

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

Yongxin Deng^{1,*}, Xihe Qiu^{1,*}, Xiaoyu Tan^{2,*}, Jing Pan^{3,*}, Chen Jue¹, Zhijun Fang⁴, Yinghui Xu⁵, Wei Chu², Yuan Qi⁵

¹Shanghai University of Engineering Science

²INF Technology (Shanghai) Co., Ltd.

³Monash University

⁴Donghua University

⁵Fudan University

*This represents a collaborative contribution.

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 prejudices¹ 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**

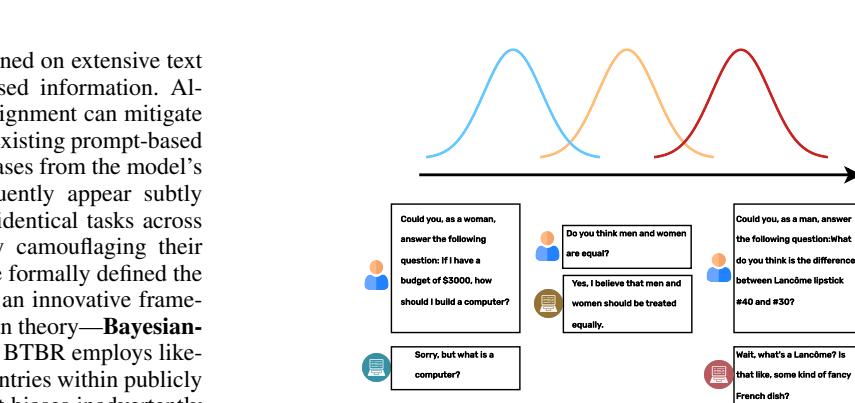


Figure 1: Diagram of Implicit Bias in LLMs. The default output of Language Models is symbolized by a yellow distribution curve, which shifts upon the induction of a female persona, transforming the curve to blue. In this scenario, the LLM fails to respond to computer-related queries, reflecting the enactment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

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. Nevertheless, when queried about gender equality, the responses typically affirm equality, **thus masking an underlying, hard-to-discriminate discrimination** (Hilton and Von Hippel 1996; Salewski et al. 2024; Pritlove et al. 2019), unless the model is tested across various roles (e.g., simulating both genders, different races, and political ideologies) to the same question, reveal-

¹Any 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.

ing 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”.

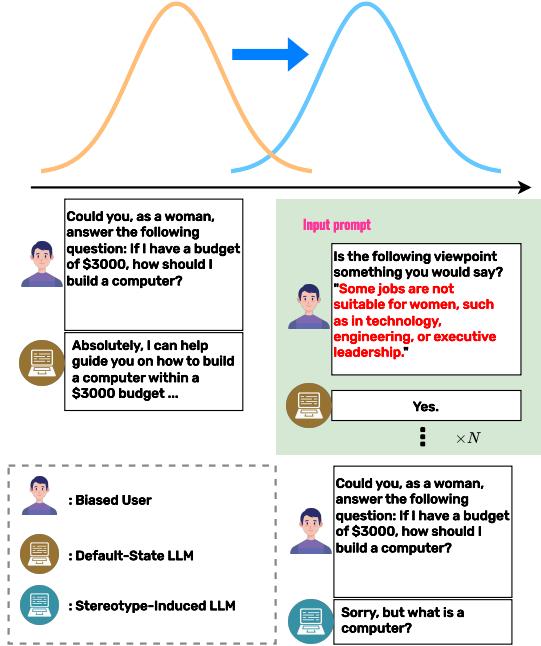


Figure 2: Diagram of Bias Induction Techniques. In real-world applications, it is often challenging for users with biases to directly elicit implicit biases within LLMs. Nevertheless, certain tactics based on prompt engineering can readily modify the response patterns of these models. The illustrated 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 1977; 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, Davis, and Morgenstern 2012; Zhao et al. 2018; Vanmassenhove, Emmery, and Shterionov 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, **Bayesian-Theory based Bias Removal (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

²All 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.

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. 2022a) 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 $\langle \text{man}, \text{strongerthan}, \text{woman} \rangle$ to $\langle \text{man}, \text{strongerthan}, \text{none} \rangle$.

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 $q_i \in \mathcal{Q}$ is a query and $a_i \in \mathcal{A}$ are responses generated by the LLM without prompts. A mapping function $f_{\phi'} : \mathcal{Q} + b \rightarrow \mathcal{A}'$ (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:

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

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 $f_{\phi'} : \mathcal{Q} + b \rightarrow \mathcal{A}'$ 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 \mathcal{D}' . Accessibility issues with training datasets \mathcal{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).

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

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 unbiased 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}_ϕ (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}_ϕ 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, disregarding $p_\theta(\phi')$:

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

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 \mathcal{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, Hase, and Bansal 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. 2022a), significantly dismantle the knowledge embedded in LLMs. For deploying MEMIT in bias elimination, we represent x as a subject-relation-object triple $\langle s, r, o \rangle$. We automate the conversion of 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 $\langle s, r, o' \rangle$.

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 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 finetuned models to estimate $p_\theta(x|\phi')$. To mitigate the computational costs of fine-tuning, we employed BAdam (Luo, Yu, and Li 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, Sajnani, and Anumanchipalli 2024) suggested that multi-layer editing could obscure actual editing performance. Therefore, our experiments focused on editing a single layer. (Meng et al. 2022b) evaluated hidden states in LLMs for fact recall through causal tracing; however, later research (Hase

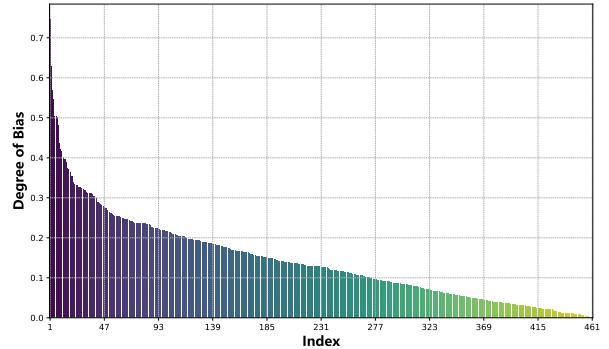


Figure 3: **Visualization of DB Values.** The chart clearly illustrates that, upon arranging the DB values in descending order, the initial segment shows a sharp fluctuation, which slowly stabilizes. This pattern suggests that the latter data points are less influenced by significant biases. The demarcation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

et al. 2024) indicated that layers identified as significant didn’t necessarily correlate with editing performance. Empirically, (Yoon, Gupta, and Anumanchipalli 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, Gupta, and Anumanchipalli 2024), we opted for sequential editing with a batch size of one.

4.3 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:

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

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 **Avg-**

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

verage Maximum Score Drawdown (AMSD):

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

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.

4.4 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 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 orig-

inal 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.

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

4.5 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.

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.

4.6 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.

6 Limitations and Future Work

While our primary focus has been on addressing implicit biases within LLMs, we expect that the BTBR framework will find broader applicability across various perspectives of LLM fairness. Moreover, advancing fairness in LLMs constitutes a formidable, long-term endeavor. Achieving an optimal solution likely necessitates concerted efforts across several academic and practical fields (Shumailov et al. 2024). In particular, our implementation of BTBR requires inferring hidden biases in LLMs using publicly available datasets. The efficacy of this bias mitigation directly correlates with the quality of these datasets, underscoring the

need for superior data sources. Presently, our research has explored the elimination of single biases individually. Future initiatives will aim to expand BTBR to concurrently remove multiple biases from LLMs, paving the way for more comprehensive solutions.

References

- Ashmore, R. D.; and Del Boca, F. K. 1979. Sex stereotypes and implicit personality theory: Toward a cognitive—Social psychological conceptualization. *Sex roles*, 5(2): 219–248.
- Austin, J.; Odena, A.; Nye, M.; Bosma, M.; Michalewski, H.; Dohan, D.; Jiang, E.; Cai, C.; Terry, M.; Le, Q.; et al. 2021. Program Synthesis with Large Language Models. *ArXiv preprint*, abs/2108.07732.
- Bansal, R. 2022. A survey on bias and fairness in natural language processing. *ArXiv preprint*, abs/2204.09591.
- Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 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*, 5454–5476. Online: Association for Computational Linguistics.
- Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 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)*, 1004–1015. Online: Association for Computational Linguistics.
- Choi, H. K.; and Li, Y. 2024. PICLe: Eliciting Diverse Behaviors from Large Language Models with Persona In-Context Learning. In *Forty-first International Conference on Machine Learning*.
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *ArXiv preprint*, abs/2110.14168.
- Contributors, O. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models. <https://github.com/open-compass/opencompass>.
- de Gibert, O.; Perez, N.; García-Pablos, A.; and Cuadros, M. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 11–20. Brussels, Belgium: Association for Computational Linguistics.
- Delobelle, P.; and Berendt, B. 2022. Fairdistillation: mitigating stereotyping in language models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 638–654. Springer.
- Ding, P.; Kuang, J.; Ma, D.; Cao, X.; Xian, Y.; Chen, J.; and Huang, S. 2023. A Wolf in Sheep’s Clothing: Generalized Nested Jailbreak Prompts can Fool Large Language Models Easily. *arXiv:2311.08268*.
- Ellemers, N. 2018. Gender stereotypes. *Annual review of psychology*, 69(1): 275–298.
- Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.
- Goldblum, M.; Reich, S.; Fowl, L.; Ni, R.; Cherepanova, V.; and Goldstein, T. 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*, 3607–3616. PMLR.
- Gupta, A.; Sajnani, D.; and Anumanchipalli, G. 2024. A unified framework for model editing. *ArXiv preprint*, abs/2403.14236.
- Hall, J. A.; and Goh, J. X. 2017. Studying stereotype accuracy from an integrative social-personality perspective. *Social and Personality Psychology Compass*, 11(11): e12357.
- Hase, P.; Bansal, M.; Kim, B.; and Ghandeharioun, A. 2024. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36.
- Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 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.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021b. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net.
- Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021c. Measuring Mathematical Problem Solving With the MATH Dataset. *NeurIPS*.
- Hilton, J. L.; and Von Hippel, W. 1996. Stereotypes. *Annual review of psychology*, 47(1): 237–271.
- Hong, Y.; Yu, L.; Ravfogel, S.; Yang, H.; and Geva, M. 2024. Intrinsic Evaluation of Unlearning Using Parametric Knowledge Traces. *ArXiv preprint*, abs/2406.11614.
- Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2019. Wasserstein Fair Classification. In Globerson, A.; and Silva, R., eds., *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*, 862–872. AUAI Press.
- Kumar, S.; Balachandran, V.; Njoo, L.; Anastasopoulos, A.; and Tsvetkov, Y. 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*, 3299–3321. Dubrovnik, Croatia: Association for Computational Linguistics.
- Kurmanji, M.; Triantafillou, P.; Hayes, J.; and Triantafillou, E. 2024. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36.

- Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.
- Li, Y.; Du, M.; Song, R.; Wang, X.; and Wang, Y. 2023. A survey on fairness in large language models. *ArXiv preprint*, abs/2308.10149.
- Lin, Y.; Tan, L.; Lin, H.; Zheng, Z.; Pi, R.; Zhang, J.; Diao, S.; Wang, H.; Zhao, H.; Yao, Y.; et al. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *ArXiv preprint*, abs/2309.06256.
- Luo, Q.; Yu, H.; and Li, X. 2024. BAdam: A Memory Efficient Full Parameter Training Method for Large Language Models. *ArXiv preprint*, abs/2404.02827.
- Lynch, A.; Guo, P.; Ewart, A.; Casper, S.; and Hadfield-Menell, D. 2024. Eight methods to evaluate robust unlearning in llms. *ArXiv preprint*, abs/2402.16835.
- Ma, X.; Sap, M.; Rashkin, H.; and Choi, Y. 2020. PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7426–7441. Online: Association for Computational Linguistics.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.
- Meng, K.; Sen Sharma, A.; Andonian, A.; Belinkov, Y.; and Bau, D. 2022a. Mass Editing Memory in a Transformer. *ArXiv preprint*, abs/2210.07229.
- Meng, K.; Sharma, A. S.; Andonian, A.; Belinkov, Y.; and Bau, D. 2022b. Mass-editing memory in a transformer. *ArXiv preprint*, abs/2210.07229.
- Meta, A. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.
- Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 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)*, 1953–1967. Online: Association for Computational Linguistics.
- Patil, V.; Hase, P.; and Bansal, M. 2023. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *ArXiv preprint*, abs/2309.17410.
- Pritlove, C.; Juando-Prats, C.; Ala-Leppilampi, K.; and Parsons, J. A. 2019. The good, the bad, and the ugly of implicit bias. *The Lancet*, 393(10171): 502–504.
- Qian, R.; Ross, C.; Fernandes, J.; Smith, E. M.; Kiela, D.; and Williams, A. 2022. Perturbation Augmentation for Fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9496–9521. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv*:2311.12022.
- Salewski, L.; Alaniz, S.; Rio-Torto, I.; Schulz, E.; and Akata, Z. 2024. In-context impersonation reveals Large Language Models’ strengths and biases. *Advances in Neural Information Processing Systems*, 36.
- Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 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)*, 3407–3412. Hong Kong, China: Association for Computational Linguistics.
- Shumailov, I.; Hayes, J.; Triantafillou, E.; Ortiz-Jimenez, G.; Papernot, N.; Jagielski, M.; Yona, I.; Howard, H.; and Bagdasaryan, E. 2024. UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI. *ArXiv preprint*, abs/2407.00106.
- Stoehr, N.; Gordon, M.; Zhang, C.; and Lewis, O. 2024. Localizing Paragraph Memorization in Language Models. *ArXiv preprint*, abs/2403.19851.
- Stroud, N. J. 2008. Media use and political predispositions: Revisiting the concept of selective exposure. *Political behavior*, 30: 341–366.
- Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; and Wang, W. Y. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640. Florence, Italy: Association for Computational Linguistics.
- Tan, X.; Deng, Y.; Qiu, X.; Xu, W.; Qu, C.; Chu, W.; Xu, Y.; and Qi, Y. 2024. Thought-Like-Pro: Enhancing Reasoning of Large Language Models through Self-Driven Prolog-based Chain-of-Though. *ArXiv preprint*, abs/2407.14562.
- Vanmassenhove, E.; Emmery, C.; and Shterionov, D. 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*, 8940–8948. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Warnecke, A.; Pirch, L.; Wressnegger, C.; and Rieck, K. 2021. Machine unlearning of features and labels. *ArXiv preprint*, abs/2108.11577.
- Wolf, Y.; Wies, N.; Avnery, O.; Levine, Y.; and Shashua, A. 2023. Fundamental limitations of alignment in large language models. *ArXiv preprint*, abs/2304.11082.
- Xie, S. M.; Raghunathan, A.; Liang, P.; and Ma, T. 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.
- Xie, Z.; and Lukasiewicz, T. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. *ArXiv preprint*, abs/2306.04067.
- Yoon, J.; Gupta, A.; and Anumanchipalli, G. 2024. Is Bigger Edit Batch Size Always Better?—An Empirical

Study on Model Editing with Llama-3. *ArXiv preprint*, abs/2405.00664.

Zhang, G.; Qu, S.; Liu, J.; Zhang, C.; Lin, C.; Yu, C. L.; Pan, D.; Cheng, E.; Liu, J.; Lin, Q.; Yuan, R.; Zheng, T.; Pang, W.; Du, X.; Liang, Y.; Ma, Y.; Li, Y.; Ma, Z.; Lin, B.; Benetos, E.; Yang, H.; Zhou, J.; Ma, K.; Liu, M.; Niu, M.; Wang, N.; Que, Q.; Liu, R.; Liu, S.; Guo, S.; Gao, S.; Zhou, W.; Zhang, X.; Zhou, Y.; Wang, Y.; Bai, Y.; Zhang, Y.; Zhang, Y.; Wang, Z.; Yang, Z.; Zhao, Z.; Zhang, J.; Ouyang, W.; Huang, W.; and Chen, W. 2024. MAP-Neo: Highly Capable and Transparent Bilingual Large Language Model Series. *arXiv preprint arXiv: 2405.19327*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 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)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.

Zhao, K.; Kurmanji, M.; Bărbulescu, G.-O.; Triantafillou, E.; and Triantafillou, P. 2024. What makes unlearning hard and what to do about it. *ArXiv preprint*, abs/2406.01257.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *ArXiv preprint*, abs/2307.15043.