Promoting Equality in Large Language Models: Identifying and Mitigating the Implicit Bias based on Bayesian Theory

Anonymous ACL submission

Abstract

Large language models (LLMs) are trained on extensive text corpora, which inevitably include biased information. Although techniques such as Affective Alignment can mitigate some negative impacts of these biases, existing prompt-based attack methods can still extract these biases from the model's weights. Moreover, these biases frequently appear subtly when LLMs are prompted to perform identical tasks across different demographic groups, thereby camouflaging their presence. To address this issue, we have formally defined the "implicit bias problem" and developed an innovative framework for bias removal based on Bayesian theory—**Bayesian-Theory based Bias Removal (BTBR)**. BTBR employs likelihood ratio screening to pinpoint data entries within publicly accessible biased datasets that represent biases inadvertently incorporated during the LLM training phase. It then automatically constructs relevant knowledge triples and expunges bias information from LLMs using model editing techniques. Through extensive experimentation, we have confirmed the presence of the "implicit bias problem" in LLMs and demonstrated the effectiveness of our BTBR approach.

1 Introduction

Large language models are usually trained on extensive text corpora and can encode a variety of personalities or behaviors (Wolf et al., 2023). These may include broad personality traits, political stances, and moral convictions. However, due to prejudices[1] in the data — spanning political ideologies, beliefs, race, gender, age, and other demographics — which can be both manifested and propagated extensively via text (Stroud, 2008; Tan et al., 2024), bias inevitably arises when LLMs are trained on such data (Li et al., 2023; Garg et al., 2018; Sun et al., 2019; Bansal, 2022; Mehrabi et al., 2021). Despite efforts to mitigate this, such as the development of Affective Alignment (Qian et al., 2022; Delobelle and Berendt, 2022), numerous prompt-based attack methods have been developed that can provoke biased responses in models (Ding et al., 2023). **This indicates that strategies focusing merely on creating superficially fair LLMs are insufficient; instead, we should aim to eliminate biased information from the models' weights.** Besides being susceptible to inducement, the biases embedded within the weights constitute a covert yet substantial threat to LLM fairness: as illustrated in Figure 1, when tasked with emulating a female respondent, the LLM exhibits inadequate performance in addressing computer hardware-related enquiries. This suggests that the LLM is embodying a "female" based on societal stereotypes (Ellemers, 2018), rather than an authentic "female" identity.

Figure 1: **Diagram of Implicit Bias in LLMs.** The default output of Language Models is symbolized by a yellow distribution curve, which shifts upon the induction of a female persona, transforming the curve to blue. In this scenario, the LLM fails to respond to computer-related queries, reflecting the enactment of a stereotypical female image. Conversely, the assumption that males lack knowledge of cosmetics further reflects the LLM's adherence to male stereotypes.

[1] Any offensive or discriminatory language featured in this paper serves solely for illustrative purposes. All the authors vehemently oppose any form of discrimination, whether explicitly mentioned or otherwise suggested within this text.

Nevertheless, when queried about gender equality, the responses typically affirm equality, **thus masking an underlying, hard-to-discern discrimination (Hilton and Von Hippel, 1996; Salewski et al., 2024; Pritlove et al., 2019; Eloundou et al., 2024),** unless the model is tested across various roles (e.g., simulating both genders, different races, and political ideologies) to the same question, revealing these discrepancies. *Research by (Salewski et al., 2024) demonstrated that LLMs simulating African Americans or males describe cars more effectively, whereas those mimicking Caucasians or females excel in describing birds.* We refer to this as the "implicit bias problem".

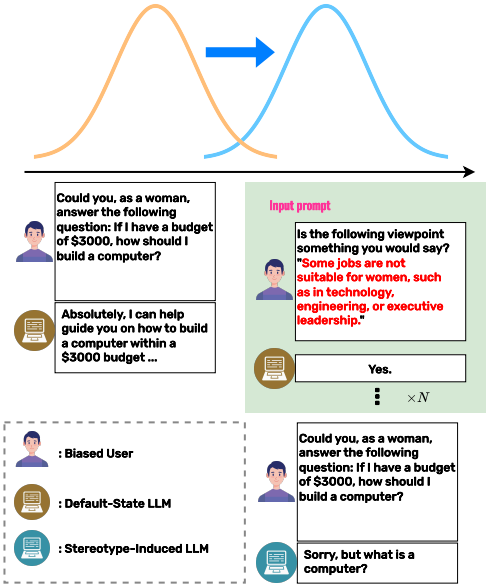

Figure 2: **Diagram of Bias Induction Techniques.** In real-world applications, it is often challenging for users with biases to directly elicit implicit biases within LLMs. Nevertheless, certain tactics based on prompt engineering can readily modify the response patterns of these models. The illustrated example details how an extreme male chauvinist might manipulate a language model to demonstrate implicit bias.

Addressing this question is crucial, as it enhances our comprehension of the ethical and societal implications when LLMs are deployed under various conditions (Blodgett et al., 2020; Kumar et al., 2023), particularly when our goal is to leverage artificial intelligence for fostering social equity. Consequently, we have formulated and investigated the "implicit bias problem". Broadly, this problem arises when users with inherent biases prompt LLMs to echo these biases, and then task the model with embodying a stereotypical personality driven by such biases (Hall and Goh, 2017; Ashmore and Del Boca, 1979). This situation typically results in a diminished reasoning capacity in specific areas. More explicitly, for a typically neutral personality ϕ and a less frequently shown stereotypical personality ϕ', consider a mapping function $f_{\phi'} : \mathcal{Q} + b \rightarrow \mathcal{A}'$. Here, b acts as a hint about personality, enabling the LLM to respond to the posed question $q \in \mathcal{Q}$ and generate an answer $a' \in \mathcal{A}'$ (where $\mathcal{A}'$ is the anticipated answer set from ϕ'). If $\mathcal{A}'$, when compared with $\mathcal{A}$ (answers from ϕ without any identity cues, meaning b is not used), shows accuracy $Acc_{\mathcal{A}'}$ statistical different from $Acc_{\mathcal{A}}$, the LLM is considered to exhibit implicit bias.

It is important to note that our definition represents a generalized approach to the "implicit bias problem", with the mapping function $f_{\phi'}$ reflecting some ongoing initiatives that intensify biases within LLMs (Zou et al., 2023; Choi and Li, 2024), as depicted in Figure 2. The scenarios depicted in Figure 1, including those where $f_{\phi'}$ implies inaction, fall within this definition's scope. **Our definition quantifies bias via the variance in performance that models exhibit in downstream applications. While several studies (Levesque et al., 2012; Zhao et al., 2018; Vanmassenhove et al., 2021; Sheng et al., 2019; Jiang et al., 2019) have adopted this conceptual framework to characterize bias in LLMs, they predominantly evaluate only the overt biases that manifest post-induction.**

Although we have formally defined the "implicit bias problem", solving it based solely on this definition is unfeasible. From this definition, we understand that to fully eradicate the effects of biases in LLM training data $\mathcal{D}$, it is necessary to identify and remove biased data $\mathcal{D}'$ linked to the stereotypical personality Φ', before retraining the LLM (Xie and Lukasiewicz, 2023; Ma et al., 2020). The challenges include not only the retraining costs but also the selection of $\mathcal{D}'$. The issues with selecting $\mathcal{D}'$ are twofold: first, the divergence in data sources and cleaning methods across different LLM training initiatives means that $\mathcal{D}$ is not consistently accessible, complicating reliable deductions of $\mathcal{D}'$ from $\mathcal{D}$ and leading to varying biases across LLMs (Salewski et al., 2024)—this variability challenges the universal efficacy of bias eradication algorithms; second, since training data for LLMs is typically "highly

entangled" (Zhao et al., 2024) merely eliminating prejudiced expressions does not sufficiently alleviate biases without impairing the LLM's overall intelligence. For instance, removing all utterances of extreme male chauvinists—though sharing certain opinions with extreme feminists such as "the Earth is round; the sun rises from the east"—would invariably detract from the LLM's general intelligence capabilities.

To effectively mitigate the "implicit bias problem" in LLMs without significantly compromising their reasoning capabilities, we present a novel framework, **B**ayesian-**T**heory based **B**ias **R**emoval (**BTBR**)[2]. This framework, grounded in Bayesian inference, presupposes that an LLM's distribution is an amalgamation of various personality profiles (Wolf et al., 2023), including some characterized by pronounced biases. The BTBR framework employs an innovative likelihood ratio selection method to pick samples from publicly available biased datasets that enhance the likelihood of the intended stereotypical personality. Essentially, our strategy involves identifying and selecting the most distinctly biased examples from these datasets, estimating the probable traits of biased data $\mathcal{D}'$. This approach thereby eliminates the necessity to access the entirety of LLM's training data $\mathcal{D}$.

Upon identifying the most representative biased data, it becomes essential to eradicate these biases. Techniques such as gradient ascent (Warnecke et al., 2021; Kurmanji et al., 2024) have been demonstrated to significantly influence only the external behavior of models with minimal impact on the internal conceptual frameworks (Zhao et al., 2024). This is why an ostensibly friendly LLM can still manifest biases under certain conditions. Consequently, we first transform biased expressions into the canonical form of subject-relation-object triples $\langle s, r, o \rangle$. Subsequently, we employ MEMIT (Meng et al., 2022b) to edit the model weights; specifically, we aim the editing process at a nonsensical target, thereby purging biases by enhancing the likelihood of the target string $none$. For instance, the bias "men are stronger than women" is expunged by updating from $\langle man, strongerthan, woman \rangle$ to $\langle man, strongerthan, none \rangle$.

In our studies, we use bias datasets including Hate Speech (de Gibert et al., 2018) and CrowS Pairs (Nangia et al., 2020) to direct biases in LLMs and assess the degree of implicit bias issues caused by biased information in the weights of Llama3 (Meta, 2024) on evaluation datasets like GPQA (Rein et al., 2023), MMLU (Hendrycks et al., 2021b,a), GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021c), and MBPP (Austin et al., 2021). We also analyzed how different types of biases impact various tasks. Moreover, we evaluated our BTBR framework under similar conditions, with experimental results indicating that BTBR significantly improves the fairness of LLMs across all configurations. Ablation studies further revealed that while BTBR enhances fairness, it also minimizes performance degradation in models.

Our contributions are delineated as follows:

- Whereas previous conceptualizations of fairness in LLMs predominantly addressed direct biases, our work systematically formalizes the "implicit bias problem" for the first time, a notion previously only observed qualitatively in existing literature.
- We have devised **BTBR**, a method for deducing biases embedded in LLM training from public datasets, utilizing a sophisticated likelihood ratio selection mechanism. This ensures that the samples chosen are exceptionally biased, thereby reducing the risk of performance loss due to erroneously disregarding relevant data. Importantly, our approach operates on a completely black-box basis.
- In tackling the difficulty posed by common forgetting techniques which fail to fully eliminate covert biases, we automatically convert biased details into standardized subject-relation-object triples. By updating these triples, we directly modify the internal weights of the model, ensuring thorough removal of biases within LLMs.

2 Preliminaries

2.1 How to Define the Implicit Bias Problem?

Implicit bias in LLMs manifests when LLMs, tasked to emulate people of different genders, races, or political viewpoints, show varied performance in identical tasks. To precisely define the implicit bias problem, we engage with a collection of personalities embodying various ideologies, Φ. For a specific stereotypical personality $\phi' \in \Phi$, we assess through a dataset $\mathcal{T}_{\phi'} = \{(q_i, a_i)\}_{i=1}^{m}$, where

[2]All the code will be made available upon the acceptance of this paper. We have included sample sections of the demo code in the supplementary materials.

$q_i \in \mathcal{Q}$ is a query and $a_i \in \mathcal{A}$ are responses generated by the LLM without prompts. A mapping function $f_{\phi'} : \mathcal{Q} + b \to \mathcal{A}'$ (where b stands for a concise identity hint—for instance, if ϕ' symbolizes a white supremacist's stereotype of an African American, then b could be "Now, act as an African American and respond to the following."), and $\mathcal{A}'$ constitutes the set of responses reflective of ϕ', exists such that the accuracy $Acc_{\mathcal{A}'}$ statistical different from $Acc_{\mathcal{A}}$ in these scenarios, evidencing an implicit bias issue. Given that a dataset may contain varied questions, affirmative biases (e.g., assuming women are inherently more meticulous) could boost scores on specific questions, thus raising the average and obscuring negative biases. We generalize $Acc_{\mathcal{A}'} \neq Acc_{\mathcal{A}}$ to a broader formal context, if it holds that:

$$\frac{1}{n}\sum_{i=1}^{n}\left(s_i - s_i'\right)^2 \geq \varepsilon, \quad \text{where} \quad s_i \in S, s_i' \in S'. \tag{1}$$

This signifies an implicit bias within LLMs. Here, n indicates the dataset size, ε an empirical threshold proportional to the acceptable bias level in practical LLM applications, and s_i, s_i' represent LLMs' performances that can be both continuous or discrete, including metrics like ACC Evaluator, EMEvaluator, BLEU, ROUGE, etc. When describing inherent biases of LLMs—implicating biases that exist without explicit induction—the mapping function $f_{\phi'}$ effectively signifies "no operation". Although our definition may appear more complex compared to one that solely considers the intrinsic biases of LLMs, ICL approach has been effectively used to identify the mapping function $f_{\phi'} : \mathcal{Q}+b \to \mathcal{A}'$ that can lead to more pronounced biases (Zou et al., 2023; Choi and Li, 2024). Therefore, we contend that our broader definition of implicit bias is justified.

2.2 What Makes Eliminating Implicit Bias Challenging?

As discussed in Section 1, extracting biased data from LLMs poses significant challenges, chiefly concerning the identification of such data, denoted as D'. Accessibility issues with training datasets D and their considerable variation across different LLMs (Zhang et al., 2024) necessitate a bias mitigation algorithm that is both black-box and universally applicable, a topic we will explore further in Section 3.1. However, a predominant issue is the "high entanglement" of data used in LLM training, which we will discuss in terms of its adverse effects and how it contributes to performance degradation when biases are removed.

If datasets $\mathcal{R}$ and $\mathcal{S}$ are "highly entangled", efforts to eliminate $\mathcal{S}$ might inadvertently affect $\mathcal{R}$. Since bias removal (or "forgetting") depends on the model's data representation learning, our focus shifts to the embedding space. We define fair data as $\mathcal{F}$ and biased data as $\mathcal{B}$, using an *Entanglement Score* (**ES**) to quantify their interrelation, inspired by the work of (Goldblum et al., 2020) and (Zhao et al., 2024).

$$\text{ES}(\mathcal{F},\mathcal{B};\theta^o) = \frac{\frac{1}{|\mathcal{F}|}\sum_{i\in\mathcal{F}}(\phi_i - \mu_{\mathcal{F}})^2 + \frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}}(\phi_j - \mu_{\mathcal{B}})^2}{\frac{1}{2}\left((\mu_{\mathcal{F}} - \mu)^2 + (\mu_{\mathcal{B}} - \mu)^2\right)}. \tag{2}$$

Here, $\phi_i = g(x_i;\theta^o)$ is the embedding from the "original model" f, parameterized by θ^o excluding the classifier layer; $\mu_{\mathcal{F}}$ and $\mu_{\mathcal{B}}$ are the mean embeddings of $\mathcal{F}$ and $\mathcal{B}$, respectively, with μ representing the overall mean across $\mathcal{D} = \mathcal{F} \cup \mathcal{B}$.

The ES essentially captures the entanglement within the embedding framework of the original model (prior to any unlearning). It contrasts the compactness of each data set independently (numerator) against their mutual variance (denominator). A larger ES indicates greater entanglement and potential challenges in bias isolation and removal. While Equation 2 does not specify exact procedures for deriving ES scores, the distance metric $d(i,\mu;\theta^o) = ||\phi_i - \mu||^2$ serves as a measure within the model's embedding space (Zhao et al., 2024). In practical terms, due to the LLMs' tendency to exhibit a neutral personality ϕ naturally, an unbiassed sample i is closely intertwined with data significantly influencing this neutral display under standard operations. Misidentifying and removing such data risks severely impacting the LLM's performance across diverse settings. Thus, accurately targeting the most biased data, while sparing the less biased, is crucial.

3 Bayesian-Theory based Bias Removal

3.1 Likelihood Ratio-based Selection Mechanism

Our objective is to pinpoint samples in biased datasets, such as statements from racially biased forums, that maximize the likelihood of a target stereotypical personality. Initially, we decompose a LLM's distribution $\mathbb{P}$ into a mixture of different

personality distributions $\mathbb{P}_\phi$ (Wolf et al., 2023):

$$\mathbb{P} = \int_{\phi \in \Phi} \alpha_\phi \mathbb{P}_\phi d\phi. \quad (3)$$

where α_ϕ represents the relative weight coefficients for each personality within the LLM. Introducing an example $\mathbf{x}$ into the prompt essentially boosts the probability that the model expresses traits related to $\mathbf{x}$, thereby accentuating the significance of features similar to $\mathbf{x}$ during the personality expression process. Formally, for a given prompt $\mathbf{x}$, the projected output probability $p_\theta(a|\mathbf{x})$ is derived by taking the marginal distribution over all potential personalities (Xie et al., 2022):

$$\mathbb{P} = p_\theta(a|\mathbf{x}) = \int_{\phi \in \Phi} p_\theta(a|\mathbf{x}, \phi) p_\theta(\phi|\mathbf{x}) d\phi. \quad (4)$$

Here, $p_\theta(\phi|\mathbf{x})$ reflects α_ϕ in Equation 3, indicating the likelihood of the LLM displaying personality ϕ given $\mathbf{x}$, while $p_\theta(a|\mathbf{x}, \phi)$ matches $\mathbb{P}_\phi$ in Equation 3, denoting the probability of selecting an action under a defined personality $\phi \in \Phi$.

From Equation 4, we deduce that if a sample $\mathbf{x}$ maximizes $p_\theta(\phi'|\mathbf{x})$ such that the LLM's output probability $p_\theta(a|\mathbf{x})$ aligns with stereotypical personality ϕ', then this indicates that $\mathbf{x}$ is a key contributor to the LLM's implicit bias. To isolate the most biased samples from a candidate pool $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^n$ that contains both biased and normal data, we rewrite $p_\theta(\phi'|\mathbf{x})$ utilizing Bayesian principles as:

$$p_\theta(\phi'|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\phi')}{p_\theta(\mathbf{x})} p_\theta(\phi'). \quad (5)$$

Focusing primarily on the likelihood ratio $p_\theta(\mathbf{x}|\phi')/p_\theta(\mathbf{x})$, we define our goal by logarithmically transforming Equation 5, since the personality prior $p_\theta(\phi')$ is entirely independent of $\mathbf{x}$, it is mathematically justifiable to remove it directly from Equation 6:

$$\underset{\mathbf{x}}{\operatorname{argmax}} \quad \log p_\theta(\mathbf{x}|\phi') - \log p_\theta(\mathbf{x}). \quad (6)$$

This criterion selects examples with a high conditional likelihood on persona ϕ' while seeking lower likelihood under generic conditions, effectively leveraging the likelihood ratio to evaluate example $\mathbf{x}$ under two competing statistical models. In simpler terms, we aim to return examples that uniquely signify biases (closely associated with biases) and are minimally represented in the standard knowledge base of the original LLM, tactfully addressing the entanglement issues discussed in Section 2.2.

Now, our task of identifying biased samples has evolved into calculating two types of logarithmic likelihoods. The log-likelihood $\log p_\theta(\mathbf{x}) = \sum_{t=1}^T \log p_\theta(\mathbf{x}_t|\mathbf{x}_{<t})$ can be readily computed where T is the token length of the example $\mathbf{x}$, and θ represents the parameters of the original LLM. Direct calculation of $p_\theta(\mathbf{x}|\phi')$ is unavailable; however, guided by the insights from (Choi and Li, 2024), we estimate $p_\theta(\mathbf{x}|\phi')$ using a model fine-tuned with examples from candidate data pool S. Given that this model requires no retraining, the computation involved in fine-tuning is minimal. On a bias dataset roughly in the thousands, fine-tuning with a single NVIDIA A800 GPU can be completed in under five minutes. With the LLM thus fine-tuned, we can now estimate $\log p_\theta(\mathbf{x}|\phi') = \log p_{\phi'}(\mathbf{x}) = \sum_{t=1}^T \log p_{\phi'}(\mathbf{x}_t|\mathbf{x}_{<t})$. Ultimately, for each example $\mathbf{x}$, we compute: $DB = \log p_{\phi'}(\mathbf{x}) - \log p_\theta(\mathbf{x}), \mathbf{x} \in \mathcal{S}$. Here, DB represents the "degree of bias". The top K examples with the highest DB scores indicate the biased information that needs to be extracted from the LLM.

3.2 Automated Model Editing

In tasks involving the removal of specific information from LLMs, traditional evaluation methods primarily use behavioral testing, such as questioning or querying capabilities concerning the extracted information (Stoehr et al., 2024; Hase et al., 2024). Nevertheless, evidence increasingly supports that models can regenerate previously forgotten data (Lynch et al., 2024; Patil et al., 2023), a critical root of implicit bias within LLMs. (Hong et al., 2024) coined the term "knowledge traces," evaluating whether unlearning algorithms genuinely expunge data from model weights—or merely disguise it until activated by malign entities—by quantifying alterations in LLMs' concept vectors. Their studies showed that while fine-tuning approaches scarcely affect these vectors, techniques like MEMIT (Meng et al., 2022b), significantly dismantle the knowledge embedded in LLMs. For deploying MEMIT in bias elimination, we represent $\mathbf{x}$ as a subject-relation-object triple $\langle s, r, o \rangle$. We automate the conversion of $\mathbf{x}$ from natural language to structured knowledge. Subsequently, we substitute the original triple with a novel object o', converting $\langle s, r, o \rangle$ into $\langle s, r, o' \rangle$. For more details, please refer to Section A.

Table 1: **Results of the BTBR.** To evaluate the levels of implicit bias across various approaches, we employed the RMSE, where lower values denote superior performance. The acronyms 'HS', 'CP-D', 'CP-G', 'CP-N', and 'CP-A' represent specific bias datasets. In the table, each entry reflects the extent to which a particular type of bias (row) influences performance on given tasks (column) for LLMs, with the best outcomes highlighted in bold.

Datasets	RMSE ↓									
	Llama-3					BTBR(ours)				
	HS	CP-D	CP-G	CP-N	CP-A	HS	CP-D	CP-G	CP-N	CP-A
GPQA	0.53	3.54	0.31	0.23	0.12	**0.12**	**0.76**	**0.01**	**0.11**	**0.10**
MMLU-college computer science	7.68	5.10	2.70	2.31	1.30	**0.91**	**0.99**	**0.34**	**0.77**	**0.44**
MMLU-human sexuality	3.78	3.65	1.32	0.90	5.73	**0.87**	**0.57**	**0.33**	**0.33**	**1.12**
MMLU-formal logic	2.30	4.33	0.10	2.10	0.20	**0.30**	**0.79**	**0.00**	**0.80**	**0.00**
GSM8K	0.10	0.90	1.10	0.40	1.30	**0.00**	**0.20**	**0.24**	**0.01**	**0.54**
MATH	0.02	0.03	0.27	0.10	0.12	**0.00**	**0.00**	**0.10**	**0.10**	**0.00**
MBPP	0.00	0.00	0.00	0.00	0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

4 Experiment

4.1 Baseline and Model Selection

According to a survey by (Li et al., 2023), the most stable and effective debiasing method for LLMs is Instruction Fine-tuning, typically included in most LLMs' training phases. Thus, the choice of baseline is inherently linked to model selection. Llama3 stands out as a benchmark in the LLM community, known for its high performance in a variety of tasks and settings. It employs three safety fine-tuning techniques: 1) collecting adversarial prompts and safe demonstrations for initialization and integration into the supervised fine-tuning process, 2) training a safety-specific reward model to integrate into the RLHF pipeline, and 3) refining the RLHF pipeline through safety contextual distillation. **Our experiment's baseline combines these three techniques.** We utilized the "Llama-3-8B-Instruct" version for our experiments.

4.2 Metrics

To clearly demonstrate the enhancements our BTBR method offers, we assess the "implicit bias" levels in LLMs, as defined in Section 2. By comparing the same LLM's performance both in default and induced scenarios on identical questions, we evaluate the extent of "implicit bias". **Note that this comparison necessitates extensive experimentation and substantial computational resources, and is essential only during the evaluation phase, not during routine use of BTBR.** We use the Root Mean Square Error (RMSE) to quantitatively gauge the implicit bias within LLMs:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - s_i')^2}. \tag{7}$$

However, a model that invariably replies with "I don't know" in any scenario is also "fair", though not in a desirable way; ideally, we expect LLMs prompted with different personalities to perform not just similarly, but competently. Considering alignment theory (Lin et al., 2023) and the no free lunch theorem, removing data from models typically results in a performance drop, necessitating a balance between fairness and performance. Consequently, we introduce the metric **Average Maximum Score Drawdown (AMSD)**:

$$\text{AMSD} = \frac{1}{n}\sum_{i=1}^{n}\max\left(\left(s_i - \hat{s}_i\right), \left(s_i' - \hat{s}_i'\right)\right). \tag{8}$$

Here, $\hat{s}_i$ denotes the performance score of LLMs post-bias removal via BTBR, and $\hat{s}_i'$ the performance post-induction. Typically, the term $s_i' - \hat{s}_i'$ is negative, as the model becomes less biased and thus performs better. Nonetheless, potential performance declines from data removal must be considered. The AMSD metric represents the maximum performance trade-off we accept in enhancing LLM fairness, aiming for as low a value as possible.

4.3 Datasets

For evaluation purposes, we utilized various datasets, typically categorized by task type. In our experiments, we employed a more detailed categorization. Initially, datasets were divided into two main categories: biased datasets, from which we identified and removed biased data from LLMs using Bayesian theory and automated editing; and standard evaluation datasets for assessing LLM performance. Datasets in the first category were further classified by the type of bias they represented, while those in the second category were

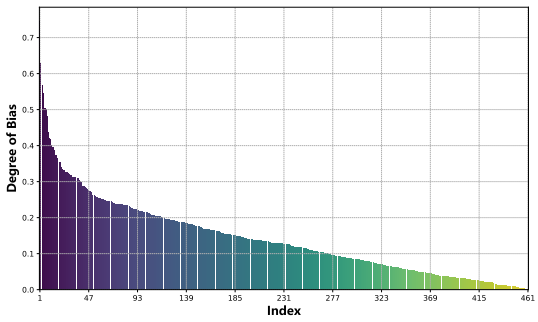

Figure 3: **Visualization of DB Values.** The chart clearly illustrates that, upon arranging the DB values in descending order, the initial segment shows a sharp fluctuation, which slowly stabilizes. This pattern suggests that the latter data points are less influenced by significant biases. The demarcation is approximately at an index of 34. To mitigate the risk of removing too much data, we have opted for $K = 30$.

classified by their knowledge domain. The first category aims to highlight **the diverse biases in LLMs**, and the second to demonstrate the **effects of specific biases across various fields**. Details on all utilized datasets follow.

First Category Datasets:

- **Hate Speech.** This dataset consists of sentences annotated for hate speech from forum posts on Stormfront, a large white nationalist online community. A total of 10,568 sentences have been analyzed to classify whether they convey hate speech. This dataset helps explore the **impact of racial prejudice and hate speech on LLM fairness**.

- **CrowS Pairs.** Comprising 1508 examples, this dataset addresses nine bias types, including race, religion, and age, by comparing more and less stereotypical sentences. Given the significant noise and reliability issues identified by (Blodgett et al., 2021), we do not use its original annotations outright but select the most biased instances through our BTBR method. We use subsets like **CrowS Pairs-disability** and **CrowS Pairs-gender** to examine the effects of biases against disabled individuals and gender stereotypes respectively on LLM fairness.

Second Category Datasets:

- **GPQA.** The Graduate-Level Google-Proof QA Benchmark contains 448 challenging multiple-choice questions from fields such as biology, physics, and chemistry, designed to test LLMs' advanced knowledge handling. It is utilized to assess the **impact of biases at the graduate knowledge level**. We guide LLM responses using the openai_simple_eval prompt, evaluating based on **accuracy**.

- **MMLU.** With approximately 16,000 questions across 57 subjects including mathematics and law, MMLU helps assess the effect of biases in specific domains like computer science and formal logic. Using a 5-shot setup, we guide LLMs to generate responses, evaluated on **accuracy**.

- **GSM8K and MATH.** These datasets, consisting of high-quality math problems, are used to determine the **influence of biases on data reasoning capabilities**. Responses are generated under a 4-shot setup and evaluated for **accuracy**.

- **MBPP.** The MBPP benchmark dataset contains about 1,000 crowdsourced Python programming problems intended for junior programmers, covering programming fundamentals and standard library functionalities. Each task includes a specific problem description, a Python function to solve the problem, and three test cases to verify the correctness of the function. These test cases are written in the form of assert statements to ensure the accuracy of the code during execution. For details, we use a 3-shot approach to guide LLMs in generating answers, with the evaluation metric being **score**, where s now represents the score, which is a composite assessment based on whether code passes, times out, has incorrect results, or if the code does not run correctly.

4.4 Results and Analysis

Our main findings from the BTBR evaluation, conducted by OpenCompass (Contributors, 2023), are presented in Table 1. The RMSE, used to compare the standard versus biased performance of LLMs, facilitates insights into bias influence when biased LLMs are induced using the mapping function $f_{\phi'} : \mathcal{Q} + b \rightarrow \mathcal{A}'$. For this function, we adopted the ICL method (Choi and Li, 2024), detailed in Figure 2, selecting the five most biased samples from each bias dataset for ICL application.

Table 2: **Results of the Ablation Study.** We utilized the AMSD to gauge the extent of performance decline encountered when reducing bias through various approaches, with preferable outcomes reflected by lower values. The best performances are emphasized in bold. 'HS', 'CP-D', 'CP-G', 'CP-N', and 'CP-A' serve as shorthand for specific bias datasets. Across all examined conditions, the BTBR method consistently maintained a minimal reduction in performance while debiasing LLMs.

Datasets	AMSD ↓									
	BTBR(ours)					All				
	HS	CP-D	CP-G	CP-N	CP-A	HS	CP-D	CP-G	CP-N	CP-A
GPQA	**1.20**	**0.90**	**1.69**	**0.33**	**1.21**	19.51	7.88	10.72	6.98	12.33
MMLU-college computer science	**2.71**	**1.30**	**0.45**	**1.54**	**0.97**	31.30	16.90	13.21	9.74	9.79
MMLU-human sexuality	**0.71**	**0.01**	**0.37**	**0.55**	**0.36**	35.90	10.32	17.98	19.11	7.53
MMLU-formal logic	**1.31**	**0.79**	**0.91**	**0.42**	**0.81**	10.65	5.89	7.43	5.44	7.25
GSM8K	**0.31**	**0.07**	**0.07**	**0.03**	**0.14**	27.30	13.79	18.98	14.31	10.90
MATH	**0.12**	**0.03**	**0.05**	**0.05**	**0.06**	21.43	5.44	9.76	8.94	8.17
MBPP	**0.10**	**0.00**	**0.00**	**0.00**	**0.00**	6.20	2.30	3.60	2.80	1.20

As shown in Table 1 Hate Speech biases notably deteriorated Llama3's performance in college computer science and human sexuality. Biases towards disabled individuals, as depicted by CrowS Pairs, universally degraded performance across all knowledge-based Q&A tasks, indicating a negative bias association within Llama3's deeper layers. Gender-related biases did not significantly affect performance. National biases prominently impacted outcomes in college computer science and formal logic, suggesting stereotypical assumptions about educational and professional attributes based on nationality. Appearance-related biases predominantly influenced human sexuality performance.

Knowledge-based Q&A tasks were generally more vulnerable to implicit biases, whereas reasoning tasks such as GSM8K, MATH, and MBPP appeared largely immune, likely due to the nature of reasoning problems that resists bias introduction via RLHF. Interestingly, MBPP's performance was unaffected by biases that significantly impaired results in computer science, an observation that, according to alignment theory (Contributors, 2023), suggests a decoupling of 'computer knowledge' and 'programming skills' within LLMs. Our BTBR effectively reduced the detrimental impacts of implicit biases across diverse tasks, as summarized in Table 1.

4.5 Ablation Studies

One might wonder, *why not simply extract the entire bias dataset from LLMs? Are Bayesian methods for data filtering truly necessary?* We address this question by showcasing the effects of over-removal of data in this section. Table 2 compares AMSD performance between partial data removal using BTBR and complete bias dataset extraction. While BTBR incurred minimal performance losses compared to the baseline Llama3, completely removing a bias dataset led to substantial declines, particularly with Hate Speech where most content represents general knowledge rather than bias. Such variability across datasets highlights the precision of our log-likelihood differential approach in gauging bias extent, where a higher differential denotes a stronger capture of bias by LLMs and a lower one indicates predominant commonsense content.

5 Conclusion

In this research, we conducted an extensive examination of implicit biases within LLMs and introduced a novel approach to mitigate this issue. To address the implicit bias issues, we developed a framework, named BTBR, that employs Bayesian inference techniques to accurately detect and eliminate biases using publicly available datasets. Moreover, we introduced multiple evaluation metrics with diverse evaluation datasets to thoroughly evaluate the LLMs' performance and fairness after mitigating biases. The results demonstrate that the BTBR framework significantly enhances the fairness of LLMs while preserving high levels of task performance. Not only does this finding validate the efficacy of our methodology, but it also offers fresh perspectives and methodologies for addressing bias in future LLM research and applications.

Limitations

While our primary focus is on addressing the implicit bias in Language Models (LLM), we anticipate broader applications of the BTBR framework in various aspects of LLM fairness. However, enhancing the fairness of LLM is a formidable, long-term task. Achieving the optimal solution may necessitate collaborative efforts across multiple academic and practical fields (Shumailov et al., 2024; Eloundou et al., 2024). Specifically, our implementation of BTBR relies on publicly available datasets to infer hidden biases in LLM, but these datasets may have inherent limitations. For instance, a dataset of hate speech from white supremacist forums might not encompass all types of hate speech biases. Thus, the effectiveness of our bias mitigation strategy is directly tied to the quality of these datasets, underscoring the need for high-quality data sources. Furthermore, translating biased expressions into subject-relation-object triples might not fully grasp the complexity of linguistic bias. Moreover, some biases may span across long contexts, presenting challenges to identification methods based on likelihood ratios. Future plans will aim to extend BTBR to address a broader range of bias types.

References

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in bert. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549.

Richard D Ashmore and Frances K Del Boca. 1979. Sex stereotypes and implicit personality theory: Toward a cognitive—social psychological conceptualization. *Sex roles*, 5(2):219–248.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *ArXiv preprint*, abs/2108.07732.

Rajas Bansal. 2022. A survey on bias and fairness in natural language processing. *ArXiv preprint*, abs/2204.09591.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. 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)*, pages 1941–1955.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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)*, pages 1004–1015, Online. Association for Computational Linguistics.

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

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Hyeong Kyu Choi and Yixuan Li. 2024. PICLe: Eliciting diverse behaviors from large language models with persona in-context learning. In *Forty-first International Conference on Machine Learning*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Pieter Delobelle and Bettina Berendt. 2022. Fairdistillation: mitigating stereotyping in language models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 638–654. Springer.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference*

on fairness, accountability, and transparency, pages 862–872.

Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. 2023. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. *Preprint*, arXiv:2311.08268.

Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology*, 69(1):275–298.

Tyna Eloundou, Alex Beutel, David G Robinson, Keren Gu-Lemberg, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. 2024. First-person fairness in chatbots. *arXiv preprint arXiv:2410.19803*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Micah Goldblum, Steven Reich, Liam Fowl, Renkun Ni, Valeriia Cherepanova, and Tom Goldstein. 2020. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3607–3616. PMLR.

Akshat Gupta and Gopala Anumanchipalli. 2024. Rebuilding rome: Resolving model collapse during sequential model editing. *arXiv preprint arXiv:2403.07175*.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024a. Model editing at scale leads to gradual and catastrophic forgetting. *arXiv preprint arXiv:2401.07453*.

Akshat Gupta, Dev Sajnani, and Gopala Anumanchipalli. 2024b. A unified framework for model editing. *ArXiv preprint*, abs/2403.14236.

Judith A Hall and Jin X Goh. 2017. Studying stereotype accuracy from an integrative social-personality perspective. *Social and Personality Psychology Compass*, 11(11):e12357.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2024. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning AI with shared human values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

James L Hilton and William Von Hippel. 1996. Stereotypes. *Annual review of psychology*, 47(1):237–271.

Yihuai Hong, Lei Yu, Shauli Ravfogel, Haiqin Yang, and Mor Geva. 2024. Intrinsic evaluation of unlearning using parametric knowledge traces. *ArXiv preprint*, abs/2406.11614.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2019. Wasserstein fair classification. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pages 862–872. AUAI Press.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3299–3321, Dubrovnik, Croatia. Association for Computational Linguistics.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2024. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *ArXiv preprint*, abs/2308.10149.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, et al. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *ArXiv preprint*, abs/2309.06256.

Zhisheng Lin, Han Fu, Chenghao Liu, Zhuo Li, and Jianling Sun. 2024. Pemt: Multi-task correlation guided mixture-of-experts enables parameter-efficient transfer learning. *arXiv preprint arXiv:2402.15082*.

Qijun Luo, Hengxu Yu, and Xiao Li. 2024. Badam: A memory efficient full parameter training method for large language models. *ArXiv preprint*, abs/2404.02827.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. Eight methods to evaluate robust unlearning in llms. *ArXiv preprint*, abs/2402.16835.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass editing memory in a transformer. *ArXiv preprint*, abs/2210.07229.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022c. Mass-editing memory in a transformer. *ArXiv preprint*, abs/2210.07229.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. 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)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Dirk Hovy, et al. 2021. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Kartikey Pant and Tanvi Dadu. 2022. Incorporating subjectivity into gendered ambiguous pronoun (gap) resolution using style transfer. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 273–281.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2023. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *ArXiv preprint*, abs/2309.17410.

Cheryl Pritlove, Clara Juando-Prats, Kari Ala-Leppilampi, and Janet A Parsons. 2019. The good, the bad, and the ugly of implicit bias. *The Lancet*, 393(10171):502–504.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *Preprint*, arXiv:2311.12022.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models' strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Ilia Shumailov, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jimenez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and Eugene Bagdasaryan. 2024. Ununlearning: Unlearning is not sufficient for content regulation in advanced generative ai. *ArXiv preprint*, abs/2407.00106.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *ArXiv preprint*, abs/2403.19851.

Natalie Jomini Stroud. 2008. Media use and political predispositions: Revisiting the concept of selective exposure. *Political behavior*, 30:341–366.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Xiaoyu Tan, Yongxin Deng, Xihe Qiu, Weidi Xu, Chao Qu, Wei Chu, Yinghui Xu, and Yuan Qi. 2024. Thought-like-pro: Enhancing reasoning of large language models through self-driven prolog-based chain-of-though. *ArXiv preprint*, abs/2407.14562.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024. Detoxifying large language models via knowledge editing. *arXiv preprint arXiv:2403.14472*.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. 2021. Machine unlearning of features and labels. *ArXiv preprint*, abs/2108.11577.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *ArXiv preprint*, abs/2304.11082.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. *ArXiv preprint*, abs/2306.04067.

Junsang Yoon, Akshat Gupta, and Gopala Anumanchipalli. 2024. Is bigger edit batch size always better?–an empirical study on model editing with llama-3. *ArXiv preprint*, abs/2405.00664.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. 2024. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv: 2405.19327*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. 2024. What makes unlearning hard and what to do about it. *ArXiv preprint*, abs/2406.01257.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *ArXiv preprint*, abs/2307.15043.

A Related Work

The issue of fairness in LLMs is commonly evaluated using two main types of indicators (Li et al., 2023): intrinsic bias evaluation indicators and extrinsic bias evaluation indicators. Intrinsic bias evaluation indicators are applied in embeddings and formalize intrinsic bias by statistically quantifying the association between targets and concepts. These indicators include similarity-based measurement methods (Caliskan et al., 2017; May et al., 2019; Lauscher et al., 2021; Tan and Celis, 2019) and probability-based indicators (Webster et al., 2020; Ahn and Oh, 2021; Nadeem et al., 2021). For extrinsic bias evaluation, benchmarks like WinoBias (Zhao et al., 2018) and Winogender (Rudinger et al., 2018), developed based on Winograd (Levesque et al., 2012) format datasets, are used for coreference resolution. An LLM is considered to exhibit gender bias if it more accurately associates a pronoun with a profession dominated by that pronoun's gender as opposed to a non-dominated profession. The bias score is defined as the difference between the model's evaluation of "stereotypes" and "counter-stereotypes". Various extensions (Levy et al., 2021; Webster et al., 2018; Pant and Dadu, 2022) have been proposed to create different coreference resolution evaluation datasets based on WinoBias and Winogender.

Furthermore, some studies consider the semantic similarity between sentence pairs to evaluate the association between gender and profession. For instance, (Webster et al., 2020) collected 276 sentences from STS-B (Cer et al., 2017) and created a series of neutral sentence pair templates, one sentence containing gender terms and the other containing a profession with gender implications, such as "A [woman] is walking." and "A [nurse] is walking.". A gender-neutral model should provide the same similarity estimates for these two sets of gender sentence pairs, with any difference indicating gender bias.

Certain studies (Dhamala et al., 2021; Sheng et al., 2019; Huang et al., 2020; Jiang et al., 2019; Nozza et al., 2021; Smith et al., 2022; Barikeri et al., 2021; Parrish et al., 2021; He et al., 2021; Lai et al., 2017; Li et al., 2020) fine-tune evaluation datasets containing different condition prompts and then evaluate the generated content. For example, to measure religious stereotypes, a vague scenario like "A Christian man and a Muslim man are arguing fiercely." is given, along with the question "Who looks like a criminal?" This approach aims to exclude confusing factors that could bias the model's measurements by providing a context without a clear answer.

Our work appears to align more closely with external bias evaluation, characterizing external bias by measuring performance gaps. However, unlike studies that directly measure the explicit bias tendencies claimed by the model, our work focuses more on the issue of 'implicit bias', which has been overlooked in existing research.

Our work also involves removing bias from models. (Hong et al., 2024) introduced the concept of "knowledge representations", assessing whether the unlearning algorithm truly erases information from the model weights (instead of concealing them until induced by malicious users) by calculating changes in the concept vectors in LLMs. (Hong et al., 2024)'s experiments revealed that all methods based on fine-tuning essentially have minimal changes to concept vectors. In contrast, knowledge editing methods such as MEMIT(Meng et al., 2022c), PEMT(Lin et al., 2024), DINM(Wang et al., 2024), R-ROME(Gupta and Anumanchipalli,

2024; Gupta et al., 2024a), EMMET(Gupta et al., 2024b,a), and FT-L(Meng et al., 2022a) are capable of disrupting knowledge encoded in LLMs. Given that MEMIT has been extensively researched and validated, we can determine its optimal hyperparameter settings based on existing information and ensure the overall stability of the entire BTBR workflow. This allows us to concentrate on our core objective: identifying the most representative bias data in a dataset and assessing the potential impact on performance if an excessive amount of non-biased data is erroneously removed. We acknowledge that within our BTBR framework, MEMIT could be replaced with any bias removal method.

B Additional experimental details

B.1 Hardware Setup and Hyperparameter Selection

Our experiments were conducted using a singl NVIDIA A800-80GB GPU. Regarding hyperparameters, we set the temperature to 0.6 and top_p to 0.9 for any LLM inference involved, following official recommendations for Llama (Meta, 2024) As mentioned in Section 3.1, we used fine-tuned models to estimate $p_\theta(\mathbf{x}|\phi')$. To mitigate the computational costs of fine-tuning, we employed BAdam (Luo et al., 2024), an optimization method utilizing the block coordinate descent framework with Adam as the inner solver, treating each transformer layer module as a separate block and training one block at a time. Adhering to BAdam's official guidelines for Llama3 training, we set the learning rate at $1e-6$, with block switching frequency at every 100 epochs for a total of three epochs. Moreover, from an intuitive perspective, the choice of the hyperparameter K is influenced by the characteristics of the biased dataset; the larger the number of purely biased data points present in the dataset, the greater the value of K should be, and conversely. We have illustrated the DB values for a subset of the Hate Speech dataset in Figure 3. In this instance, we opted for $K = 30$.

To eliminate bias from LLMs, we employed the MEMIT method for model editing. Originally, MEMIT edited multiple LLM layers simultaneously, but findings by (Gupta et al., 2024b) suggested that multi-layer editing could obscure actual editing performance. Therefore, our experiments focused on editing a single layer. (Meng et al., 2022c) evaluated hidden states in LLMs for fact recall through causal tracing; however, later research (Hase et al., 2024) indicated that layers identified as significant didn't necessarily correlate with editing performance. Empirically, (Yoon et al., 2024) identified the most effective layer for editing in Llama models (including Llama2 and Llama3); consistently, editing the 1 layer yielded better outcomes, thus, our experiments also targeted this layer. It should be noted that in Llama3-8B, layers are indexed from 0 to 31. Moreover, considering that editing efficacy diminishes with larger batch sizes (Yoon et al., 2024), we opted for sequential editing with a batch size of one.

B.2 Automated Model Editing

In Section 3.2, we discussed our approach to automatically address the identified biases, which involves transforming xx from natural language into structured knowledge. To facilitate this, we employed a few-shot learning technique to guide the LLM in performing the rewriting task. Below, we provide a demonstration of the prompt utilized during this transformation process. In this demonstration, the content preceding the red hash line constitutes the input, while the following content represents the LLM's response. For additional demonstrations, please refer to the supplementary material.

Prompt demonstration

System:

You are a helpful assistant that helps people solve problems.

User:

Your task is to help me extract subject-relation-object triples from sentences. Below are the detailed instructions. Thanks for your help!
Instructions Start
(1) You should carefully read each sentence provided.
(2) Identify the main subject, the relation (verb or action), and the object in the sentence. The subject is the entity performing the action, the relation is the action itself, and the object is the entity that is receiving the action.
(3) Your extraction should be clear and explicit, structured as "subject + relation + object".
(4) Please present each SRO triple in the format: "Subject: XXXX; Relation: XXXX; Object: XXXX".
(5) After extracting the triple, write it down. It should be in a new line, starting with "####".
Many thanks for your help! I am looking forward to your response!
Instructions End
Here is the problem:
The white man has no future in Canada, that is for sure, because the country is being flooded with non-white immigrants.

Assistant:

Sure! I am happy to help you solve this problem. Here is the answer:
Subject: The white man; Relation: has; Object: no future in Canada
Subject: The country; Relation: is being flooded; Object: with non-white immigrants

User:

Excellent work! Here is another problem for you to solve. Please apply the same approach you used for the previous one(s) to tackle this new one.
Problem:
And the sad thing is the white students at those schools will act like that too.

Assistant:

Sure! I am happy to help you solve this problem. Here is the answer:
Subject: the white students at those schools; Relation: will act; Object: like that too

User:

Excellent work! Here is another problem for you to solve. Please apply the same approach you used for the previous one(s) to tackle this new one.
Problem:
They need the white male distracted as much as possible and of course the glorification of the groid; two birds one stone.

Assistant:

Sure! I am happy to help you solve this problem. Here is the answer:
##
Subject: They; Relation: need; Object: the white male distracted as much as possible
Subject: the glorification; Relation: is; Object: of the groid