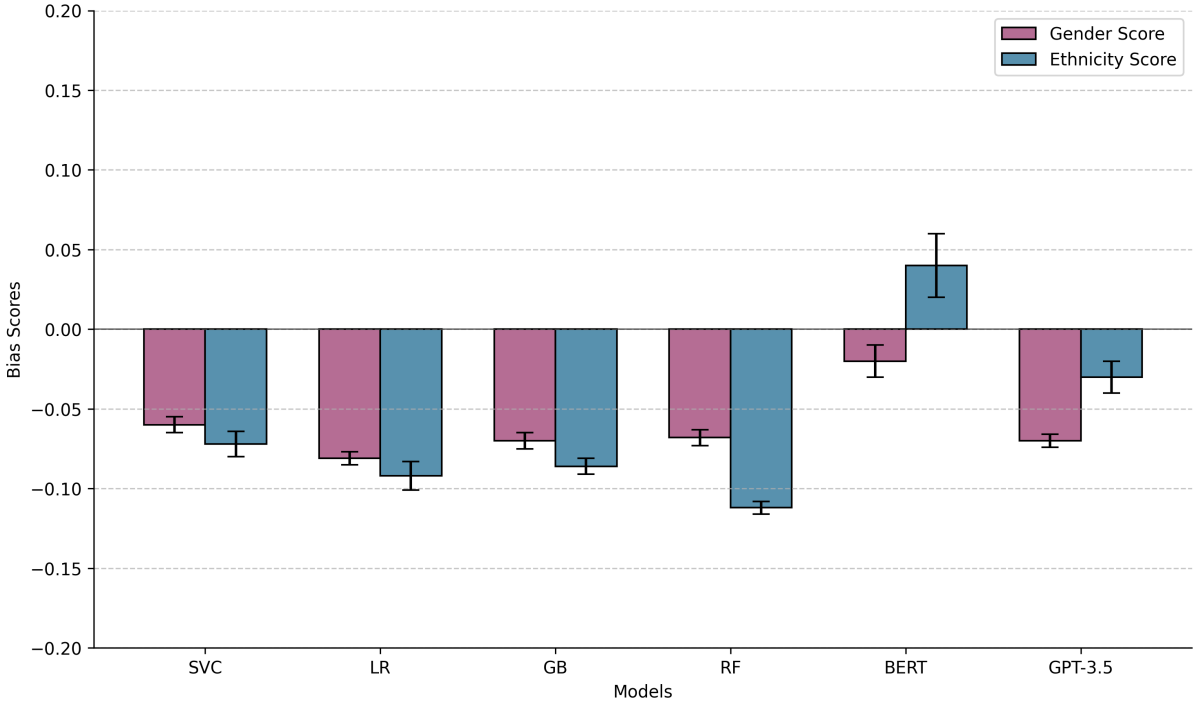

Figure 5.3: Effects of Applying a Biased Model to Biased Data

B. Unbiased Model Applied to Biased Data

Gender:

Lastly, in this scenario all six models privilege female candidates, which the dataset is, in theory, biased against. Surprisingly, the amount of bias observed in all models surpasses the bias from the previous scenario, with all models but GPT-3.5 Turbo exhibiting a bias of 17 to 19%. In this scenario, however, GPT-3.5 Turbo is not immune to biased data and also privileges female candidates by almost 6%.

Some bias was expected in this scenario, but it is surprising that it is this high despite the models being trained on unbiased data. Furthermore, this bias is opposite to the bias introduced in the dataset in all models. The pattern is too clear to be explained by randomness or general variability of the results. One plausible explanation is that the models might have learnt this bias from the dataset and, therefore, underestimated the expected unprivileged groups on the validation data. This could come across as privilege via the metrics. For example, the likelihood of a model predicting 'very good' for a female candidate would probably be higher than in the biased dataset (albeit lower than in the balanced one). This could present as a privilege for the female group in the DPD or the FDRD. This effect is exacerbated by the fact that in the process of biasing the datasets, the male group is treated in the opposite way, which probably leads to the models underestimating male candidates, which translates as gender bias against men.

Ethnicity:

Similarly, the traditional models and BERT present a bias of around 17 to 20% in favour of the (H, BA AF) candidates. As can be seen in figure 5.4, GPT-3.5 Turbo yet again has a lower bias of almost 5%.

As for the gender attribute, it is unexpected that the observed bias favours the ethnicities the dataset is supposedly biased against. As established above, a plausible explanation could be that the models' predictions reflect the unbiased training data. Yet, when tested on biased data, the groups that are expected to be underprivileged are overestimated with respect to the validation data.

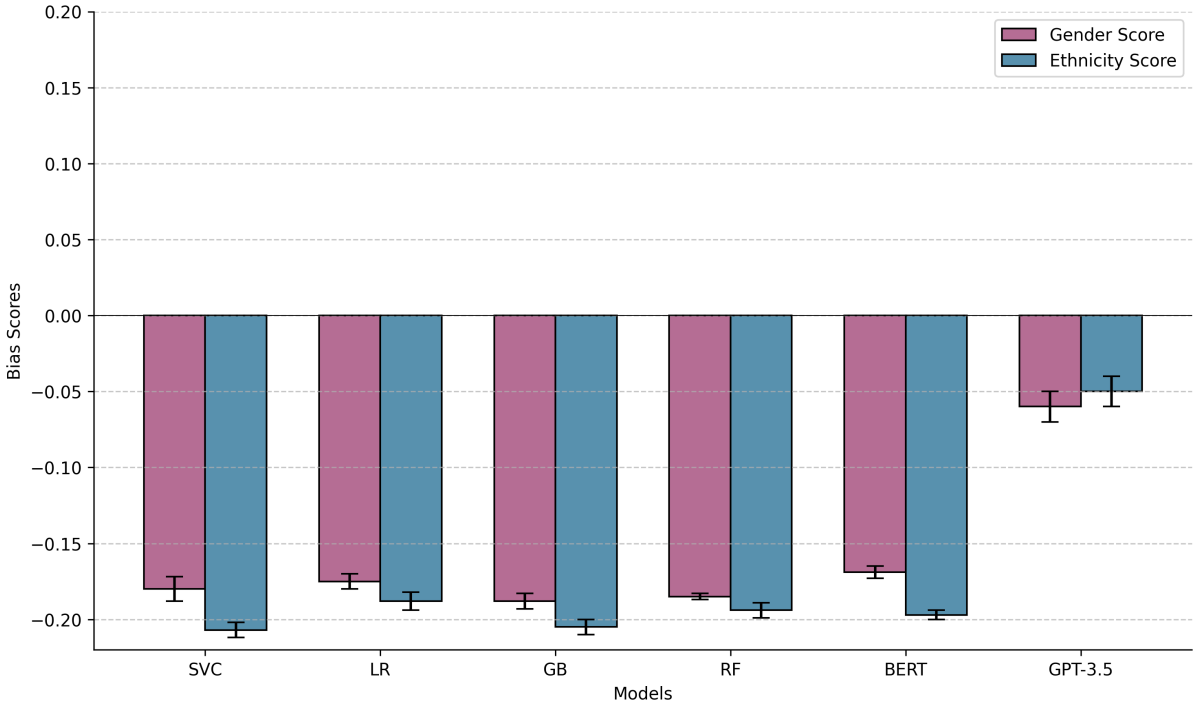

Figure 5.4: Effects of Applying an Unbiased Model to Biased Data

5.2 Performance

The accuracies of the six models for each scenario are shown in figure 5.5 for the gender attribute and in figure 5.6 for ethnicity. The complete accuracy results can be found in Appendix B. These results indicate that for Scenario 1, all models perform similarly, with an accuracy ranging from 67% to 79% for gender and from 68% to 84% for ethnicity. There is no notable difference in performance between the traditional models and the LLMs. All models perform slightly better in Scenario 2, with accuracies between 50% and 60% for 5 of them. However, GPT-3.5 Turbo is significantly ahead with an accuracy of over 70% for both sensitive attributes. The accuracies for all models decrease drastically in Scenario 3 except for GPT-3.5 Turbo, reaching their lowest point in Scenario 3.A, in which they all drop below 50% for both sensitive attributes. The accuracy of GPT-3.5 Turbo is also somewhat lower in this scenario but remains above 65%, and thus ahead of the other models by almost 20%.

Regarding accuracy-based performance, GPT-3.5 Turbo distinguishes itself from the other models, which all follow similar patterns. This also holds for BERT, which performs very similarly to the traditional models and not to GPT-3.5 Turbo as expected. Consequently, there is no

benefit to using BERT instead of one of the traditional models, especially since it is significantly more time-consuming to train (2 hours vs. a few seconds) and complex to set up and implement. However, the GPT-3.5 Turbo model has a significantly higher accuracy in all scenarios except for Scenario 1, which is not of the utmost importance since it does not represent the status quo at this time. These performance results align with the bias results; there is a clear inverse correlation between bias levels and accuracy. While GPT-3.5 Turbo is also more costly in terms of time and computation than traditional models (12 minutes to train and a cost of a few cents), the significant difference in performance, combined with the clear bias advantage, confirms that switching to GPT-3.5 Turbo for resume classification could be worth it.

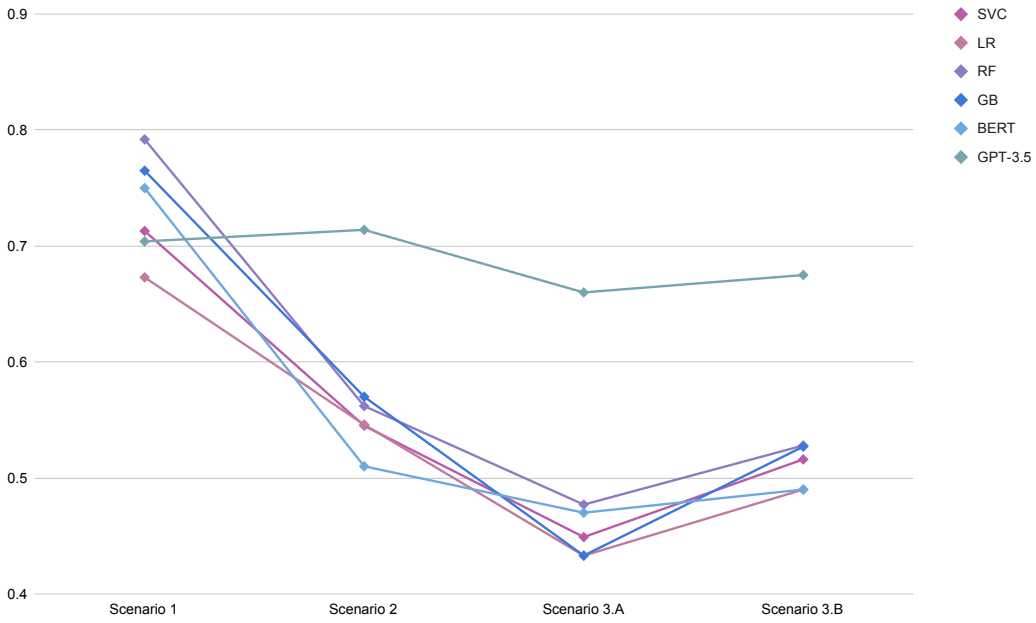

Figure 5.5: Accuracy: Gender

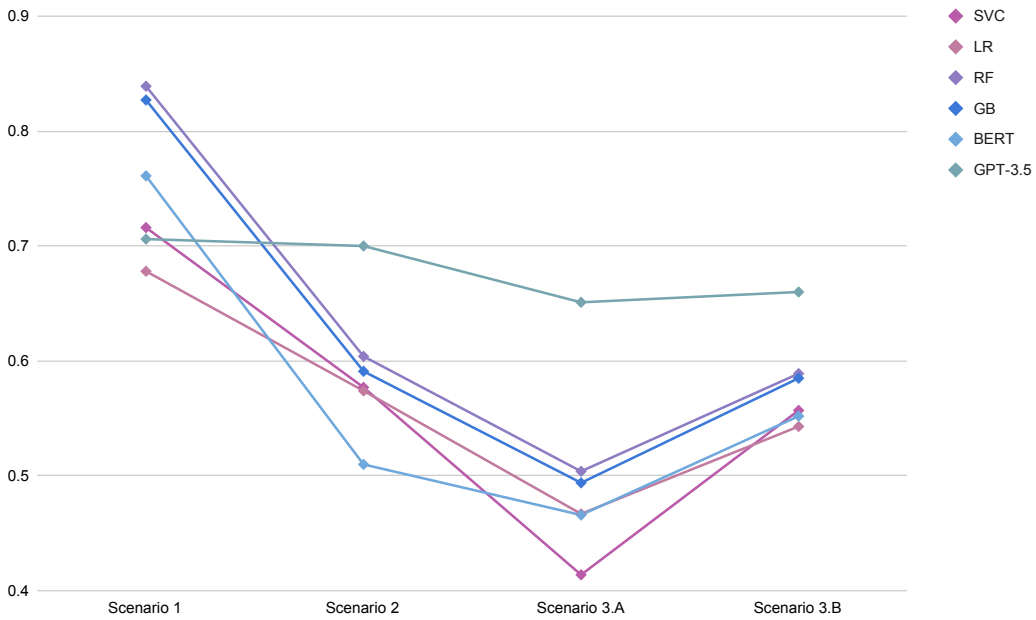

Figure 5.6: Accuracy: Ethnicity

Chapter 6

Discussion

6.1 Comparison of Traditional Models and LLMs

Overall, the four traditional models present similar levels of bias in each of the test scenarios. The bias exists but is very low when the models are trained and tested on balanced data, and it is significantly higher in all the other scenarios, which include biased data in the train/test process. This is reflected in the accuracy of the models, which plummets when bias is present in the data. Interestingly, **the expected distinction between LLMs and traditional models is not observed. Instead, an obvious distinction between the GPT-3.5 Turbo model and the other models appears. While the data visibly influences the bias in other models, GPT-3.5 Turbo appears almost immune to bias**, often having a bias of less than 1% and never reaching more than 7%. In contrast, the other models, including BERT, can reach bias values of almost 20%. This contrast between GPT-3.5 Turbo and the other models can also be confirmed through the performance of the models in the various scenarios. While GPT-3.5 Turbo does not have the highest accuracy in the balanced scenario, it also remains virtually unaffected by the introduction of bias, seeing a maximum decrease of 4%, whereas the performance of the other models unanimously drops by roughly 25%.

The results concerning the BERT model are in line with previous studies, which have shown in many contexts the presence of discriminatory bias in the outcomes generated by BERT-based architectures [16, 66, 118]. Google has, however, claimed to make efforts to prioritise fairness and avoid discrimination through their models [54], so it is somewhat surprising that the bias is just as high as in the traditional models, which do not contain any safeguards against bias. But this could indicate that the alleged safeguards compensate for the bias integrated during pre-training phases, without making up for the newly learnt bias during fine-tuning. Sure enough, BERT is one of the least biased models in the balanced data scenario. Furthermore, there has been research focused on the BERT architecture, finding that the model is susceptible to biased data, which is exacerbated by its architecture [94].

The robustness of GPT-3.5 Turbo to data-induced bias it is in line with OpenAI's declared commitment to fairness in their models [87]. The model's robustness could be explained by the extremely high number of internal parameters (20 billion), which by far surpasses the number of parameters the other models rely on, as shown in figures 6.1 and figure 6.2. This is in line with other studies that have found GPT-3.5 Turbo to be resilient to bias [103, 114], especially for the gender and ethnicity sensitive attributes that are evaluated

here. Nevertheless, many studies have found non-negligible amounts of bias in GPT-based models [8, 11, 118, 78, 117]. This contrast in findings can be attributed to many factors, ranging from the type of task and data evaluated, to the high sensitivity GPT-3.5 Turbo has to the prompts used [8, 29].

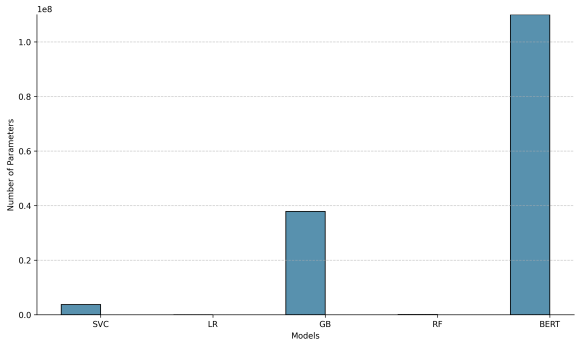

Figure 6.1: Number of Parameters Without GPT-3.5 Turbo

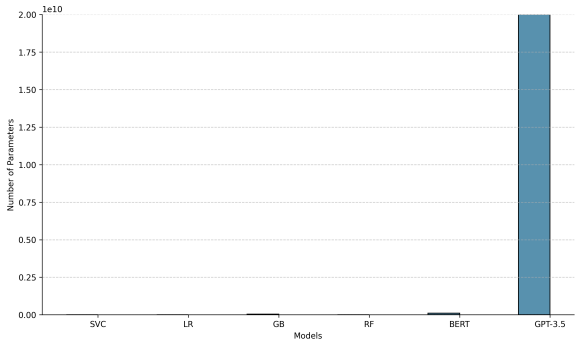

Figure 6.2: Number of Parameters Including GPT-3.5 Turbo

6.2 Impact of the Data

As expected, the balanced dataset translates to a near lack of bias in all models, confirming that there is **no significant bias inherent to these models**. Except GPT-3.5 Turbo, all models see a strong increase in bias when trained on biased data. A dataset biased against female resumes gives rise to a model with gender-biased predictions, putting female candidates at a disadvantage. Identically, a dataset biased against Black, Hispanic, and African individuals leads to a model biased against those same ethnicities and in favour of the other ones. As expected, this confirms the impact of the data in a model's fairness and the importance of prioritising the creation and use of balanced datasets free of prejudice [80].

Another interesting point is the effect of testing the models on biased data, reflecting a real-life scenario in which a model would be applied to a set of resumes that had already been processed in a biased manner, likely in a previous step of the hiring process. This setup leads to the highest bias for all models (even GPT-3.5 Turbo to a lesser extent). Surprisingly, the bias is consistently in favour of the demographic groups that are disadvantaged in the biased datasets. This holds for models trained on balanced data and models trained on biased data. This scenario has not been researched, and there is no proven explanation for this phenomenon. However, it can be hypothesized that the models do not learn the presented bias perfectly (reflected in the low accuracies), and are therefore less biased than the dataset they are tested on. As a result, the biased predictions will still appear to overestimate the populations that the test dataset is biased against, resulting in a bias opposite to the expected one. **This highlights the importance of ensuring that not only is the training data fair and devoid of biases, but that the data hiring models are applied to have not been compromised by bias in previous steps of the hiring process** [18].

Lastly, it should also be noted that general claims about bias cannot be made from the study of a single type of bias. While this research was focused on only two sensitive attributes, there were already clear differences in the bias levels. Despite both biases being introduced in the same manner and the same amount in the datasets, the ethnic bias was consistently higher than the gender bias across the different scenarios. This is in line with findings claiming that ethnic bias is more difficult to eliminate than gender bias [9]. The difference in bias between the two sensitive attributes could also result from the amount of data for each (720 resumes per gender,

but only 240 per ethnicity) and how the resume data reflects these attributes. Beyond ethnicity and gender, other types of bias are even more prevalent [114], possibly because safeguards and mitigation strategies for biases around which there is less awareness are less researched and implemented. Consequently, **such studies should be extended to more types of bias in order to better understand and eliminate them.** Even more so, this draws attention to intersectional biases, which are often more complex and often overlooked [111, 56].

6.3 Recommendations

This study highlights the impact of data on discrimination in resume classification and, more generally, in AI-based hiring. As expected, the findings show that **regardless of the choice of model, the ideal scenario is training it on balanced data, devoid of social biases, and applying it to data that was not previously corrupted by bias**, which could happen during preceding phases of the hiring process (such as advertisement of the job offer, application fee, etc.). If balanced data is ensured, the choice of model is not highly relevant.

Another key takeaway is that **a more complex and novel model is not always better in terms of fairness or performance**. The findings from this study show that there is no benefit to using BERT for resume classification because it presents very similar bias levels and performance scores as more traditional models. Furthermore, BERT is significantly more complex to implement and more time-consuming to train; it takes almost two hours, contrary to traditional models, which attain the same results within seconds.

However, while GPT-3.5 Turbo is also more complex to implement and more computationally expensive than traditional models (still considerably faster to train than BERT), the trade-off might be worth considering since it is significantly less biased and more performant than all other models in this study. This especially holds in scenarios where balanced data cannot be guaranteed, as GPT-3.5 Turbo is exceptionally robust to bias, whether in training data or in the data the model is ultimately applied to. While balanced data is preferable, it is not always a possibility, and consequently, **using GPT-3.5 Turbo for resume classification is the safest choice to ensure fairness in the hiring process, without compromising performance.**

6.4 Limitations and Directions for Future Research

The limitations of this study are mostly related to its size; the results obtained, while relevant, remain limited in scope, and it would be useful to extend the study to obtain more general conclusions. First, two LLMs are insufficient to draw conclusions about this type of model, especially given the large bias difference observed between BERT and GPT-3.5 Turbo. **Including other state-of-the-art models** such as Llama, Mistral, Gemini and other varieties of BERT and GOT would be useful. Furthermore, the findings and existing works suggest that all biases are not created equal, and it is impossible to extrapolate information about all biases from just a few. For this reason, there is a **need to study a larger variety of biases**, especially less known ones, including intersectional biases. **More resources** could also help provide more comprehensive and reliable results. In this case, having more than five or ten trials for each model would be preferable, as would a larger dataset. While the results across trials are relatively consistent, and therefore, the patterns observed here are unlikely to change drastically, more resources could lead to more nuanced conclusions and better insight into explanations of those results.

While resume classification is a core task in AI-based hiring, it is not the only one, and it would be useful to **undertake a similar study focused on other tasks from the hiring process** such as interview analysis, social media screening, or cognitive assessments.

While using five metrics provides reliable results, the chosen metrics only reflect group fairness. Although this was a deliberate choice in this study, individual and subgroup fairness should also be explored. Additionally, these metrics are merely mathematical approximations of more complex social phenomena. Hence, it could be beneficial to **complement the metrics by another form of assessment, such as a form of human bias assessment.**

Bias detection in AI-based hiring is in itself an exciting topic to study, but its primary relevance is its contribution to understanding and confronting bias. Therefore, the next logical steps are to **extend the research on bias mitigation strategies and create accessible and efficient tools that employers and companies can and want to implement in the models that intervene in their hiring processes**.

Chapter 7

Conclusion

Regarding hiring-related processes, such as resume classification, AI has been the status quo for years in most companies due to its efficiency and high performance on a myriad of different tasks. The social and ethical implications of automating important decisions have been vastly studied, and by now, the dangers of relying so heavily on AI are well-known. However, the AI landscape is changing rapidly and drastically with the development of LLMs and their wide adoption rates. Given the novelty and complexity of these models, it is becoming increasingly important to be aware of their limitations and to actively search for ways to ensure that they are safe, ethical, and fair.

One crucial issue in AI-based hiring is discrimination, which often arises from the common use of biased training data. This bias reflects stereotypes prevalent in our society (gender bias, racial bias, etc.), which are then learnt by the model and perpetuated through its outcomes. This study aimed to evaluate such biases in LLMs (specifically BERT and GPT-3.5 Turbo) and to compare them to the amounts of biases in more traditional models (SVC, LR, RF, and GB), specifically in resume classification. Ultimately, the goal was to determine whether transitioning to novel and 'trendy' LLM-based approaches in hiring tasks is wise from a fairness perspective. More concretely, this study attempted to compare the two types of model, in terms of inherent bias and bias learnt from the data, in order to provide a recommendation to companies wondering whether the switch to LLMs is the best choice.

To this end, three synthetic datasets were generated due to a lack of appropriate data for this study. The first dataset was fully balanced in terms of gender and ethnicity, thus serving as a baseline and a means to measure the inherent bias present in the models. A predefined amount of bias was introduced in the other two datasets, resulting in one dataset containing gender bias and another containing ethnic bias. The six models were tested in three different scenarios, aiming to highlight the inherent bias in the models, their robustness to biased training data, and the effect of applying them to previously biased data.

The findings indicate that all models present extremely low levels of inherent bias, as can be seen when trained and tested on balanced data. However, when trained on biased data, all models, except GPT-3.5 Turbo, are found to be highly biased in their predictions, discriminating against the same demographic groups that the training data is biased against. These findings are somewhat in line with existing literature. However, a surprising phenomenon was observed when the models were tested on biased data, reflecting scenarios in which a model is applied to a group of applications already processed or selected through biased processes. In these test cases, high levels of bias opposite to the bias present

in the datasets were observed for all models (though significantly lower for GPT-3.5 Turbo). **This highlights the need to prioritise fairness along the whole hiring process because one biased step can suffice to corrupt ulterior outcomes**, even when a fair and unbiased model is used. Another takeaway from this study is the difference in bias levels corresponding to the two sensitive attributes, despite bias being introduced in the same manner and amount in both datasets. This shows **the need to consider as many sensitive attributes as possible** in fairness studies. Lastly, the findings are less binary than expected; the expected clear distinction between traditional models and LLMs is not observed. Rather, **all models are heavily affected by biased data except for GPT-3.5 Turbo, which is highly robust to bias. Out of the models included in this study, GPT-3.5 Turbo is, therefore, by far the best choice in terms of avoiding discrimination in resume classification**. It does, however, require more time and computational resources, but not to an extent that could be problematic at the level of a company (unlike other LLMs such as BERT).

This study does have limitations that must also be mentioned. Many of them are related to the size of the study; more time and resources would be necessary for a more comprehensive comparison, with more models, data, sensitive attributes, metrics, and, of course, a larger number of trials for more reliable results. This topic could be further investigated by looking into other hiring-related tasks such as interview analysis or cognitive assessments. And, of course, research on bias avoidance or mitigation strategies is also welcome and very much needed.

At a time when AI is increasingly entrusted with important decisions, research on bias detection and mitigation is essential. **This thesis points out the impact of biased training data and the importance of selecting models wisely. It also highlights the need to ensure fairness throughout all steps of the hiring process.** Moving forward, we must ensure that the AI processes we rely on contribute positively to society beyond their performance and efficiency. To leverage the power of AI without causing harm or infringing on human rights, we have to prioritise fairness, transparency, and accountability. This thesis attempted to take a modest step in that direction, and my hope is that it will inspire further research and rapid progress in this area.

Appendix A

Full Bias Results for Each Scenario

Table A.1: Scenario 1: Inherent Bias

	Score Gender	Privileged Group	Score Ethnicity	Privileged Group
SVC	0.001 ± 0.004	M	0.021 ± 0.004	WA, WE, EA
LR	0.014 ± 0.004	F	0.053 ± 0.005	WA, WE, EA
GB	0.006 ± 0.003	F	0.063 ± 0.003	WA, WE,EA
RF	0.006 ± 0.003	F	0.044 ± 0.004	WA, WE, A
BERT	0.002 ± 0.005	M	0.021 ± 0.006	H, BA, AF
GPT-3.5	0.005 ± 0.008	F	0.00 ± 0.01	H, BA, AF

Table A.2: Scenario 2: Robustness to Bias

	Score Gender	Privileged Group	Score Ethnicity	Privileged Group
SVC	0.147 ± 0.007	M	0.132 ± 0.008	WA, WE,EA
LR	0.111 ± 0.004	M	0.11 ± 0.01	WA, WE,EA
GB	0.125 ± 0.006	M	0.134 ± 0.007	WA, WE,EA
RF	0.122 ± 0.006	M	0.131 ± 0.004	WA, WE,EA
BERT	0.14 ± 0.04	M	0.17 ± 0.05	WA, WE,EA
GPT-3.5	0.002 ± 0.006	F	0.01 ± 0.01	H, BA, AF

Table A.3: Scenario 3.A: Biased Model Applied to Biased Data

	Score Gender	Privileged Group	Score Ethnicity	Privileged Group
SVC	0.060 ± 0.005	F	0.072 ± 0.008	H, BA, AF
LR	0.081 ± 0.004	F	0.092 ± 0.009	H, BA, AF
GB	0.070 ± 0.005	F	0.086 ± 0.005	H, BA, AF
RF	0.068 ± 0.005	F	0.112 ± 0.004	H, BA, AF
BERT	0.02 ± 0.01	F	0.04 ± 0.02	WA, WE, AF
GPT-3.5	0.070 ± 0.004	F	0.03 ± 0.01	H, BA, AF

Table A.4: Scenario 3.B: Unbiased Model Applied to Biased Data

	Score Gender	Privileged Group	Score Ethnicity	Privileged Group
SVC	0.180 ± 0.008	F	0.207 ± 0.005	H, BA, AF
LR	0.175 ± 0.005	F	0.188 ± 0.006	H, BA, AF
GB	0.188 ± 0.005	F	0.205 ± 0.005	H, BA, AF
RF	0.185 ± 0.002	F	0.194 ± 0.005	H, BA, AF
BERT	0.169 ± 0.004	F	0.197 ± 0.003	H, BA, AF
GPT-3.5	0.06 ± 0.01	F	0.05 ± 0.01	H, BA, AF

Appendix B

Full Accuracy Results for each Scenario

Table B.1: Scenario 1: Inherent Bias

	Accuracy Gender	Accuracy Ethnicity
SVC	0.713 ± 0.008	0.716 ± 0.008
LR	0.673 ± 0.006	0.678 ± 0.008
GB	0.765 ± 0.007	0.827 ± 0.003
RF	0.792 ± 0.007	0.839 ± 0.005
BERT	0.75 ± 0.04	0.761 ± 0.009
GPT-3.5	0.704 ± 0.004	0.706 ± 0.005

Table B.2: Scenario 2: Robustness to Bias

	Accuracy Gender	Accuracy Ethnicity
SVC	0.545 ± 0.006	0.577 ± 0.008
LR	0.546 ± 0.007	0.574 ± 0.005
GB	0.57 ± 0.01	0.591 ± 0.006
RF	0.562 ± 0.008	0.604 ± 0.009
BERT	0.51 ± 0.03	0.51 ± 0.03
GPT-3.5	0.714 ± 0.006	0.70 ± 0.01

Table B.3: Scenario 3.A: Biased Model Applied to Biased Data

	Accuracy Gender	Accuracy Ethnicity
SVC	0.449 ± 0.005	0.414 ± 0.007
LR	0.433 ± 0.007	0.467 ± 0.008
GB	0.433 ± 0.006	0.494 ± 0.005
RF	0.477 ± 0.009	0.504 ± 0.005
BERT	0.47 ± 0.01	0.466 ± 0.009
GPT-3.5	0.66 ± 0.01	0.651 ± 0.004

Table B.4: Scenario 3.A: Unbiased Model Applied to Biased Data

	Accuracy Gender	Accuracy Ethnicity
SVC	0.516 ± 0.009	0.557 ± 0.006
LR	0.490 ± 0.006	0.543 ± 0.006
GB	0.527 ± 0.003	0.585 ± 0.005
RF	0.528 ± 0.004	0.589 ± 0.002
BERT	0.49 ± 0.02	0.552 ± 0.002
GPT-3.5	0.675 ± 0.005	0.660 ± 0.006

Appendix C

Dataset Sample

Name	Gender	Email	Ethnicity	Education	GPA	Work Experience	Skills	Awards	Fit	Job
Zara Abubakar	female	zara.abubakar@email	AF	Zara completed her hairdressing degree at the New York School of Cosmetology.	3.3	Zara has worked at a busy salon in New York City for three years, where she honed his skills in precision cutting and advanced color techniques. She is known for her attention to detail and ability to customize styles to suit each client's needs.	Proficient in highlighting, texturizing, and men's grooming. Excellent problem-solving skills and ability to work efficiently in a fast-paced environment.	Zara received the 'Outstanding Stylist' award at the New York Hair Expo for her exceptional talent and dedication to the craft.	average	Hairdresser
Tyrone Williams	male	tyrone.williams@example.com	BA	Associate Degree in Cosmetology from New York Beauty College	3.2	Tyrone has 3 years of experience working as a hairdresser at Trendy Cuts Salon, specializing in men's haircuts and grooming services.	Skilled in fades, beard trims, and classic barbering techniques. Excellent customer service and interpersonal skills. Knowledgeable about current hair trends.	Tyrone was recognized as a top hairstylist at the New York Hair and Beauty Show in 2020.	good	Hairdresser
Latasha Bennett	female	lbennett@email.com	BA	Bachelor's in Chemical Engineering from State University	2.9	Worked as a laboratory assistant during college, assisting with experiments and data analysis.	Basic knowledge of chemical processes, laboratory techniques, attention to detail.	Dean's List recognition for academic achievement	bad	Engineer
Roberto Nunez	male	robertonunez@email.com	L	Diploma in Office Management from Miami Technical College	2.7	Office Clerk at a law firm for 1 year, assisting with filing, copying documents, and organizing case files.	Knowledge of legal terminology, detail-oriented, proficiency in data entry.	Certificate of Completion for professional development training.	average	Secretary
Sydney Palmer	female	sydney.palmer@example.com	WA	Master's in Materials Science and Engineering, University of Michigan	3.8	Sydney has 6 years of experience in materials research, focusing on advanced materials for aerospace applications. She has a proven track record of leading interdisciplinary teams in research projects.	Expert in materials characterization techniques, project management skills, strong technical writing abilities.	Sydney was recognized with the Research Excellence Award for her contributions to the development of a high-performance composite material.	very good	Engineer

Figure C.1: Resume Dataset Sample

Bibliography

[1] Scikit learn library.

[2] Elias Abdollahnejad, Marilynn Kalman, and Behrouz H. Far. A deep learning bert-based approach to person-job fit in talent recruitment. *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec 2021.

[3] Arpita Agnihotri and Saurabh Bhattacharya. *Artificial Intelligence for hiring and induction: The unilever experience*, 2024.

[4] Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in bert. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[5] Ibrahim M Alabdulmohsin, Jessica Schrouff, and Sanmi Koyejo. A reduction to binary approach for debiasing multiclass datasets. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2480–2493. Curran Associates, Inc., 2022.

[6] Elham Albaroudi, Taha Mansouri, and Ali Alameer. A comprehensive review of ai techniques for addressing algorithmic bias in job hiring. *AI Systems: Theory and Applications*, 5(1):383–404, 2024.

[7] Karim Amzile, Mohamed Beraich, Imane Amouri, and Cheklekbire Malainine. Towards a digital enterprise: the impact of artifical intelligence on the hiring process. *Journal of Intelligence Studies in Business*, 13(3):18–26, 2023.

[8] Haozhe An, Christabel Acquaye1, Colin Kai Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? *Computation and Language*, Jun 2024.

[9] Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. Measuring gender and racial biases in large language models, 2024.

[10] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese. In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI amp; SOCIETY*, 35(3):611–623, Jan 2020.

[11] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. The silicon ceiling: Auditing gpt's race and gender biases in hiring. *Computers and Society*, May 2024.

[12] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Measuring implicit bias in explicitly unbiased large language models, 2024.

[13] Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal*, 8(2), Jul 2021.

[14] Jack Bandy. Problematic machine behavior. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–34, Apr 2021.

[15] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. 2018.

[16] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018, May 2021.

[17] Sarah Bird, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. 2020.

[18] Miranda Bogen and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Technical report, Upturn, 2018.

[19] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[21] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[22] Ian Burke, Robin Burke, and Goran Kuljanin. Fair candidate ranking with spatial partitioning: Lessons from the SIOP ML competition. In Mesut Kaya, Toine Bogers, David Graus, Katrien Verbert, and Francisco Gutiérrez, editors, *Proceedings of the Workshop on Recommender Systems for Human Resources (RecSys in HR 2021) co-located with the 15th ACM Conference on Recommender Systems (RecSys 2021), Amsterdam, The Netherlands, 27th September - 1st October 2021*, volume 2967 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[23] Longbing Cao. Ai in finance: A review. *SSRN Electronic Journal*, 2020.

[24] Ishita Chakraborty, Khai Chiong, Howard Dover, and K. Sudhir. Ai and ai-human based salesforce hiring using interview videos. *SSRN Electronic Journal*, 2022.

[25] Ashish Virendra Chandak, Hardik Pandey, Gourav Rushiya, and Harsh Sharma. Resume parser and job recommendation system using machine learning. *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, Feb 2024.

[26] Xinyu Chang. Gender bias in hiring: An analysis of the impact of amazon's recruiting algorithm. *Advances in Economics, Management and Political Sciences*, 23(1):134–140, Sep 2023.

[27] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *Association for Computing Machinery*, 2024.

[28] Zhisheng Chen. Collaboration among recruiters and artificial intelligence: Removing human prejudices in employment. *Cognition, Technology amp; Work*, 25(1):135–149, Sep 2022.

[29] Benjamin Clavié, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, and Thomas Brightwell. Large language models in the workplace: A case study on prompt engineering for job type classification. *Natural Language Processing and Information Systems*, page 3–17, 2023.

[30] European Commission. What personal data is considered sensitive?, 2024. Accessed: 2024-06-30.

[31] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[32] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–242, 1958.

[33] David Danks and Alex John London. Algorithmic bias in autonomous systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Aug 2017.

[34] Asmita Deshmukh and Anjali Raut. Applying bert-based nlp for automated resume screening and candidate ranking. *Annals of Data Science*, Mar 2024.

[35] Ketki V. Deshpande, Shimei Pan, and James R. Foulds. Mitigating demographic bias in ai-based resume filtering. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, Jul 2020.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[37] V. V. Dixit, Trisha Patel, Nidhi Deshpande, and Kamini Sonawane. Resume sorting using artificial intelligence. *International Journal of Research in Engineering, Science and Management*, 2(4), Apr 2019.

[38] Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. Disclosure and mitigation of gender bias in llms. 2024.

[39] Eleanor Drage and Kerry Mackereth. Does ai debias recruitment? race, gender, and ai's "eradication of difference". *Philosophy Technology*, pages 35–89, 2022.

[40] Yingpeng Du, Di Luo, Rui Yan, Hongzhi Liu, Yang Song, Hengshu Zhu, and Jie Zhang. Enhancing job recommendation through llm-based generative adversarial networks. *Information Retrieval*, 2023.

[41] European Commission. Data protection working party. directive 95/46/ec § article 29, 2018. Accessed on [June 15th 2024].

[42] European Parliament and Council. General data protection regulation (gdpr). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX Recital 51, Regulation (EU) 2016/679.

[43] European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

[44] European Parliament Think Tank. Artificial intelligence act: Proposal for a regulation laying down harmonised rules on artificial intelligence. European Parliament Think Tank Briefing, 2021.

[45] European Union. Article 21 - non-discrimination. https://fra.europa.eu/en/eu-charter/article/21-non-discrimination: :text=Any Accessed: 2024-06-28.

[46] Fairlearn. Common fairness metrics. https://fairlearn.org/main/user$_guide/assessment/common_fair$ $2024 - 06 - 28$.

[47] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, Dec 2023.

[48] Xavier Ferrer, Tom van Nuenen, Jose M. Such, Mark Cote, and Natalia Criado. Bias and discrimination in ai: A cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2):72–80, Jun 2021.

[49] J. H. Friedman. Greedy function approximation: A gradient boosting machine. Technical report, Department of Statistics, Stanford University, 1999. This paper laid the groundwork for the gradient boosting framework.

[50] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. 2024.

[51] Chegguang Gan, Qinghao Zhang, and Tastunori Mori. Application of llm agents in recruitment: A novel framework for resume screening. 2024.

[52] Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. Ai pitfalls and what not to do: Mitigating bias in ai. *The British Journal of Radiology*, 96(1150), Sep 2023.

[53] Michael Gira, Ruisu Zhang, and Kangwook Lee. Debiasing pre-trained language models via efficient fine-tuning. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar, editors, *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[54] Google. Responsibility: Our principles.

[55] Mayuri Gund, Divyashree Chavan, Sourabh Magar, Shreya Ghadage, and Rohini Jadhav. Transforming hr practices: Resume ranking using bert embeddings. *International Research Journal of Modernization in Engineering Technology and Science*, 5, 2023.

[56] Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Jul 2021.

[57] Maryam Amir Haeri and Katharina Anna Zweig. The crucial role of sensitive attributes in fair classification. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec 2020.

[58] Abid Haleem, Mohd Javaid, Mohd Asim Qadri, Ravi Pratap Singh, and Rajiv Suman. Artificial intelligence (ai) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 3:119–132, 2022.

[59] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323, 2016.

[60] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, Aug 2023.

[61] Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. Don't judge me by my face : An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews, 2021.

[62] Lennart Hofeditz, Sünje Clausen, Alexander Rieß, Milad Mirbabaie, and Stefan Stieglitz. Applying xai to an ai-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets*, 2022.

[63] IBM Cloud. Fairness metrics overview. https://dataplatform.cloud.ibm.com/docs/content/wsj/mode fairness-metrics-ovr.html?context=cpdaas. Accessed: 2024-06-28.

[64] Index Mundi. Netherlands demographics profile, 2021.

[65] Index Mundi. Usa demographics profile, 2021.

[66] Sophie Jentzsch and Cigdem Turan. Gender bias in bert - measuring and analysing biases through sentiment rating in a realistic downstream classification task. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 2022.

[67] Ratnesh Kumar Joshi, Arindam Chatterjee, and Asif Ekbal. Saliency guided debiasing: Detecting and mitigating biases in lms using feature attribution. 2024.

[68] Elisabeth K. Kelan. Algorithmic inclusion: Shaping the predictive algorithms of artificial intelligence in hiring. *Human Resource Management Journal*, Apr 2023.

[69] Aislinn Kelly-Lyth. Challenging biased hiring algorithms. *Oxford Journal of Legal Studies*, 41, 2021.

[70] Yeqing Kong and Huiling Ding. Tools, potential, and pitfalls of social media screening: Social profiling in the era of ai-assisted recruiting. *Journal of Business and Technical Communication*, 38(1):33–65, Sep 2023.

[71] Girish L, Raviprakash M L, Gurushankar H B, Kotramma Mathada, and Merlin B. Intelligent resume scrutiny using named entity recognition with bert. *International Conference on Data Science and Network Security (ICDSNS)*, pages 1–8, 2023.

[72] Max Langekamp, Allan Costa, and Chris Cheug. Hiring fairly in the age of algorithms. *Human-Computer Interaction*, 2020.

[73] Byoung Chol Lee and Bo-Young Kim. Development of an ai-based interview system for remote hiring. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 12(3):654–663, Mar 2021.

[74] Bingbing Li, Hongwu Peng, Rajat Sainju, Junhuan Yang, Lei Yang, Yueying Liang, Weiwen Jiang, Binghui Wang, Hang Liu, and Caiwen Ding. Detecting gender bias in transformer-based models: A case study on bert,,,, lei yang,,,, liu,. *Computation and Language*, Oct 2021.

[75] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. 2023.

[76] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654, Mar 2022.

[77] Morgan Livingston. Preventing racial bias in federal ai. *Journal of Science Policy amp; Governance*, 16(02), May 2020.

[78] Li Lucy and David Bamman. Gender and representation bias in gpt-3 generated stories. *Proceedings of the Third Workshop on Narrative Understanding*, 2021.

[79] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

[80] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, Jul 2021.

[81] Sushruta Mishra, Pradeep K Mallik, Hrudaya K Tripathy, Lambodar Jena, and Gyoo-Soo Chae. Stacked knn with hard voting predictive approach to assist hiring process in it organizations. *International Journal of Electrical Engineering Education*, 2021.

[82] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PLOS ONE*, 15(8), Aug 2020.

[83] Rizwan Qureshi Abbas Shah amgad muneer Muhammad Irfan Anas Zafar Muhammad Bilal Shaikh Naveed Akhtar Jia Wu Seyedali Mirjalili Mubarak Shah Muhammad Usman Hadi, qasem al tashi. A survey on large language models: Applications, challenges, limitations, and practical usage. 2023.

[84] Sydney Myers. 2023 applicant tracking system (ats) usage report: Key shifts and strategies for job seekers, May 2024.

[85] Selin E. Nugent and Susan Scott-Parker. Recruitment ai has a disability problem: Anticipating and mitigating unfair automated hiring decisions. *Intelligent Systems, Control and Automation: Science and Engineering*, page 85–96, 2022.

[86] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016.

[87] OpenAI. How should ai systems behave, and who should decide?, 2023.

[88] Devah Pager, Bruce Western, and Bart Bonikowski. Discrimination in a low-wage labor market: A field experiment. *American Sociological Review*, 74(5):777–799, 2009.

[89] Carlos M. Parra, Manjul Gupta, and Denis Dennehy. Likelihood of questioning ai-based recommendations due to perceived racial/gender bias. *IEEE Transactions on Technology and Society*, 3(1):41–45, 2022.

[90] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. Investigations of performance and bias in human-ai teamwork in hiring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12089–12097, Jun 2022.

[91] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. 55(3), 2022.

[92] Prasanna Kumar R, Rithani M, Bharathi Mohan G, and Venkatakrishnan R. Empirical evaluation of large language models in resume classification. *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Jan 2024.

[93] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: evaluating claims and practices. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

[94] Maryam Ramezanzadehmoghadam, Hongmei Chi, Edward L. Jones, and Ziheng Chi. Inherent discriminability of bert towards racial minority associated data. *Computational Science and Its Applications – ICCSA 2021*, page 256–271, 2021.

[95] M. S. Mohamad Raub. Bots, bias and big data: Artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. 2018.

[96] P. V. Raveendra, Y. M. Satish, and Padmalini Singh. Changing landscape of recruitment industry: A study on the impact of artificial intelligence on eliminating hiring bias from recruitment and selection process. *Journal of Computational and Theoretical Nanoscience*, 17(9):4404–4407, Jul 2020.

[97] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023.

[98] Jahan Mohan Reddy, Sirisha Regella, and Srinivasa Reddy Seelam. Recruitment prediction using machine learning. *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, pages 1–4, 2020.

[99] Jonas Rieskamp, Lennart Hofeditz, Milad Mirbabaie, and Stefan Stieglitz. Approaches to improve fairness when deploying ai-based algorithms in hiring – using a systematic literature review to guide future research. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2023.

[100] CPRW Robert Henderson. Applicant tracking systems: Everything you need to know, May 2024.

[101] Arthur Sakamoto, Isao Takei, and Hyeyoung Woo. The myth of the model minority myth. *Sociological Spectrum*, 32, 07 2012.

[102] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. 2019.

[103] Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. Addressing cognitive bias in medical language models, 2024.

[104] Joel Silas, Prajakta Udhan, Pranali Dahiphale, Vaibhav Parkale, and Poonam Lambhate. Automation of candidate hiring system using machine learning. *International Journal of Innovative Science and Research Technology*, 8, 2023.

[105] Vasudha Singh. Exploring the role of large language model (llm)-based chatbots for human resources. *The University of Texas at Austin*, Dec 2023.

[106] Helen Smith. Clinical ai: Opacity, accountability, responsibility and liability. *AI amp; SOCIETY*, 36(2):535–545, Jul 2020.

[107] Beata Stefanowicz. Ai recruitment statistics: What is the future of hiring?, May 2024.

[108] Dianna L. Stone, Kimberly M. Lukaszewski, and Richard D. Johnson. Will artificial intelligence radically change human resource management processes? *Organizational Dynamics*, 53(1):101034, Jan 2024.

[109] K Sri Surya, Reguri Sharanya, Afrah Zilani, and DrCh. Niranjan. Smart applicant tracking system using gen ai. *International Journal for Innovative Engineering Management Research*, 13, 2024.

[110] Brittany Swift. Artificial constraints on opportunity: Artificial intelligence and gender discrimination in automated hiring practices from an information fiduciary perspective. *BUJ Sci. & Tech. L.*, 28:215, 2022.

[111] Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[112] Zarina Tasheva and Vitali Karpovich. Transformation of recruitment process through implementation of ai solutions. *Journal of Management and Economics*, 04(02):12–17, Feb 2024.

[113] Daniel Varona and Juan Luis Suárez. Discrimination, bias, fairness, and trustworthy ai. *Applied Sciences*, 12(12):5826, Jun 2022.

[114] Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. Are emily and greg still more employable than lakisha and jamal? investigating algorithmic hiring bias in the era of chatgpt. *Nature Reviews Neuroscience*, 18(7):404–418, 2023.

[115] Andrew C. Wicks, Linnea P. Budd, Ryan A. Moorthi, Helet Botha, and Jenny Mead. Automated hiring at amazon. 2021.

[116] Paris Will, Dario Krpan, and Grace Lordan. People versus machines: Introducing the hire framework. *Artificial Intelligence Review*, 56(2):1071–1100, May 2022.

[117] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, and et al. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digital Health*, 6(1), Jan 2024.

[118] Sijing Zhang, Ping Li, and Ziyan Cai. Are male candidates better than females? debiasing bert resume retrieval system. *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2022.

[119] Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. Gptbias: A comprehensive framework for evaluating bias in large language models, 2023.

[120] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, Feb 2022.

[121] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[122] Jianlong Zhou, Fang Chen, and Andreas Holzinger. Towards explainability for ai fairness. *xxAI - Beyond Explainable AI*, page 375–386, 2022.