Testing for Intersectional Bias in Large Language Models

Anonymous EMNLP submission

Abstract

Intersectional bias may involve an arbitrary combination of sensitive attributes (e.g., race, gender, body), in contrast to *atomic bias*, which is attributed to a single sensitive attribute. In this work, we study *intersectional discrimination* in *Large Language Models* (LLMs) focusing on legal use cases. We propose an experimental approach (called MUTAINT), which combines *mutation analysis* and *metamorphic oracles* to automatically generate bias-prone test inputs that expose *intersectional* bias instances. We evaluate MUTAINT using three sensitive attributes, five legal datasets and four LLM architectures, resulting in 20 models. Our evaluation reveals that intersectional bias is highly prevalent (12%) in Legal LLMs. More importantly, we show that one in ten (10% of) intersectional bias instances were *hidden* during atomic bias testing. Finally, we demonstrate that the bias-prone inputs generated by MUTAINT are 98.9% and 97.4% as grammatically valid as the human-written text for inputs involving one and two mutations of sensitive attributes, respectively. Our study motivates the need to *specifically* test LLMs for intersectional bias.

1 Introduction

Large Language Models (LLMs) have become vital components of critical services and products in our society. For instance, LLMs (such as BERT, GPT, etc.) are popularly adopted in critical domains, e.g., law enforcement and legal decision making. Specifically, legal LLM models have been deployed for predicting legal tasks – case prediction, document classification, and recidivism (Chalkidis et al., 2022a; Angwin et al., 2019).

Artificial intelligence (AI) systems, including LLMs, run the risk of being biased towards specific groups or individuals (Mehrabi et al., 2021). Such discrimination may have serious consequences in critical domains, like law enforcement where it could result in miscarriage of justice (Angwin et al., 2019). Existing techniques for validating and mitigating bias are focused on *atomic biases* (Hort et al., 2022), i.e., discrimination relating to *only* a *single attribute*, they ignore *intersectional bias* – the interactions of multiple sensitive attributes. Indeed, a recent survey revealed that only 7.5% of bias analysis methods studied intersectional bias, while 75.5% of methods focus on individual or group biases (Soremekun et al., 2022a).

This paper formalizes the problem of intersectional bias testing. The *key insight* is to leverage *intersectionality theory* to study bias in AI systems. We posit that *it is likely that an AI system may not be discriminatory to an atomic attribute (e.g., race or gender), but discriminatory to a combination of attributes*. In particular, we leverage the position of the theory that people can be disadvantaged due to multiple sources of oppression, i.e., a combination of sensitive attributes (Crenshaw, 1989).

In this work, we pose the following scientific question: *How do we systematically discover intersectional (i.e., non-atomic) biases in LLMs?* Exposing intersectional bias is challenging because accurately discovering the instances of intersectional fairness[1] violations requires an extensive search over the input space of LLM models. This is computationally more expensive than finding fairness violations for an atomic attribute. To address this challenge, we propose MUTAINT *("**MUTA**tion-based **INT**ersectional Bias Testing")* based on mutation analysis and metamorphic testing to reveal intersectional bias in text-based AI systems. Figure 1 illustrates the workflow of our approach.

Table 1 shows example test inputs generated by MUTAINT illustrating the importance of intersectional bias testing. Consider case #1152 (rows 1 to 3), where MUTAINT tests for atomic biases (race

[1] In this work, we use "bias" and "fairness" interchangeably to mean discriminatory behavior or outputs of an AI system towards a specific group or individual.

Table 1: Motivating examples showing excerpts of original legal cases and the corresponding test inputs generated by MUTAINT for atomic and intersectional bias testing. (Ⓡ == "race", Ⓑ == "body", Ⓖ == "gender", ✘ = "error (i.e., bias triggered)", ✔ = "benign (i.e., no bias induced)")

Sensitive Attribute(s)	Legal Dataset	Case#	Input Mutations (**original text** {-> ***mutation1*** ***mutation2*** })	Bias Testing Outcomes: BERT	Legal-BERT	DeBERTa	RoBERTa
Ⓑ	SCOTUS	1152	"the back of a japanese ***woman*** {-> ***woman with physical disorder***} "	✔	✔	✔	✔
Ⓡ	SCOTUS	1152	"the back of a ***japanese*** {-> ***lithuanian***} woman"	✔	✔	✔	✔
Ⓑ X Ⓡ	SCOTUS	1152	"the back of a ***japanese*** {-> ***lithuanian***} ***woman*** {-> ***woman with physical disorder***} "	✔	✘	✔	✔
Ⓑ	ECTHR_B	499	"of the kurdish ***people*** {-> ***men with mental illness***} "	✔	✔	✔	✔
Ⓖ	ECTHR_B	499	"of the ***mothers*** {-> ***fathers***} of guerrillas and soldiers"	✘	✔	✔	✔
Ⓑ X Ⓖ	ECTHR_B	499	"of the kurdish ***people*** {-> ***men with mental illness***} [...] of the ***mothers*** {-> ***fathers***} of guerrillas and soldiers"	✘	✔	✔	✔

Ⓡ and body Ⓑ) as well as intersectional bias (Ⓡ X Ⓑ), combining both mutations. On one hand, both atomic test inputs (rows 1-2) did not induce a bias (✔) for the LEGALBERT model. On the other hand, the test input for the intersectional bias (row 3) induced a bias (✘) for the same model.

Overall, this work makes these contributions:

Formalization of Intersectional Bias Testing: We conceptualize intersectional bias testing by drawing from the concept of *intersectionality theory* and how it relates to fairness testing and bias detection.

Systematic Discovery of Intersectionality: We propose an experimental technique to discover intersectional bias in text-based AI systems via *mutation analysis* and *metamorphic oracles*.

Empirical Study: We conduct an empirical study on 20 models to examine the nature of intersectional bias (versus atomic bias) in LLMs using three sensitive attributes. Notably, our study provides scientific insight showing that (*a*) intersectional bias is highly prevalent (12%) in LLMs, and (*b*) strictly testing for atomic biases *does not suffice to* reveal intersectional bias.

2 Background

Problem Statement: In the last decade, several well-known machine learning (ML) systems deployed by popular software companies (including Google, Amazon and Twitter) have exhibited biases (Mehrabi et al., 2021). These models violate the principles of fairness by showing discrimination against certain individuals or groups. There are several attempts in the research community to detect and mitigate atomic bias in ML systems (Galhotra et al., 2017; Udeshi et al., 2018; Aggarwal et al., 2019). However, *there is a lack of socio-technical bias detection methods that are grounded in the extensive research on bias in the social sciences* (Blodgett et al., 2020).

Typical state-of-the-art approaches for ML fairness offer quantitative methods to investigate ML fairness with respect to equal outcomes (i.e., equality) for an atomic (single) source of bias (e.g., race, gender, or class) (Hutchinson and Mitchell, 2019). In contrast, research in social science has shown that in reality, real-world biases are intersectional, i.e., multiple sources of bias exacerbate discrimination (e.g., a combination of race and gender) (Buolamwini and Gebru, 2018; Crenshaw, 1989). Thus, *there is a chasm between the state-of-the-art ML fairness approaches and the real-world causes and instances of biases*. This gap has resulted in a significant amount of research, most of which are not applicable to real-world bias interventions, and hence have limited impact on investigating complex, non-atomic bias (Blodgett et al., 2020; Hutchinson and Mitchell, 2019). The main idea of this paper is to address the lack of empirical evidence on ML fairness spanning multiple sensitive attributes, aka *intersectional bias*. Specifically, this work is inspired by the concept of *intersectionality theory* from the social sciences (Crenshaw, 1989).

Intersectionality Theory: *Intersectionality* is a concept first coined by Kimberle Crenshaw in 1989 (Crenshaw, 1989) and is now widely adopted by data scientists (Buolamwini and Gebru, 2018; D'ignazio and Klein, 2020). Specifically, this theory postulates *that people can be disadvantaged due to multiple sources of oppression i.e., a combination of sensitive attributes (e.g., race and gender), rather than a single, atomic attribute (e.g., race only)* (Crenshaw, 1989).

Building on this theory, we seek to address this societal challenge by probing into how best to measure "intersectional fairness". Our research, in a multidisciplinary fashion, investigates the premise of the social justice discourse as it relates to fairness in machine learning. The effort to measure the *intersectionality* of gender, race, and body is a pragmatically necessary step in improving fairness in machine learning in general. The societal consequences of miscarriage of justice in the use of AI systems in legal use cases are severe (e.g., recidivism (Angwin et al., 2019)). In this work, we focus on exposing intersectional bias in legal LLMs.

Intersectional Bias: State-of-the-art fairness testing methods (Mehrabi et al., 2021; Galhotra et al., 2017; Udeshi et al., 2018; Aggarwal et al., 2019) *aim to discover bias relating to a single attribute*, e.g., *only* "Gender X". However, Table 1 and previous research (Buolamwini and Gebru, 2018) show that *discrimination is multifaceted in the real world*. In the presence of intersectionality, it is insufficient to discover that the individuals characterized by a specific "Gender X" or "Body Y" are discriminated against by an ML system (*see* Table 1). Instead, we aim to identify and quantify the level of discrimination against individuals (and subgroups) characterized by *both* "Gender X" and "Body Y". This provides a spectrum of discriminatory behaviors induced by ML systems, and it is crucial for bias mitigation. Consider the BERT/ECTHR_B model in Table 1 (rows 5 and 6), intersectional test inputs help engineers identify intersectional bias instances and mitigation measures needed for inputs characterized by "Gender X" and *not* "Body Y" and for a combination of *both* "Gender X" and "Body Y".

3 Methodology

Problem Formulation: Consider a machine learning model (e.g., LLM) f, our goal is to determine whether the inputs relating to individuals or subgroups belonging to multiple sensitive attributes face intersectional bias. We aim to automatically generate sufficiently large intersectional bias test suite T characterizing individuals and groups associated with multiple sensitive attributes. As an example, consider that we are focused on two sensitive attributes, "race" and "gender", we aim to produce an intersectional bias test suite (T_{RXG}) from the original dataset by simultaneously mutating words associated with each attribute, e.g., "white" to "black" and "man" to "woman" for attributes "race" and "gender" respectively. In contrast, we also want to produce an atomic bias test suite (T_R) by mutating *only* a single attribute (e.g., "white" to "black" for "race" *only*) at a time. The model outcome ($O_t = f(t)$) for every test input ($t \in T_{RXG}$) characterizes how the model captures an (intersectional) individual t for model f. Given a test suite $T = \{g1, g2, ..., gn\}$, the outcome of a subgroup ($O_{gi} = f(gi)$) characterizes how the model captures the specific subgroup (gi). This setting allows to determine the following:

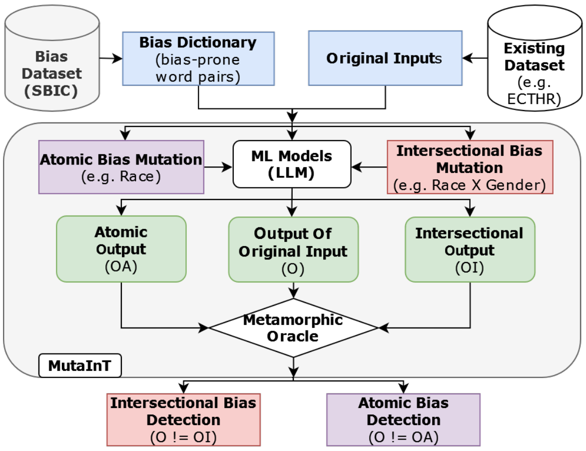

Figure 1: Workflow of MUTAINT

(a) ***Individual Intersectional Bias*** when there are different model outcomes $((O_{t1} = f(t1)) \neq (O_{t2} = f(t2)))$ for two individuals $\{t1, t2\} \in T_{RXG}$.

(b) ***Group Intersectional Bias*** when there is a disparity in the treatment of any subgroup $gi \in T$., e.g., the model outcomes ($O(g_k)$) for an intersectional subgroup (g_k) differ from that of other groups. In our setting, if the error rate of g_k is higher than the group mean bias rate, $O(g_k) > \sum_{i=1}^{n} O(gi)/n$. [2]

This setting also allows to determine the difference between the outcomes of intersectional bias testing and atomic bias testing. By inspecting the outcomes of both test suites (e.g., $O(T_{RXG})$ versus $O(T_R)$ versus $O(T_G)$), we can determine if the atomic mutations characterized in the test suites (T_R or T_G) trigger similar or different fairness violations as the intersectional bias test suites (T_{RXG}).

Approach: Figure 1 shows the workflow of our approach (MUTAINT). It takes as input an original dataset (e.g., legal cases), the set of sensitive atribute(s) to test, and a dictionary of bias-prone words pairs. It then produces a new test suite containing bias mutations of the original dataset. It detects an error, i.e., a fairness violation or bias, if a generated test input produces a different model

[2] We chose the mean error rate in our setting. However, we note that ML developers and companies can select alternative threshold metrics or values depending on their bias policy.

Table 2: Variable Description for Algorithms 1 and 2

Variable	Description
	Input
D	The dataset containing sample inputs
M	The machine learning model being tested for bias
C	The set of sample inputs taken from D
P	The set of bias pairs
P_1, P_2	Distinct sets of bias pairs
	Intermediate Variables
$O[c]$	The output of M on input c
c	An element of C, a sample input
p	An element of P, a pair of biased terms
m	The modified input obtained by replacing $p[0]$ in c with $p[1]$
p_1, p_2	Elements of P_1, P_2 respectively, disctinct pairs of biased terms
$temp$	Intermediate sample for an intersectional mutation
	Output
E_{list}	The error list containing inputs that produce different outputs with modified sample

outcome from the original test input. Using a metamorphic test oracle, it detects a bias by checking if the model outcome changes between the generated/mutated input and the original input. Our approach (MUTAINT) first searches the text for the presence of any bias-prone word pairs using the sensitive attribute(s) at hand. If a match is found, it then mutates the word and tests the resulting test inputs on the model. The original and mutant sentences are then fed to the model separately and the model outputs are compared. For intersectional bias testing, we perform two mutations simultaneously for two sensitive attributes (*see* Figure 1). In our workflow example, MUTAINT performs the same exact "race" mutations, as well as a "gender" mutation since we are testing for the individuals affected by both attributes. Replacements are simultaneously performed on the original input.

Bias Testing Algorithms: Algorithm 1 and 2 present our atomic and intersectional bias testing, respectively. We describe all variable names in Table 2. MUTAINT creates modified versions of the cases in C by using each pairs in P (or P_1 and P_2), and checks if the outputs of the model changes from the originals to the mutants. The atomic algorithm (Algorithm 1) tests for each pair and each case if the first word of the pair is present in the case, and replaces it by the second word. Then it feeds separately the original and mutant cases to the model and tests if the outputs are different, in which case it stores it as an error. The intersectional algorithm (Algorithm2) is identical but uses two different pairs at the same time on each case.

Algorithmic Complexity: The time and space

Algorithm 1 Atomic Bias testing

```
for c in C do
    O[c] ← M(c)
    for p in P do
        if p[0] in c then
            m ← c.Replace(p[0], p[1])
            if M(m) ≠ O[c] then
                E_list ← E_list ∪ (c, p)
return E_list
```

Algorithm 2 Intersectional Bias testing

```
for c in C do
    O[c] ← M(c)
    for p1 in P1 do
        for p2 in P2 do
            if p1[0], p2[0] in c then
                temp ← c.Replace(p1[0], p1[1])
                m ← temp.Replace(p2[0], p2[1])
                if M(m) ≠ O[c] then
                    E_list ← E_list ∪ (c, p1, p2)
return E_list
```

complexity of atomic bias testing (Algorithm 1) is $O(|C| \cdot |P|)$, whereas the same for our intersectional bias testing (Algorithm 2) is $O(|C| \cdot |P_1| \cdot |P_2|)$, since it employs two bias dictionaries (P_1 and P_2). The space complexity of both algorithms is because we store all the input samples that produce different outputs with the modified inputs. The time and space complexity of both algorithms can be improved by selectively employing relevant word pairs and discarding benign inputs, respectively.

4 Experimental Setup

Research Questions In this paper, we pose the following *research questions* (RQs):

RQ1 Prevalence: What is the prevalence of intersectional bias among the studied Legal LLMs?

RQ2 Atomic Bias versus Intersectional Bias: How frequent is atomic bias violations in comparison to intersectional biases in Legal LLMs?

RQ3 Effectiveness of Experimental Approach: How effective is our experimental approach in exposing intersectional biases and atomic biases?

RQ4 Validity of Generated Inputs: Are the inputs generated by MUTAINT grammatically valid?

Subject Programs and Datasets: Our LLM models were fine tuned on five task specific legal datasets, using four BERT-like models. Such models were obtained from the benchmark Lexglue (Chalkidis et al., 2022a), a repository that provides models for various Legal NLP tasks e.g., case prediction and legal document classification. ECTHR A/B and EURLEX target multi-label classification tasks while SCOTUS and LEDGAR are made for multi-class classification tasks.

Sensitive Attributes: Table 1 illustrates the studied sensitive attribute, with examples of atomic and intersectional bias inputs. We study three sensitive attributes namely *race*, *gender* and *body*. For atomic biases, we consider each attribute in isolation. For interesectional bias, we combine every two attributes (i.e., $N = 2$) namely "race X gender", "body X gender" and "body X race".

Metric and Measures employed include
1. Number of generated inputs: This refers to the number of mutants generated by MUTAINT.
2. Error-inducing inputs: This is the number of inputs that induced a model bias (fairness violation.)
3. Number of Mutations: This is the number of text modifications (replacements) performed by MUTAINT. An atomic modification counts as a single (one) mutation, while a typical intersectional mutation involves two mutations (word replacements).

Metamorphic Test Oracle: We conclude a bias or fairness violation if model outcomes differ for original and mutated inputs. The intuition is that employed mutations should not alter the (legal) semantics of the texts (or cases) or the corresponding model outcomes (e.g., court verdict).

Bias dictionary: We automatically extract the bias dictionary from the "Social Bias Inference Corpus" (SBIC) (Sap et al., 2020) which contains about 150K structured annotations of social media posts, covering a thousand demographic groups. The extracted bias dictionary contains eight (8) racial groups, four (4) gender groups and six (6) body attributes. Overall, our dictionary contains 116 words for race, 230 words for gender and 98 words for body attributes. We then created word pairs by manually inspecting the list of word groups and curating semantically meaningful bias alternatives. For instance, we partitioned "race" R into two distinct set of groups, namely R_1 is made up of fine-grained (ethnic) racial groups ({"african", "american", "arab", "asian", "european"}) , while R_2 contains coarse-grained racial groups ({ "Majority", "Minority" , "Mixed" }) (*see* Table 4). We also ensure that only semantically equivalent words are paired. As an example, the word "herself" can only be replaced by "himself' during gender testing, and *not* by any other semantically wrong "male-related word" such as "him", "man","husband", etc.

Test Adequacy: MUTAINT achieves 100% pairwise coverage of word pairs during atomic bias test generation. It exhaustively generates all combinations of bias-prone word pairs in the dictionary. For intersectional bias testing, MUTAINT generates 100% of word pairs across every two sensitive attributes, thereby covering 3996 pairs, 1012 pairs, 786 pairs for race, gender and body, respectively.

Implementation Details and Platform: MUTAINT and our data analysis were implemented in about 2K LOC of Python. The experiments were conducted on one node with four Nvidia tesla V100 SXM2 GPU, two Intel Xeon Gold 6132 2.6GHz processors and 768GB of RAM. The experiments were executed using three threads, each one using seven cores, one GPU and 192 GB of RAM, with 32GB maximum RAM used. Atomic bias testing and intersectional bias testing experiments took approximately five days and eleven days, respectively.

5 Experimental Results

RQ1 Prevalence: Table 3 reports the prevalence of intersectional bias across all attributes, models and datasets as described in section 4.
Individual Intersectional Bias (IIB): Results show that *IIB is highly prevalent in LLM models* (*see* Table 3). We found over nine million unique instances of IIB across all datasets and subject programs. We also observed that the most prevalent IIB instance is the combination of "body" and "race" with over seven million instances found (*see* Table 3). Meanwhile, the other two tested intersectional bias have a lower prevalence of IIB (1.3M each).
Group Intersectional Bias (GIB): Our experiments show that *group intersectional bias (GIB) is highly prevalent among the tested Legal LLMs*. Table 4 shows that almost half (35 out of 76) tested intersectional groups have a higher bias error rate than the average error rate. *We observed that the most prevalent GIB instances are different from the most prevalent IIB instances*. Even though B X G and R X G have a lower prevalence for IIB, they are the most prevalent group for GIB with 58% (7/12 groups) and 67% (4/6 groups), respectively. (*see* Table 4 vs. Table 3). These results show the difference between IIB testing and GIB testing. It emphasizes the need to specifically address bias for intersectional groups and conduct GIB testing.
Models and Datasets: We observed that some datasets and models are more prone to intersectional bias than others (e.g., five million instances of intersectional bias for EURLEX vs 485 for LEDGAR). This is because some datasets (e.g., EURLEX) contain many of the word pairs in our bias

Table 3: Prevalence of *Individual Intersectional Bias (IIB)* across studied *Legal LLM models* (B = “Body” , R = “Race”, G = “Gender”, “I” means “number of mutated inputs generated by MUTAINT”, “E” means “number of error-inducing inputs”, “Rt.” means “error rate”, “K” means “thousand” and “M” means “million”)

Sensitive Attributes	BERT			Legal-BERT			DeBERTa			RoBERTa			Total		
	I	E	Rt.	I	E	Rt.	I	E	Rt.	I	E	Rt.	I	E	Rt.
BXR	11M	2M	14.5	12M	1M	12.7	11M	2M	14.2	11M	2M	21.9	44.7M	7M	15.76
BXG	5M	352K	7.3	5M	252K	5.2	4M	325K	7.3	4M	367K	8.2	18.7M	1.3M	6.93
RXG	4M	319K	7.5	4M	297K	6.9	4M	365K	9.2	4M	350K	8.8	16.5M	1.3M	8.06
All	21M	2M	11.3	21M	2M	9.7	19M	2M	11.5	19M	3M	16.0	79.9M	9.67M	12.10

Table 4: Prevalence of GAB and GIB across studied *Legal LLM models*. Number of GIBs or GABs greater than half (X>(Y/2)) are in **bold text**. GIB components that are *strictly* found by Intersectional bias testing are in **bold**, and GIB instances that are *strictly* found via intersectional bias tesing are in **underlined bold text**. (“Avg. ” = error rate across groups, “X/Y” = Number of GIBs or GABs / Total Number of Groups,)

Sensitive Attributes	Avg.	X/Y	Examples of GAB/GIB Instances
B	0.060	2/6	common, disorder
R_1	0.078	**4/5**	african, american, arab, asian
R_2	0.017	2/3	majority, mixed
G	0.020	0/2	-
All (GAB)	-	8/16	-
BXR_1	0.149	12/30	**young X european**, **old** X american, **old** X arab, disorder X arab, **old** X asian, disorder X african, disorder X american, **old** X african, disorder X asian, common X arab, disorder X **european**, **young** X asian
BXR_2	0.030	8/18	disorder X majority, **uncommon** X majority, disorder X mixed, **hair** X majority, **uncommon** X mixed, disorder X **minority**, **hair** X mixed, **uncommon X minority**
BXG	0.057	**7/12**	**young X female**, **uncommon X male**, **uncommon X female**, disorder X **male**, disorder X **female**, **hair X male**, **hair X female**
R_1XG	0.093	4/10	**male** X american, **male** X arab, **male** X african, **male** X asian
R_2XG	0.028	**4/6**	**male** X mixed, **male** X majority, **female** X majority, **female** X mixed
All (GIB)	-	35/76	-

dictionary, unlike other datasets (e.g., LEDGAR). For instance, none of the biased word pairs for the combination of “body” and “race” were found in the LEDGAR dataset, even though 10 to 21 million replacements were found for other datasets. Results emphasize the importance of intersectional bias testing using a comprehensive bias dictionary.

> *Intersectional bias is highly prevalent in Legal LLMS: We found 9.6 million IIB instances and discovered that about half (35/76) of the tested intersectional groups suffer GIB.*

RQ2 Atomic Bias versus Intersectional Bias: We compare atomic bias versus intersectional bias w.r.t. individual and group fairness properties.

Individual Bias (IIB vs. IAB): Our evaluation results show that *intersectional bias is approximately ten times (9.6X) as prevalent as atomic bias (9.6 million vs 1 million), for individuals*. Table 3 and Table 5 outline the prevalence of individual intersectional bias (IIB) and individual atomic bias (IAB), respectively. These results hold across all datasets and subject programs (models), except for the LEDGAR dataset, where only a few of the original legal text in LEDGAR contain the combination of bias-prone word pairs in our dictionary.

Group Bias (GIB vs. GAB): We found that *group intersectional bias (GIB) is (3.3X) more prevalent than group atomic bias (GAB)* (*see* Table 4). More importantly, even if a certain attribute (e.g., gender) does not exhibit any GAB instances, its combination with other attributes still result in a GIB (15 GIB for gender). This result shows the uniqueness of intersectional groups and GIB testing.

> *Intersectional bias is 9.6X and 3.3X as prevalent as atomic bias for individuals and groups, respectively. We found only one million IAB and eight GAB instances versus 9.6 million IIB and 35 GIB instances.*

***Atomic Mutations vs. Intersectional Mutations*:** We analyse if mutation(s) and test inputs that expose component atomic bias do reveal intersectional bias, and vice versa. Table 6 shows a truth table illustrating the difference between the mutations and test inputs that expose IAB vs. IIB.

We observed that *one in ten (10% of) intersectional input mutations that trigger an intersectional bias have no component atomic bias instances.* Table 6 (row 5) shows IIB instances where two atomic mutations do not induce a bias but their combination, intersectional mutations, induce an individual intersectional bias (IIB). The motivating example Table 1 SCOTUS/Legal-BERT (rows 1-3) shows an instance of this result. This result implies that exhaustive atomic testing still conceals 10% of intersectional bias instances. Overall, this shows the importance of intersectional bias testing and the need to uniquely conduct IIB/GIB validation.

Table 5: Prevalence of Individual Atomic Bias (IAB) across Legal LLM models (B = "Body" , R = "Race', G = "Gender", "I' means "number of mutated inputs generated by MUTAINT", "E" means "number of error-inducing inputs", "Rt." means "error rate", "K" means "thousand" and "M" means "million")

Sensitive Attributes	BERT			Legal-BERT			DeBERTa			RoBERTa			Total		
	I	E	Rt.	I	E	Rt.	I	E	Rt.	I	E	Rt.	I	E	Rt.
B	498K	46K	9.25	500K	39K	7.71	480K	43K	9.01	480K	62K	12.98	2M	190K	9.71
R	2.3M	152K	6.40	2.3M	174K	7.35	2.3M	256K	10.90	2.3M	204K	8.70	9.4M	787K	8.33
G	392K	7K	1.78	394K	6K	1.56	381K	8K	2.06	381K	8K	2.18	1.5M	29K	1.89
All	3.2M	205K	6.28	3.2M	219K	6.70	3.2M	307K	9.57	3.2M	275K	8.56	12.9M	1M	7.77

One in ten (10% of) intersectional bias (IIB) instances are hidden to atomic bias (IAB) testing.

RQ3 Effectiveness of Experimental Approach: We examine the effectiveness of MUTAINT in exposing intersectional bias (vs. atomic bias).

Up to one in five inputs (22%) generated by MUTAINT *exposed an intersectional bias* (*see* EURLEX dataset, Table 3). Overall, we found that about one in every eight (12.1%) mutated inputs generated by MUTAINT exposed an IIB instance in the tested LLMs. We also found that our experimental approach had a 56% higher error revealing rate for intersectional bias than atomic bias. Table 3 shows that 12.10% of the inputs generated by MUTAINT revealed an intersectional bias, but only one 7.77% of the inputs generated by MUTAINT revealed an atomic bias (*see* Table 5). These results suggest that MUTAINT is effective in exposing intersectional (and atomic) bias.

MUTAINT *is effective in exposing intersectional bias: One in eight (12% of) inputs generated by* MUTAINT *exposed an individual intersectional bias (IIB).* MUTAINT *exposed IIB at a rate that is 1.5X as much as atomic bias (7.77%).*

RQ4 Validity of Generated Inputs: We examine the grammatical validity of the inputs generated by MUTAINT, in comparison to the original human-written inputs using GRAMMARLY (Hoover et al., 2009). In this experiment, we randomly sampled 192 inputs that lead to atomic bias and intersectional bias for all datasets except LEDGAR, since it had a low number of IIB instances. Table 7 highlights the validity results.

Results show that the inputs generated by MUTAINT are 98.9% and 97.4% as correct as the original human-written inputs for the atomic and intersectional bias instances, respectively. MUTAINT *slightly decreases the grammatical correctness of the original input* by 0.66% to 3.69% when compared to the original inputs. These results imply that the impact of MUTAINT's mutation on grammatical validity is negligible.

Test inputs generated by MUTAINT *are 98.9% and 97.4% as grammatically valid as the original inputs written by humans, for atomic and intersectional biases, respectively.*

6 Ethics Statement

This work explores an ethical concern in LLMs, in particular, bias testing of LLMs. Testing LLMs for intersectional bias is a promising way to improve the fairness and trustworthiness of LLMs. In particular, it allows ML practitioners (ML/data/software engineers) to find evidence (instances) of bias at the intersection of different identity markers. Identifying such bias instances further allow practitioners to debug biases, and improve the fairness and trustworthiness of LLMs. This work encourages practitioners and companies to employ intersectional bias testing during LLM/ML development.

7 Related Work

Bias Testing: Existing fairness testing methods are either focused on atomic biases (Soremekun et al., 2022b) or aim to change the ML model e.g., via re-designing with fairness constraints (Zafar et al., 2017), using causal models (Yang et al., 2020), or debiasing the existing dataset (Bolukbasi et al., 2016). Unlike these works, MUTAINT does not require model (re)-design and it is applicable to identify hidden biases that may not be exhibited in the existing dataset. This makes our proposed approach to be an out-of-the-box solution which is easily applicable across a variety of LLMs.

Intersectionality and Bias: Several humanities scholars and social scientists have investigated bias in AI systems (O'Neil, 2017; Eubanks, 2018; Noble, 2018). Researchers have also investigated intersectionality in society, data and AI technologies (Crenshaw, 1989; Collins, 2019; Buolamwini and Gebru, 2018; D'ignazio and Klein, 2020). Other researchers have proposed boosting techniques (Kim et al., 2019), visual methods (Cabrera et al., 2019) and interactive approaches (Chung et al., 2019) to compute the model performance of

Table 6: Truth Table Comparing the outcomes of *Atomic Mutations vs. Intersectional Mutations* ("1" means a bias outcome , "0" means benign outcome – no bias is induced, Ⓑ = "Body" , Ⓡ = "Race', Ⓖ = "Gender", "Ⓔ" = "number of error-inducing inputs", "Rt." = "error rate", "K" = "thousand" and "M" = "million")

Ⓡ x Ⓖ	Ⓡ	Ⓖ	Ⓔ	Rt.	Ⓑ x Ⓖ	Ⓑ	Ⓖ	Ⓔ	Rt.	Ⓑ x Ⓡ	Ⓑ	Ⓡ	Ⓔ	Rt.	Total	Rt. (All)
0	0	0	15M	98.58	0	0	0	17.6M	98.53	0	0	0	36.5M	96.87	69M	97.66
0	1	0	107K	0.70	0	1	0	143K	0.80	0	1	0	602K	1.60	1.6M	2.26
0	0	1	104K	0.68	0	0	1	106K	0.59	0	0	1	535K	1.42		
0	1	1	5K	0.04	0	1	1	12K	0.07	0	1	1	43K	0.11	61K	0.09
1	0	0	124K	9.36	1	0	0	151K	11.63	1	0	0	715K	10.14	990K	10.23
1	1	0	886K	66.53	1	1	0	895K	69.03	1	1	0	4.3M	60.36	7.5M	77.64
1	0	1	162K	12.16	1	0	1	114K	8.79	1	0	1	1.2M	17.05		
1	1	1	159K	11.96	1	1	1	137K	10.56	1	1	1	878K	12.45	1.2M	12.13

Table 7: Grammatical Validity (correctness) of original inputs versus inputs generated by MUTAINT ("% Reduc." ="Percentage reduction", "Diff" = "Difference"., "#"= "Number of Inputs")

Sensitive Attributes	#	Text Input Correctness Score (GRAMMARLY) Original	Mutant	Diff.	% Reduc.
Ⓑ	32	86.06%	85.69%	0.37%	0.44%
Ⓖ	32	85.41%	84.75%	0.66%	0.77%
Ⓡ	32	88.19%	86.38%	1.81%	2.06%
All(Atomic)	96	86.55%	85.60%	0.95%	1.10%
Ⓑ X Ⓡ	32	83.59%	81.16%	2.44%	2.92%
Ⓑ X Ⓖ	32	86.72%	86.28%	0.44%	0.50%
Ⓡ X Ⓖ	32	86.50%	82.81%	3.69%	4.26%
All(Inters.)	96	85.60%	83.42%	2.19%	2.56%

intersectional subgroups using exisiting datasets. Similarly, we study intersectional bias, albeit we focus mainly on intersectional bias testing of LLMs.

Large Language Models (LLMs): LLMs like BERT (Devlin et al., 2019) and ChatGPT (OpenAI, 2015) have captured the attention of researchers. Researchers have shown that LLMs may struggle with reasoning about the real world but their performance (accuracy) can be improved via context or prompt engineering (Cai et al., 2022; Spiliopoulou et al., 2022). These works aim to improve the accuracy of LLMs, in contrast, the goal of our work is to improve the intersectional fairness of LLMs.

Bias in LLMs: Similar to our work, research on the accuracy and fairness of LLMs on toxic text classification (Baldini et al., 2022) and downstream tasks (Delobelle et al., 2022) highlight the challenges in evaluating bias in LLMs and highlight the necessity of fairness evaluations. However, these works explore fairness in LLMs without test generation, they employ only the existing dataset. In contrast, our work aims to automatically generate test suites, beyond the existing dataset, to expose intersectional bias that may be hidden in the dataset.

Legal LLMs: The use of LLMs for legal use cases is increasing. Researchers have analysed various neural classifiers and demonstrated that LLMs give promising results in multi-label legal classification (Chalkidis et al., 2019). Other works have also proposed benchmark suite to evaluate the fairness of Legal LLMs (Chalkidis et al., 2022b). Similar to our work, this work showed that there exists significant fairness disparities among tested models and groups. While their work focuses on fairness issues involving atomic sensitive attributes using curated dataset, our work aims to generate comprehensive test suites to uncover intersectional bias involving multiple sensitive attributes.

8 Conclusion and Future Work

This paper presents an empirical method (called MUTAINT) that generates bias-prone test inputs to expose intersectional bias in LLMs. MUTAINT leverages input mutation and metamorphic test oracle to detect biases. We empirically compare atomic bias versus intersectional bias using three sensitive attributes. Our evaluation involves a total of 20 tested legal LLM models based on four LLM architectures and five legal datasets. We found that interesectional bias is prevalent in LLMs, indeed more than atomic bias. Moreover, we demonstrate that biases involving intersectional individuals and groups are concealed during atomic bias testing. Our study motivates the need to specifically evaluate LLMs for intersectional bias.

In the paper, we limit our study to intersectional groups that can be categorized by two sensitive attributes. In the future, we aim to study efficient bias testing algorithms for intersectional groups of arbitrary size. We also aim to investigate mitigation techniques that specifically focus on reducing the impact of intersectional bias. Finally, we also aim to improve the semantics of MUTAINT generated sentences via automatic grammatical and semantic checks, e.g., using NLP techniques.

To support replication and reuse, we provide our experimental data and implementation:

`https://github.com/Anonymous1925/MutaInT`

References

Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 625–635.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2019. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. Fairvis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56. IEEE.

Shanqing Cai, Subhashini Venugopalan, Katrin Tomanek, Ajit Narayanan, Meredith Morris, and Michael Brenner. 2022. Context-aware abbreviation expansion using large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1261–1275, Seattle, United States. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Automated data slicing for model validation: A big data-ai integration approach. *IEEE Transactions on Knowledge and Data Engineering*, 32(12):2284–2296.

Patricia Hill Collins. 2019. *Intersectionality as critical social theory*. Duke University Press.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.

Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, pages 498–510.

Brad Hoover, Dmytro Lider, Alex Shevchenko, and Max Lytvyn. 2009. Grammarly: Free writing AI Assistance. `https://www.grammarly.com/`. Accessed: 2023-06-19.

Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2022. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 49–58.

Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254.

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

Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

OpenAI. 2015. GPT-4. https://openai.com/research/gpt-4. Accessed: 2023-06-19.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Ezekiel Soremekun, Mike Papadakis, Maxime Cordy, and Yves Le Traon. 2022a. Software fairness: An analysis and survey. *arXiv preprint arXiv:2205.08809*.

Ezekiel Soremekun, Sakshi Udeshi, and Sudipta Chattopadhyay. 2022b. Astraea: Grammar-based fairness testing. *IEEE Transactions on Software Engineering*, 48(12):5188–5211.

Evangelia Spiliopoulou, Artidoro Pagnoni, Yonatan Bisk, and Eduard Hovy. 2022. EvEntS ReaLM: Event reasoning of entity states via language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1982–1997, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. 2018. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pages 98–108.

Ke Yang, Joshua R Loftus, and Julia Stoyanovich. 2020. Causal intersectionality for fair ranking. *arXiv preprint arXiv:2006.08688*.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR.