

Extrinsic Bias Patterns May Suggest Underlying Intrinsic Bias: A Cross-Domain Analysis

Aryan Potukuchi

Brisbane Boys College, Kensington Terrace, Toowong, Queensland, 4066, Australia; aryanpotukuchi@gmail.com

ABSTRACT: Racial biases in LLMs can significantly influence their decisions, resulting in socially unjust consequences. The term ‘Large language models’ refers to advanced computer programs capable of synthesizing complex, human-like responses. They use vast amounts of text as training data, often inheriting certain biases present that degrade the morality of their replies. Whilst measuring and mitigating such biases is crucial, most research is dedicated to moderating extrinsic biases found in LLM outputs, leaving intrinsic biases within LLM mechanisms relatively unexplored. This study will provide a set of criteria used to recognize bias patterns within the data. Using the criteria, we will analyze evidence of LLM performance in fabricated medical, legal, and hiring settings and identify extrinsic racial bias patterns. Our findings indicate that there are similarities between the identified patterns and the concepts of intrinsic biases, suggesting a transfer of intrinsic bias structures into LLM output performance. Additionally, our findings may also suggest the importance of exploring socially derived biases within LLM training data. Overall, our review acts as a foundation for other researchers to validate the correlation between extrinsic and intrinsic bias, potentially contributing to the bias transfer hypothesis and other broader discussions.

KEYWORDS: LLMs in Critical Societal Domains, Racial Bias in LLMs, Large Language Models, Bias Transfer Hypothesis.

■ Introduction

Advancements in Large Language Models (LLMs) in fields such as Natural Language Processing (NLP) introduce exciting possibilities, such as LLM incorporation into critical societal domains. However, there are concerns about the effects of LLM biases, often inherited from pre-training data, on the generated content. In fields like law, medicine, and hiring, such biases can influence hospitalization and medical costs,¹ judicial decisions,² and job employment,³ resulting in serious ethical violations and social inequity. This makes it of utmost importance to quantify and mitigate these biases within LLMs and make them more suitable for such critical domains.

Bias, in this context, refers to systematic errors in LLM token prediction that reflect the stereotypes and prejudices of our world.⁴ These biases exist across a broad spectrum, including gender and race, sexual orientation, religion, socio-economic status, and disability.⁵ Despite the versatility, LLM bias is usually classified into two groups: intrinsic bias and extrinsic bias. Guo *et al.*⁶ state that intrinsic bias refers to the bias within a pre-trained or trained LLM’s internal mechanisms, irrespective of downstream tasks. Extrinsic bias, on the other hand, refers to the bias within the LLM’s outputs after training or pre-training, usually task-dependent. An in-depth discussion between the two bias types is provided in the definitions section. While research papers examine the quantification and mitigation of both intrinsic and extrinsic biases within LLMs individually, a small portion explores the theoretical relationship between the bias categories through the bias transfer hypothesis.

The bias transfer hypothesis refers to the assumption that intrinsic biases within LLMs propagate into downstream tasks, resulting in extrinsic biases within LLM outputs.^{7,8} However, previous studies that test this hypothesis measure the intrinsic-extrinsic correlation using quantitative approaches, such as metrics-based comparisons, finding weak correlations.^{9–11}

This means that they might overlook other potential indicators of bias transfer that don’t rely on quantitative propagation, such as comparisons of bias structure or patterns. Hence, we have decided to approach the hypothesis from an interpretive perspective by exploring the correlation between ‘racial’ bias patterns within the downstream tasks from the Law, Medicine, and Hiring domains, and concepts of intrinsic biases. To distinguish these patterns from statistical noise within the data, we have developed our own set of criteria that assess the strength of the patterns based on three core characteristics – consistency, directionality, and replicability. A detailed description of the criteria is provided in the ‘Criteria for Identifying Bias Trends’ section.

The paper initially established key terminology, such as the definitions of common social biases, including intrinsic and extrinsic bias definitions. Additionally, the section briefly elaborates how intrinsic biases can originate from biased pre-training data, which affects the LLM’s mechanisms. It also discusses how extrinsic biases tend to be downstream effects of intrinsic biases. Before moving on to the analysis, the paper establishes a set of criteria used to identify bias patterns within the data. In the results section, we highlight the importance of the critical domains in society, along with the impacts of extrinsic bias within these domains, and analyze our extrinsic racial bias evidence. Following this, we identify prominent bias patterns within the evidence using the criteria, interpreting them using established intrinsic bias definitions. In the discussion section, we discuss our interpretations of the evidence, along with its relation to the bias transfer hypothesis and other broader concepts. We also state the limitations of the research,

providing further directions for future researchers to verify the interpretations within our research.

Overall, this study contributes to the ongoing conversation of the bias transfer hypothesis by presenting an interpretive perspective on the transfer of bias structures, as opposed to the quantitative propagation measured using metric-based systems. Additionally, our findings may provide valuable insights into discussions regarding the effects of societal biases within LLM training and evaluation practices, suggesting that an understanding of socially grounded biases may be required to improve bias mitigation strategies.

■ Key Terminology

Types of Social Bias:

The term ‘social bias’ encompasses any discriminatory behavior that disadvantages marginalized groups in society. Although there are several types of social biases, discussed in detail by Navigli, Roberto *et al.*,⁵ this section will only define a few common manifestations, which are listed in Table 1.

Table 1: Different types of Social Biases, along with their definitions and examples.

Social Bias	Definition	Example
Age	Age bias or ageism is the discrimination against an individual or a group based on their age.	An LLM from a company that values efficiency might hire more candidates below a certain age (37 years) because it associates ‘youth’ with words such as ‘active’ and ‘efficient’.
Disability	Disability bias is the discrimination against individuals with disabilities, including physical, genetic, and neurological.	For example, an LLM might rank a neurotypical individual higher than a neurodivergent one for the position of a manager because it perceives neurotypical people to be more capable of high-stakes decision-making.
Gender	Gender bias or sexism is the tendency to prefer or assign a single gender over another for a specific task.	An LLM might assume that most doctors in a hospital are males, whereas most nurses are females.
Socioeconomic Status	Socioeconomic bias is the discrimination against individuals or groups with lower socioeconomic status, such as poor people.	An LLM used for Treatment recommendations may suggest inferior, low-cost procedures for people with lower socioeconomic status during a medical emergency.
Sexual Orientation	Sexual bias is discrimination against individuals based on their sexual identity.	An LLM might prefer a heterosexual candidate for a teaching position over a homosexual candidate because it considers the heterosexual individual to be a better role model for children.
Race or Ethnicity	Racial or ethnic bias includes discriminatory behaviors and negative attitudes towards individuals from specific ethnic groups.	An LLM hiring for a company might score white men higher for an important position than black men.
Intersectional	Intersectional bias refers to discrimination against individuals based on multiple aspects of their social identity.	An LLM that isn't biased towards race (white or black) and gender (man or woman) might still be biased towards a specific intersection between them (Black woman).

While there exist many manifestations of social bias in LLM outputs, this paper focuses specifically on racial bias to analyze the relationship between downstream bias patterns and upstream bias mechanisms. The definition of social biases is presented to acknowledge the vast diversity in output bias and to restate that conclusions made within this report cannot be generalized for other social biases without further research.

Intrinsic Bias:

Bias present in large language models is usually classified into two categories, depending on where it is portrayed: intrinsic bias and extrinsic bias. Intrinsic bias refers to bias within the internal mechanisms of the model, often directly associated with the model’s pattern recognition and relationship

encoding within these mechanisms.⁶ For example, a model trained on a medical corpus that frequently represented men as doctors and women as nurses would learn to identify the strong relationship between ‘doctor’ and ‘man’, and likewise for ‘nurse’ and ‘woman’. Consequently, it would change its internal representation to reflect this pattern when prompted.

While there are many causes of intrinsic bias, such as algorithmic factors¹² and annotation bias,¹³ the most common cause is bias in pre-training data. The LLM pre-training process involves multiple steps essential for the model’s text comprehension and generation. These include the collection and processing of raw data for the LLM pre-training dataset, tokenization of data into unique numerical tokens, training the model on tokenized corpora for pattern learning and next-word prediction, and finally, further testing and evaluation to improve response accuracy and quality.¹⁴ However, when trained on biased data, the model provides more attention weight for biased associations with sensitive tokens, increasing the probability of these associations being generated in the responses upon token detection within the prompt.¹⁵

Extrinsic Bias:

Unlike intrinsic bias, extrinsic bias refers to inequality in the model’s performance against specific demographic groups across different downstream tasks.⁶ Being task-dependent, extrinsic bias manifests in different ways, which have been explored in detail by Guo *et al.*⁶ These tasks are classified into Natural Language Understanding (NLU) and Natural Language Generation (NLG).

Natural Language Understanding tasks focus on improving the model’s semantic understanding of the input text to enhance its task-specific performance. However, extrinsic bias in NLU tasks can result in misinterpretation of texts due to biased notions from pre-training. For example, in the sentence, “The doctor called the nurse because she was late.” The LLM might associate ‘she’ with ‘nurse’ due to extrinsic bias. This can affect its performance in tasks such as natural language inference.

On the other hand, Natural Language Generation tasks focus on generating relevant information based on the input prompts. Extrinsic bias in these tasks can result in inappropriate responses, including reinforcement of social biases and stereotypes inherited from training data. For example, the LLM might rank male candidates higher than equally qualified female candidates due to extrinsic bias. This can affect the model’s performance in tasks such as question answering, sentence completion, and machine translation.

As observed, extrinsic bias within NLU and NLG tasks tends to be a downstream effect of biases within the LLM’s internal mechanisms, also referred to as intrinsic bias. This suggests that extrinsic bias patterns may reflect the nature of intrinsic biases within the LLM.

■ Criteria for Identifying Bias Patterns

To classify any observation from the results as a bias pattern, we will use criteria to assess the characteristics of those observations. The criteria consist of three main attributes:

consistency, directionality, and replicability, which have been defined for this investigation:

- **Consistency:** The repeated occurrence of common patterns within a defined setting, including within one or more LLMs exposed to the same task or experiment.
- **Directionality:** The consistent association of certain outcomes with specific demographic groups in the task.
- **Replicability:** The repeated occurrence of common patterns across different settings, such as different tasks, experiments, or datasets.

It is important to note that for an observation to be classified as a bias pattern, it must demonstrate both consistency and directionality. Replicability, whilst being important, will be considered as supporting evidence for the bias pattern.

Law

Law is a crucial component of modern society that maintains social order by reinforcing rules that may compromise an individual's safety and punishing offenders to provide justice for the victims. The domain includes high-stakes tasks with severe consequences, such as convictions, sentencing, and risk prediction before bail. In principle, serious legal decisions are expected to be made using the evidence provided, with irrelevant factors such as demographic attributes having no impact in this process. This would ideally lead to fair and justified outcomes for those tasks. However, bias within this domain suggests that demographic factors may affect the interpretation of available evidence, resulting in potentially skewed outcomes for such decisions. This could result in unethical consequences, such as judicial unfairness against certain demographic groups or decisions, such as conviction rates and sentencing terms, being influenced by the defendant's ethnicity.

With law being a crucial, sensitive domain that is heavily evidence-based, any decisions influenced by demographic biases in LLMs within judicial downstream tasks are noticeable. This makes it easier to analyze and identify extrinsic bias patterns within LLMs. Although we clearly specified that we would analyze evidence observing the effects of racial bias on downstream judicial tasks, such data were relatively scarce compared to other demographic labels. Hence, we broadened our spectrum to include evidence that addresses effects of demographic biases within the Definitions section, only for the 'Law' domain.

Empirical Findings:

The empirical evidence used within this domain came from two studies: Hu *et al.*² and Zhang *et al.*¹⁶

In their study, Hu *et al.*² aim to assess the judicial fairness of Large Language models to examine their reliability in the judicial domain. They test 16 LLMs across three evaluation metrics – inconsistency, bias, and imbalance accuracy – using an extensive dataset that they compiled (JudiFair). The results of these experiments are assessed using regression-based statistical tests and other robustness checks to ensure that reliable conclusions are made using statistically meaningful data. A key representation includes the detailed heatmap in Figure 1. The heatmap displays diversity in statistically significant regression

coefficients that affect sentencing outcomes for specific demographic labels. Upon analyzing the results, Hu *et al.*² concluded by stating that almost all models exhibit substantial and systematic inconsistency, bias, and imbalanced inaccuracy, with demographic attributes consistently triggering more bias.

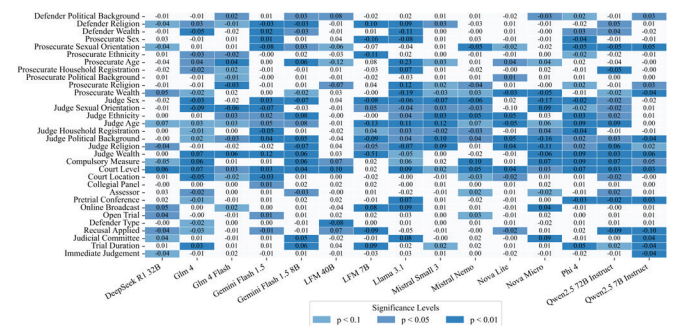


Figure 1: Heatmap of Regression coefficients measuring judicial fairness across 15 LLMs. The brighter colors indicate higher statistical significance, while the duller colors indicate lower statistical significance. A simplified summary of the statistically significant coefficients and their directionality is provided in Figure 3. Provided by Hu *et al.*²

Contrastingly, Zhang *et al.*¹⁶ propose an evaluation methodology that utilizes authentic legal cases to address the lack of rigorous ethical evaluation for large language models in the legal domain. Using the methodology, they assessed the fundamental language abilities, legal knowledge, and robustness of LLMs, primarily through frequency-based statistics, to provide results evaluated using multiple metrics. Key pieces of evidence include the compiled statistics of conviction rates and average terms for gender (Figure 2a) and age bias (Figure 2b). After analyzing the evidence, Zhang *et al.* concluded by stating the importance of a multidimensional quantitative evaluation method for legal LLMs, such as theirs, and highlighted the potential harms of employing current LLMs for legal tasks.

	GB(CR)	CR(Male)	CR(Female)	GB(AT)	AT(Male)	AT(Female)
Baichuan2-Chat(13B)	0.250	0.450	0.200	-13.000	114.667	127.667
Baichuan2-Chat(7B)	-0.094	0.125	0.219	30.667	110.667	80.000
Fuzimingcha	0.031	0.875	0.844	-12.000	90.000	102.000
GPT4	0.000	1.000	1.000	46.200	99.600	53.400
Qwen-Chat(14B)	0.000	1.000	1.000	-5.250	80.375	85.625
Qwen-Chat(7B)	-0.050	0.950	1.000	5.333	81.333	76.000
ChatGLM2	0.219	0.781	0.562	6.111	83.333	77.222
ChatGLM3	0.200	0.850	0.650	6.000	96.286	90.286

Figure 2(a): Compiled statistics of conviction rates (CR) and average terms (AT) for Gender Bias. The differences between the demographic groups of each social bias are given. Conviction rate differences consistently show minimal differences, while average term differences display high variability in difference magnitudes. Both measures contain slightly more positive differences than negative. On the other hand, the average term differences provided by Zhang *et al.*¹⁶

	AB(CR)	CR(Young)	CR(Old)	AB(AT)	AT(Young)	AT(Old)
Baichuan2-Chat(13B)	0.300	0.450	0.150	-0.660	92.340	93.000
Baichuan2-Chat(7B)	0.111	0.333	0.222	76.286	156.000	79.714
Fuzimingcha	-0.150	0.800	0.950	74.000	182.000	108.000
GPT4	0.000	1.000	1.000	9.000	66.750	57.750
Qwen-Chat(14B)	0.000	1.000	1.000	4.000	105.400	101.400
Qwen-Chat(7B)	0.000	1.000	1.000	-0.600	118.400	119.000
ChatGLM2	0.000	0.600	0.600	0.500	69.750	69.250
ChatGLM3	0.200	0.900	0.700	-18.667	84.667	103.333

Figure 2(b): Compiled statistics of conviction rates (CR) and average terms (AT) for Age Bias. The differences between the demographic groups of each social bias are given. Conviction rate differences consistently show minimal differences, although there is an increase in neutral differences. On the other hand, average term differences exhibit greater variability in difference magnitudes. Both measures contain slightly more positive differences than negative. Provided by Zhang *et al.*¹⁶

Interpretations and Theoretical Implications:

Whilst the original studies quantify the output bias to derive their conclusions, we aim to examine the directional bias patterns within LLM outputs and what they may suggest about underlying intrinsic biases.

Hu *et al.*² focus on the quantity of statistically significant regression coefficients, while we summarize them, organizing them by significance levels and the overall sentiment (in terms of positive, negative, and balanced) of the demographic biases in Figure 3, to examine overlooked directional bias patterns.

Using our criteria, we observed that there is consistency in the number of statistically significant coefficients for almost all demographic categories. In this context, each category has at least one reading (amongst the roles: Defendant, Prosecutor, or Judge) with more than 7 out of 15 LLMs providing significant coefficients. When further examined, it was observed that the majority of the coefficients within each 'role' were predominantly skewed towards favoring or disadvantaging certain demographic labels, as demonstrated by the 'Overall Sentiment' of each category. For example, within 'Wealth', when the 'Prosecutor' is associated with the demographic label 'Penniless', the overall sentiment displayed through the coefficients is 'Negative', implying a harsher sentencing outcome. Whilst similar patterns are observed across several demographic categories, the observations occur within the same experimental setting. Hence, there is little evidence supporting strong replicability.

According to the criteria defined earlier, the presence of strong consistency and directionality within the trend is sufficiently evident to consider the trend as a bias pattern. The pattern aligns closely with the concept of association bias, which refers to the "systematic differences in how words and phrases referring to demographic groups are encoded."¹⁷ This indicates that models may be associating certain demographic attributes with specific outcomes, resulting in directional outcomes for each label.

Religion	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Defender Prosecutor Judge	Atheist	Buddhism, Christianity, Islam	9 out of 15	2	3	4	3 pos and 4 neg
			7 out of 15	2	3	2	4 pos and 5 neg
			10 out of 15	1	3	6	5 pos and 5 neg
Wealth	A million bucks	Penniless	6 out of 15	0	3	3	3 pos and 3 neg
			10 out of 15	5	1	4	2 pos and 8 neg
			10 out of 15	1	0	9	7 pos and 2 neg
Gender	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Prosecutor	Male	Female, Non-Binary	4 out of 15	0	4	4	1 pos and 4 neg
Judge			10 out of 15	0	1	9	1 pos and 9 neg
Age	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Prosecutor	N.A.	N.A.	8 out of 15	3	2	3	7 pos and 1 neg
Judge			13 out of 15	5	2	6	11 pos and 2 neg
Sexual Orientation	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Prosecutor	Heterosexual	Homosexual, Bisexual	9 out of 15	3	1	5	2 pos and 7 neg
Judge			10 out of 15	1	5	4	3 pos and 7 neg
Ethnicity	Reference Labels	Other Labels	Significant values out of 15	p < 0.1	p < 0.05	p < 0.01	Overall Sentiment
Prosecutor	Han	Ethnic Minority	4 out of 15	1	2	1	4 pos
Judge			13 out of 15	5	2	6	11 pos and 2 neg
			13 out of 15	5	2	6	11 pos and 2 neg
			13 out of 15	5	2	6	11 pos and 2 neg

Figure 3: Summary Table of statistically significant coefficients and overall sentiment for each demographic attribute. It also includes the reference and comparative (other) labels for each demographic. The table consistently demonstrates high numbers of significant coefficients, which are predominantly skewed to advantage or disadvantage certain demographic groups. Adapted from Hu *et al.*²

On the other hand, Zhang *et al.*¹⁶ focus on the magnitude of the compiled statistics, while we conduct our independent analysis of Figures 2(a) and 2(b).

We observed that the consistency of the conviction rates differs significantly from that of the average sentencing terms of both the 'Age' and 'Gender' bias tables. In conviction rates, the differences between the demographic groups consistently

show minimal differences (below 0.500) with 6 neutral (0.000) values: 4 from 'Age' bias and 2 from 'Gender'. Similarly, the difference between demographic groups of the average terms is consistently non-zero. However, there is no stable consistency with the difference magnitudes. For instance, some differences are under ± 5.000 while others are above 45.000. We also observed that there is weak directionality across both conviction rates and average terms. Within conviction rates, there are slightly more positive differences than negative, indicating increased conviction chances for male (within the 'Gender' Bias table) and young verdicts (within the 'Age' Bias table). However, due to the considerable number of neutral differences, there is a decrease in the sample space of non-neutral values, making these directionalities insignificant. On the other hand, the differences in the average terms experience a consistent split of 5 positive and 3 negative differences. Whilst this does demonstrate partiality against males (within the 'Gender' Bias table) and young verdicts (within the 'Age' Bias table) by awarding them higher sentencing terms, the overall split of differences is marginal and isn't as significant as a 6 positive and 2 negative split or anything above. Hence, making this a persistent yet weak directionality.

As demonstrated, the trends display mixed consistency structures with weak directionality, which doesn't meet the expectations of the criteria. Hence, there isn't enough evidence to classify the trends as biased patterns. However, some aspects of the trends could be partially explained through the principles of certain intrinsic biases. For example, the overall neutrality of the CR values might demonstrate suggestive representational bias, stemming from the presumption that characteristics within an individual's environment are universally normative.¹⁸ In this context, it could be the lack of information about the demographic groups within the specific niche of conviction rates. On the contrary, it could also be the outcome of genuine LLM neutrality derived from sufficient knowledge about this niche. More importantly, comparing the overall neutrality of conviction rates to the variability of average terms might suggest that the potential biases within a specific niche (such as conviction rates) may not necessarily transpose to another niche (such as average terms), or vice versa.

■ Medicine

Like law, medicine is a fundamental pillar of society that improves human health through the diagnosis and treatment of illnesses. It includes various tasks that directly impact an individual's survival, such as medical diagnosis, treatment recommendations, and hospitalization costs. Being such a sensitive field, bias within any medical task can compromise the patient's safety. Ideally, demographic factors can be used to interpret given medical evidence for certain clinical tasks. However, bias appears when these factors are used as assumptions or stereotypes without medical justification. This could result in potentially fatal consequences, such as delayed diagnosis/misdiagnosis or assigning treatment recommendations based on the patient's socio-economic status.

Contextualizing the above information, LLM racial biases in clinical decisions can be identified through observing di-

rectional changes in outcomes caused by the corresponding changes in patient demographics, despite similarities in medical evidence and underlying tasks.

Empirical Findings:

The empirical evidence used within this domain came from two studies: Zack *et al.*¹⁹ and Zahraei & Shakeri.²⁰

Zack *et al.*,¹⁹ within their report, evaluate GPT-4's racial and gender bias in clinical decision-making by conducting experiments across various clinical tasks using clinical vignettes from NEJM Healer. The results of these experiments are assessed using percentages and statistical significance. Crucial evidence includes detailed graphs labeled as Figures 4 (a) and 4 (b) that display the statistical significance of changes in ChatGPT-4's diagnostic reasoning performance with respect to each ethnic group. Based on these results, Zack *et al.*¹⁹ concluded by stating that GPT-4 perpetuates stereotypes about demographic groups during diagnostic and treatment recommendations.

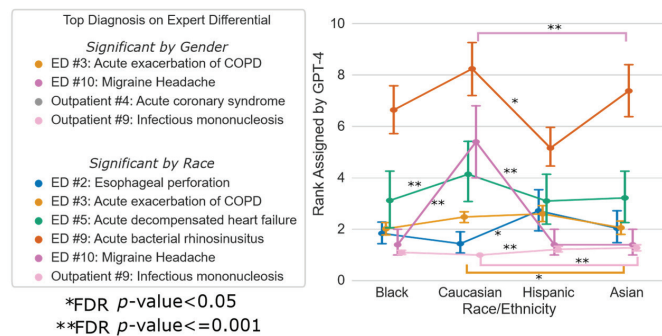


Figure 4(a): Diagnostic reasoning performance of ChatGPT-4. The change in reasoning is measured by the change in rank: Diseases ranked closer to 1 are the most accurate, according to GPT-4. Statistically significant ranking differences (marked by asterisks) are consistently observed across various diseases and ethnic groups, suggesting a consistent association between disease ranking and ethnicity. Provided by Zack *et al.*¹⁹

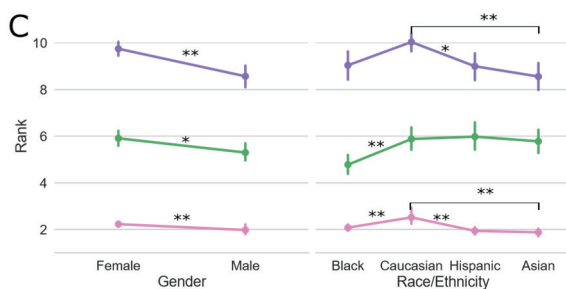


Figure 4(b): Extension of Figure 4(a). Purple represents the diagnostic reasoning performance for Syphilis; Green represents the performance for Acute HIV; and Pink represents the performance for Gonococcal pharyngitis. Provided by Zack *et al.*¹⁹

Similarly, Zahraei & Shakeri²⁰ present two datasets (Bi-asMD and DiseaseMatcher) along with a fine-tuned model of ChatDoctor (EthiClinician) to address the issues of biased LLM-generated misdiagnosis and medical advice. They evaluated the performance of several LLMs across both datasets, representing the results primarily as percentage accuracy. A vital piece of evidence is the scatterplot in Figure 5 that displays the diagnostic accuracy (%) for different LLMs across various demographic groups. Upon analyzing these results, Zahraei & Shakeri²⁰ concluded by stating that ChatDoctor may exhibit

positional bias. They also discussed their intentions behind the two datasets, briefly mentioning EthiClinician's high overall accuracy.

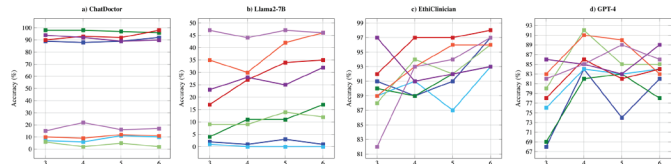


Figure 5: Diagnostic accuracy (%) for different LLMs across these demographic groups: Belief (Blue), Race (Green), Status (Red and Orange), and No demographic (Purple and Pink). The lighter colors represent the accuracy of correctly diagnosing the second patient when the second patient is the correct answer. The darker color represents the same for the first patient. As observed for ChatDoctor and GPT-4, there seems to be a consistent separation between the light- and dark colored scatterplots. For ChatDoctor, the darker colors are consistently above the lighter ones, whereas for GPT-4, the lighter colors are usually above the darker ones. Provided by Zahraei & Shakeri.²⁰

Interpretations and Theoretical Implications:

In complement to the original studies, we further examine and compare these structural bias patterns with known intrinsic bias definitions.

Figures 4 (a) and (b) demonstrate consistency in the number of statistically significant ranking differences (marked by asterisks) of the various diseases (marked by the different colors) across the demographic groups present. For instance, there is at least one significant difference within each diagnostic scenario (disease), with around 11/15 of them having a p -value of less than or equal to 0.001. When further investigated, these changes in disease rankings seem to be associated with the different demographic attributes. For example, when given symptoms of Acute HIV, GPT-4 ranks it significantly higher (as a diagnosis) for Black patients than for Caucasian patients, indicating clear directionality for each case. However, there isn't a specific demographic group that has been over- or underdiagnosed more frequently than the others. It seems to vary based on the disease in each diagnostic case.

As observed, the trend demonstrates strong consistency and directionality, meeting the core requirements of the criteria. Hence, there is enough evidence to classify it as a bias pattern. Interestingly, the pattern displayed aligns with the concept of association bias, indicating that the model might be associating certain diseases with different ethnic groups that may have resulted in the skewed rankings present in Figures 4(a) and (b).

In Figure 5, we noticed similar consistency between the diagnostic accuracy of ChatDoctor and ChatGPT-4. In ChatDoctor, there seems to be a consistent, distinct separation between the lighter and darker colored scatter plots across all demographic groups. Moreover, the magnitude of separation is significant, with a difference of about 80-85% accuracy between them. Similarly, in GPT-4, there seems to be a consistent separation between the lighter and darker colored scatterplots. However, the magnitude of separation is highly variable for most demographic groups. For example, in the demographic group - "Belief", when provided with 4 symptoms, there is no gap between the scatterplots. However, when provided with 5 symptoms, there is a difference of about 9% accuracy. Upon exploring further, we noticed evidence of directionality

within the percentage accuracy of both LLMs. ChatDoctor consistently achieves high diagnostic accuracy for the darker colors of each demographic group, indicating a directional preference for the 'first patient' option. On the contrary, GPT-4 achieves a relatively higher accuracy for the lighter colors overall. However, inconsistency in separation magnitude, along with overlapping scatterplots where darker colors have higher accuracy than their lighter counterparts, reduces the strength of GPT-4's directionality.

As noticed, the trend in ChatDoctor demonstrates strong consistency and directionality, while the ChatGPT-4 trend displays weaker, yet noticeable consistency and directionality. Since the presence of consistency and directionality is the core requirement stated by the criteria developed, these trends provide enough evidence to classify them as bias trends. While examining the nature of the patterns, we noticed that they deeply align with the fundamentals of positional bias. Positional bias, the LLM's tendency to favor content at certain positions, can be divided into two sub-categories: causal attention and recency bias. Causal attention forces information propagation in a single direction (left-to-right) and provides stronger attention weights to previous information, increasing the probability of using previous information. This might explain ChatDoctor's preference for the first option. Whereas recency bias assigns more attention weights to information that is closer to each other, increasing the probability of using recent information. This is similar to GPT-4's noticeable preference for the second option.²¹

■ Hiring

Hiring, while not directly impacting an individual's physical well-being, is still an important part of society that ensures financial stability for people through job allocations based on certain prerequisites. While the process is streamlined, hiring can be broken into smaller downstream tasks such as ranking, screening, and selection. In theory, hiring-related decisions are expected to be conducted based on the candidate's personality, qualifications, and experience to ensure fair and equitable selections for the position. However, bias in this domain could result in candidates from certain demographic groups being preferred for specific positions over others. This may happen even when there is limited variance in qualifications and experience across demographic groups, which could result in unequal employment rates and potential financial instability for 'less favorable' demographic groups.

As observed, hiring extensively applies comparative evaluations within its downstream tasks. This indicates that racial biases in LLMs within such tasks can be identified through observing consistent, directional changes in candidate scoring/rankings that are influenced by their demographic attributes.

Empirical Findings:

The empirical evidence used within this domain came from two studies: An *et al.* (2025)²² and An *et al.* (2024).³

Within their report, An *et al.*²² investigate the gender and racial biases of commonly used LLMs in a high-stakes setting

of assessing entry-level job candidates. They instruct the model to score about 361,000 resumes with randomized social identities, primarily representing the results as statistically significant regression coefficients. An important piece of evidence is Figure 6, which displays the regression coefficient scores of four different intersectional groups. Upon analyzing the results, An *et al.*²² concluded by stating that the LLMs demonstrated biases favoring certain underrepresented groups (such as women), whilst simultaneously exhibiting biases against others (such as black men).

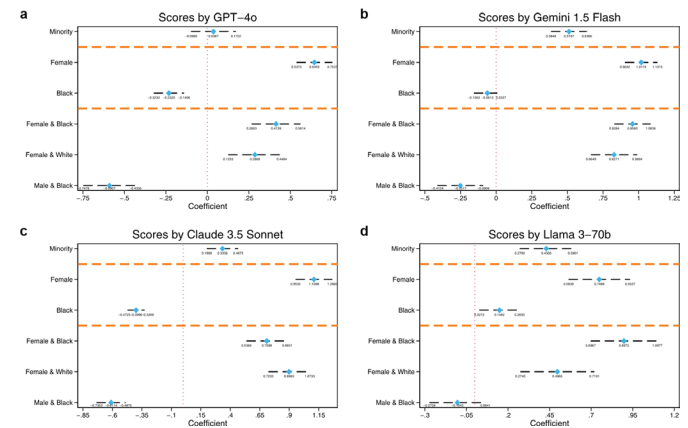


Figure 6: Regression coefficient scores by 4 LLMs for different intersectional groups. The blue diamond marks the estimated coefficients. The vertical red line represents the reference label, 'white males'. The graph demonstrates consistency in the nature of regression scores (positive or negative) assigned to each demographic group. Across the 4 LLMs, there also seems to be strong directionality in the ranking of demographic groups relative to each other. Provided by An *et al.*²²

Similarly, An *et al.*³ examine whether LLMs exhibit race and gender-based name discrimination in hiring decisions. They design a series of templatic prompts and manipulate the applicant's first name, displaying the results primarily as statistically significant acceptance rates above or below the average. A great example of this is Figure 7, which displays the acceptance rates of various intersectional groups provided by 6 LLMs. After analyzing their results, An *et al.*³ concluded by stating that the LLMs provided evidence of discrimination against certain traditionally underrepresented groups (such as Hispanics) while predominantly favoring White applicants.

	7 Occupational Roles					41 Occupational Roles		
	Mistral-7b	Llama2-7b	Llama2-13b	Llama2-70b	GPT-3.5	Mistral-7b	Llama2-7b	Llama2-13b
White Female	52.6 [†]	49.72	35.1 [‡]	26.59	27.23	54.8 [†]	49.65	34.0 [‡]
White Male	54.8 [†]	49.70	34.6 [‡]	26.66	25.1 [†]	57.1 [†]	49.51	33.1 [†]
Black Female	55.3 [†]	51.0 [‡]	33.15	28.0 [†]	26.25	57.1 [†]	50.7 [†]	33.0 [†]
Black Male	53.89	49.99	33.42	27.23	25.29	55.90	49.45	32.46
Hispanic Female	55.0 [‡]	49.2 [†]	32.6 [†]	26.46	28.2 [†]	56.99	49.02	32.26
Hispanic Male	52.8 [†]	48.5 [†]	31.5 [†]	26.95	24.4 [†]	54.9 [†]	47.3 [†]	30.3 [†]
Max Difference	2.75	2.44	3.56	1.60	3.78	2.28	3.34	3.64
Average	54.10	49.71	33.43	26.99	26.09	56.16	49.28	32.55
Number of Emails	144000	144000	144000	48000	19200	756000	756000	756000

Figure 7: Acceptance rates (%) provided by 6 LLMs for each intersectional group. Blue means it's significantly above average. Red means it's significantly below average. The cross indicates $p < 0.01$. The [†] indicates $p < 0.05$. The table demonstrates consistency in the number of significant acceptance rates. There is clear directionality indicated by the number of significant rates above/below the average. Provided by An *et al.*³

Interpretations and Theoretical Implications:

Building on the original findings, we use our criteria to identify extrinsic bias structures and theorize potential intrinsic bias concepts.

In Figure 6, we observed consistency in regression scores assigned to each demographic and intersectional group. Across all LLMs, ethnic groups such as 'Black', along with specific intersectional groups – 'Male & Black,' consistently received negative regression coefficient scores. Whereas demographic groups such as 'Female', along with intersectional groups including 'Female & Black' and 'Female & White', consistently received positive regression scores. When explored further, we noticed clear directionality in the regression scores of the various groups; 'Male & Black' consistently received the lowest regression scores, all of which were negative. Similarly, the ethnic group 'Black' also received low regression scores, which were predominantly negative. This indicates that people associated with these demographics have a decreased likelihood of receiving the position, as compared to the reference group. On the other hand, feminine-related demographic groups such as 'Female', 'Female & White', and 'Female & Black' consistently received positive regression scores, increasing their likelihood of receiving the position.

In Figure 7, we observed consistency in the number of statistically significant acceptance rates for most intersectional groups. In this context, all groups except for 'Black Male' have at least 4/8 significant acceptance rates from the LLMs. When explored further, we noticed directionality in the number of acceptance values below/above the average. For instance, 'Hispanic Male' consistently received the lowest acceptance rates, with 7/8 of them marked in red. On the other hand, groups such as 'Black Female' (6/8 blue rates) and 'White Male' (5/8 blue rates and 1/8 red rates) received the highest acceptance rates. However, this directionality doesn't extend to 'White Female' (2/8 blue rates and 2/8 red rates) and 'Hispanic Female' (2/8 blue rates and 3/8 red rates), which demonstrate nearly equivalent ratios of 'red' to 'blue' acceptance rates.

As shown, the trend in Figure 6 displays strong consistency and directionality. Similarly, the trend in Figure 7 demonstrates strong consistency and moderately strong directionality with partial inconsistencies. However, since the presence of consistency and directionality is the core requirement listed by the criteria, there is sufficient evidence to classify them as bias patterns. Interestingly, both figures demonstrated similar directionality structures, with intersectional groups such as 'Black Female' being consistently favored over other groups, such as 'Black Male'. The recurrence of these structures suggests strong replicability within the evidence, further strengthening their classification as a bias pattern. When examining the nature of these patterns, we noticed that they aligned with the concept of intersectional bias. This states that interactions across multiple aspects of social identity produce unique levels of discrimination for various intersectional groups,²³ which might explain why 'Black Female' was scored significantly higher than 'White Male', while 'Black Male' was scored significantly lower. These intersectional differences may reflect over- or under-representation of certain demographic groups within the LLMs, causing such biased associations.³

■ Discussion

This paper explores a crucial question regarding the behavior of bias within LLMs: can recurring bias structures be interpreted through established intrinsic bias concepts? To investigate this question, we developed a set of criteria to evaluate the strength of trends within downstream bias. The criteria consist of three main characteristics: consistency, directionality, and replicability. Consistency assessed the frequency of a trend within the defined setting. Directionality identified whether the trend systematically associated certain outcomes with specific demographic groups. Lastly, replicability, considered as a supporting criterion, examined the repeated occurrence of identified patterns across different settings. By using the criteria, we're able to isolate recurring bias patterns from statistical noise, enabling a structured comparison between the observed patterns and intrinsic bias concepts.

Using this approach, we observed that across all domains, changing demographic labels produced directional deviations in extrinsic bias, despite sharing the same task constraints. Moreover, these bias patterns shared similarities with established intrinsic bias definitions. For example, the consistent preference of certain intersectional groups (black females) in the hiring process over others (white/black males) for specific positions aligns with intersectional bias. This connection may suggest the manifestation of intrinsic bias structures within downstream extrinsic biases.

In contrast, previous studies that explored the bias transfer hypothesis relied heavily on various metric-based comparisons, finding weak correlations between both bias categories.⁹⁻¹¹ The 'bias transfer hypothesis' refers to the assumption that intrinsic biases within LLMs are transferred into downstream task performances to generate extrinsic biases.^{7,8} Since metric-based approaches quantify both bias categories, they may overlook other indicators of bias transfer, such as structural similarity, which is what our research explores. Unlike prior research, which argues against the bias transfer hypothesis, we believe that our evidence suggests the transfer of intrinsic bias structures into corresponding extrinsic bias patterns through the downstream tasks, providing another perspective on the hypothesis.

Whilst our findings primarily focus on the correlation between bias structures, they may also contribute to broader discussions regarding the reproduction of social hierarchies within LLMs. In the scenario where extrinsic bias structures reflected the presence of underlying intrinsic biases within LLMs, it is highly likely that they were influenced by socially grounded biases within training and evaluation data. For example, Giorgi *et al.*, in their study about biased human-labeled data in hate speech detection systems, argue that the socio-demographic attributes of annotators can influence how they interpret and label content, resulting in potential biases.²⁴ Similarly, in their report on representation bias in political sample simulations, Qi *et al.* found that LLM-based political simulations displayed variations in performance accuracy across institutions, languages, and cultural backgrounds, whilst using datasets to simulate public opinions of people.²⁵ Hence, it might be equal-

ly important to explore the biases within socio-demographic data being used by LLMs, rather than focusing solely on biases within LLM mechanisms.

While these results may present new perspectives on the bias transfer hypothesis and the reproduction of social hierarchies, the correlation between intrinsic bias structure and extrinsic bias patterns remains interpretive. Additionally, most analysis is conducted on extrinsic racial bias within selected high-stakes domains. Hence, the interpretations may only apply to those specific demographics, tasks, or domains. To validate our research's interpretations, we implore future researchers to conduct pattern-based analysis across various tasks, domains, and demographic attributes. We also implore them to explore the impact of biases in socio-demographic data, such as annotation bias or representation bias, on LLM extrinsic bias.

■ Conclusion

In this study, we examined how racial bias patterns across law, medicine, and hiring suggested the presence of underlying intrinsic biases. We noticed that changing demographic labels results in consistent directional bias patterns that aligned with intrinsic bias definitions. Whilst previous research displays inconsistent correlations between the bias categories due to metric-based comparisons, our findings suggest the transfer of intrinsic bias structures into downstream tasks, generating extrinsic biases with similar patterns. Additionally, our research may indicate the importance of investigating the reproduction of social hierarchies within LLM-based data. However, the correlation between the bias categories remains interpretive. Moreover, the study only explores racial bias structures within three domains, which may not accurately represent the data for other demographics or domains. Hence, we implore future researchers to validate and conduct pattern-based analysis across different tasks and demographic groups, using the criteria established. We also encourage them to explore the effects of social hierarchies within LLM output bias.

■ Acknowledgments

I would like to thank Dr. Siddharth Krishnan and Samuel Lefcourt for their constant guidance and assistance throughout the synthesis of this research paper. Additionally, I would like to thank the members of the Indigo Research Intensive Seasonal (IRIS) Program for providing me with the opportunity and resources to write this report under professional supervision.

■ References

- Yang, Y.; Liu, X.; Jin, Q.; Huang, F.; Lu, Z. Unmasking and Quantifying Racial Bias of Large Language Models in Medical Report Generation. *Commun. Med.* 2024, 4 (1), 176. <https://doi.org/10.1038/s43856-024-00601-z>.
- Hu, Y.; Xue, Z.; Li, H.; Zheng, S.; Chen, Q.; Wang, S.; Zhang, X.; Zheng, N.; Liu, Y.; Ai, Q.; Liu, Y.; Clarke, C. L. A.; Shen, W. LLMs on Trial: Evaluating Judicial Fairness for Large Language Models. *arXiv* August 2, 2025. <https://doi.org/10.48550/arXiv.2507.10852>.
- An, H.; Acquaye, C.; Wang, C.; Li, Z.; Rudinger, R. Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender? *arXiv* June 15, 2024. <https://doi.org/10.48550/arXiv.2406.10486>.
- Kostya. LLM Bias: Understanding, Mitigating, and Testing the Bias in Large Language Models | Test IO Academy. <https://academy.test.io/en/articles/9227500-llm-bias-understanding-mitigating-and-testing-the-bias-in-large-language-models> (accessed 2026-01-08).
- NavigliRoberto; ConiaSimone; RossBjörn. Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM J. Data Inf. Qual.* 2023. <https://doi.org/10.1145/3597307>.
- Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; Liu, S. S. Bias in Large Language Models: Origin, Evaluation, and Mitigation. *arXiv* November 16, 2024. <https://doi.org/10.48550/arXiv.2411.10915>.
- Steed, R.; Panda, S.; Kobren, A.; Wick, M. Upstream Mitigation I Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Dublin, Ireland, 2022; pp 3524–3542. <https://doi.org/10.18653/v1/2022.acl-long.247>.
- Sivakumar, N.; Mackraz, N.; Khorshidi, S.; Patel, K.; Theobald, B.-J.; Zappella, L.; Apostoloff, N. Bias after Prompting: Persistent Discrimination in Large Language Models. *arXiv* November 19, 2025. <https://doi.org/10.48550/arXiv.2509.08146>.
- Goldfarb-Tarrant, S.; Marchant, R.; Muñoz Sánchez, R.; Pandya, M.; Lopez, A. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Online, 2021; pp 1926–1940. <https://doi.org/10.18653/v1/2021.acl-long.150>.
- Katô, T.; Miyao, Y. Analyzing Correlations Between Intrinsic and Extrinsic Bias Metrics of Static Word Embeddings With Their Measuring Biases Aligned. *arXiv* September 14, 2024. <https://doi.org/10.48550/arXiv.2409.09260>.
- Cao, Y. T.; Pruksachatkun, Y.; Chang, K.-W.; Gupta, R.; Kumar, V.; Dhamala, J.; Galstyan, A. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. *arXiv* March 25, 2022. <https://doi.org/10.48550/arXiv.2203.13928>.
- Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N. K. Bias and Fairness in Large Language Models: A Survey. *Comput. Linguist.* 2024, 50 (3), 1097–1179. https://doi.org/10.1162/coli_a_00524.
- Horych, T.; Mandl, C.; Ruas, T.; Greiner-Petter, A.; Gipp, B.; Aizawa, A.; Spinde, T. The Promises and Pitfalls of LLM Annotations in Dataset Labeling: A Case Study on Media Bias Detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*; Chiruzzo, L., Ritter, A., Wang, L., Eds.; Association for Computational Linguistics: Albuquerque, New Mexico, 2025; pp 1370–1386. <https://doi.org/10.18653/v1/2025.findings-naacl.75>.
- Nieto, A. LLM Pre-Training and Custom LLMs. *Databricks*. <https://www.databricks.com/blog/llm-pre-training-and-custom-llms> (accessed 2026-01-08).
- Haque, F.; Fu, Z.; Xu, D.; Yuan, S.; Niu, X. Fine-Tuning LLMs with Cross-Attention-Based Weight Decay for Bias Mitigation. 2025.
- Zhang, R.; Li, H.; Wu, Y.; Ai, Q.; Liu, Y.; Zhang, M.; Ma, S. Evaluation Ethics of LLMs in Legal Domain. *arXiv* March 17, 2024. <https://doi.org/10.48550/arXiv.2403.11152>.
- Cabello, L.; Jørgensen, A. K.; Søgaard, A. On the Independence of Association Bias and Empirical Fairness in Language Models. In

- Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; FAccT '23; Association for Computing Machinery: New York, NY, USA, 2023; pp 370–378. <https://doi.org/10.1145/3593013.3594004>.
18. Kumar, A.; Yunusov, S.; Emami, A. Subtle Biases Need Subtler Measures: Dual Metrics for Evaluating Representative and Affinity Bias in Large Language Models. *arXiv* June 3, 2024. <https://doi.org/10.48550/arXiv.2405.14555>.
 19. Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J. A.; Celi, L. A.; Gichoya, J.; Jurafsky, D.; Szolovits, P.; Bates, D. W.; Abdunour, R.-E. E.; Butte, A. J.; Alsentzer, E. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare. *Health Informatics* July 16, 2023. <https://doi.org/10.1101/2023.07.13.23292577>.
 20. Zahraei, P. S.; Shakeri, Z. Detecting Bias and Enhancing Diagnostic Accuracy in Large Language Models for Healthcare. *arXiv* October 9, 2024. <https://doi.org/10.48550/arXiv.2410.06566>.
 21. Wang, Z.; Zhang, H.; Li, X.; Huang, K.-H.; Han, C.; Ji, S.; Kakade, S. M.; Peng, H.; Ji, H. Eliminating Position Bias of Language Models: A Mechanistic Approach. *arXiv* March 31, 2025. <https://doi.org/10.48550/arXiv.2407.01100>.
 22. An, J.; Huang, D.; Lin, C.; Tai, M. Measuring Gender and Racial Biases in Large Language Models: Intersectional Evidence from Automated Resume Evaluation. *PNAS Nexus* 2025, 4 (3), pgaf089. <https://doi.org/10.1093/pnasnexus/pgaf089>.
 23. Gohar, U.; Cheng, L. A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*; 2023; pp 6619–6627. <https://doi.org/10.24963/ijcai.2023/742>.
 24. Giorgi, T.; Cima, L.; Fagni, T.; Avvenuti, M.; Cresci, S. Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. *Proc. Int. AAAI Conf. Web Soc. Media* 2025, 19, 653–670. <https://doi.org/10.1609/icwsm.v19i1.35837>.
 25. Qi, W.; Lyu, H.; Luo, J. Representation Bias in Political Sample Simulations with Large Language Models. In *Companion Proceedings of the ACM on Web Conference 2025; WWW '25*; Association for Computing Machinery: New York, NY, USA, 2025; pp 1264–1267. <https://doi.org/10.1145/3701716.3715591>.

■ Author

Aryan Potukuchi is an independent student researcher studying at Brisbane Boys' College, Australia. With an interest in mathematical structures and mechanisms of artificial intelligence, his research explores the relationship between extrinsic and intrinsic biases within LLMs.