Debiasing Large Language Models with Structured Knowledge

Congda Ma[1] **Tianyu Zhao**[2] **Manabu Okumura**[1]
[1]Tokyo Institute of Technology [2]Sakana AI
{ma, oku}@lr.pi.titech.ac.jp tianyu@sakana.ai

Abstract

Due to biases inherently present in data for pre-training, current pre-trained Large Language Models (LLMs) also ubiquitously manifest the same phenomena. Since the bias influences the output from the LLMs across various tasks, the widespread deployment of the LLMs is hampered. We propose a simple method that utilizes structured knowledge to alleviate this issue, aiming to reduce the bias embedded within the LLMs and ensuring they have an encompassing perspective when used in applications. Experimental results indicated that our method has good debiasing ability when applied to existing both autoregressive and masked language models. Additionally, it could ensure that the performances of LLMs on downstream tasks remain uncompromised. Our method outperforms state-of-the-art (SOTA) baselines in the debiasing ability. Importantly, our method obviates the need for training from scratch, thus offering enhanced scalability and cost-effectiveness.[1]

1 Introduction

There have recently been rapid developments in natural language processing (NLP) with the emergence of pre-trained Large Language Models (LLMs). Fine-tuning these models can significantly improve their performance in downstream tasks because, during the fine-tuning process, the knowledge acquired during pre-training on large corpora can be awakened and effectively applied to the downstream tasks. However, biases present in LLMs (e.g., gender and occupation biases) can be a serious problem because of being propagated to the downstream tasks. Therefore, their analysis and mitigation in LLMs have become a critical issue.

Noun in context	**Neutral**	**Positive**	**Negative**
"laborer"	97	1	1
"CEO"	52	44	2

Table 1: Regard scores (Sheng et al., 2019) of generated texts from contexts, containing "laborer" or "CEO". We used GPT2-large to generate the texts. The detail of the score is described in Sec. 4.5.

Analysis of the text generated by LLMs has shown that the bias exists at the word-level (Sheng et al., 2019; Nozza et al., 2021). The first phenomenon entails that when given a context, LLMs prefer generating words of a certain category with a higher probability. For example, when the input context is "The CEO believes that", the distribution of output probabilities shows a higher likelihood of the next word being "he" rather than "she". However, if the word "CEO" in the context is replaced with "nurse", the bias leads to a higher probability of generating "she" over "he" (Hewitt et al., 2023). The second phenomenon is that when generating text related to a particular noun, LLMs tend to prefer generating content with a specific attribute. For example, as observed in our experiments (see Table 1), when LLMs generate text related to "laborer" and "CEO" separately, the text related to "laborer" exhibits a neutral sentiment, whereas the text related to "CEO" possesses a higher positive sentiment.

It is commonly believed that these biases are caused by the biases inherent in the pre-training data (Brunet et al., 2019; Dev et al., 2020; Papakyriakopoulos et al., 2020). We assume the heightened association between "CEO" and the positive sentiment is attributed to the co-occurrence of "CEO" with many positive connotations. Consequently, LLMs incorporate positive features into the representation of "CEO". During the inference process, LLMs tend to reproduce and magnify the bias inherent in the pre-training data (Kurita et al., 2019; Sheng et al., 2019), which causes a rise in the prob-

[1] Our code is at https://github.com/KGDebias/KGDebias.

Findings of the Association for Computational Linguistics ACL 2024, pages 10274–10287
August 11-16, 2024 ©2024 Association for Computational Linguistics

ability of generating positive content, resulting in the production of biased text.

Previous research has sought to alleviate the bias by controlling the pre-training data (Touvron et al., 2023a) or by adjusting embeddings (Hewitt et al., 2023). However, these approaches require models to be pre-trained from scratch, which significantly increases the training cost. To address this issue, we propose a simple method characterized by diminished computation cost. It mitigates the inherent bias in LLMs by incorporating structured knowledge during a second phase of pre-training. The structured knowledge contains hypernyms for a word as one of the relationships between words. Since the hypernyms tend to cover broader or more general concepts, even when a word causes a bias for surrounding words, the second pre-training by incorporating the information of its hypernyms can enhance its representation to take into account the hypernyms, thus decreasing the bias caused by the word. For example, from the knowledge piece "CEO is-a employee", we can construct a sentence "a CEO is an employee.", which signifies that a "CEO" belongs to the superordinate concept of "employee". By training with these sentences for the hypernym information, we can incorporate the representation for "employee" into that for "CEO", that makes LLMs generate content more related to "employee" and reduce the focus on positive content related to "CEO", achieving the goal of debiasing. Since our method is implemented only with a second phase of pre-training and needs no pre-training from scratch, it reduces the cost of training compared to the previous methods.

Our experiments demonstrated the efficacy of our method in debiasing across various LLMs while preserving their generalization ability. Specifically, our results for autoregressive and masked language models indicated that our method can lower bias. Concurrently, models with our method did not exhibit significant performance degradation in downstream tasks, affirming the preservation of their generalization ability. Furthermore, comparing to strong debiasing baselines, our method yielded superior scores and exhibited enhanced bias control.

2 Related Work

2.1 Bias in LLMs

While the majority of tasks have been accomplished using the LLMs (Devlin et al., 2019; Radford and Wu, 2019; Touvron et al., 2023a,b) with the advancement of LLMs, an increasing number of researchers have discovered bias embedded within them. Several previous studies have demonstrated the presence of the bias in word embeddings (Bolukbasi et al., 2016; Brunet et al., 2019; Papakyriakopoulos et al., 2020). When applying the LLMs to downstream tasks, such as automatic summarization or web search, the bias can lead to significant harm. Jentzsch and Turan (2022) and Kirk et al. (2021) identified obverse gender bias in BERT and GPT2, where their prediction often reinforces gender-based stereotypes, e.g., doctors are assumed to be male and nurses to be female. Standard benchmarks utilized to evaluate the bias consist of various kinds of stereotypical and anti-stereotypical sentence pairs. CrowSPairs (Nangia et al., 2020) is a commonly used dataset to measure whether a model generally prefers stereotypical sentences.

2.2 Debiasing Methods

Researchers have proposed various methods to mitigate the bias in NLP models. Park et al. (2018); Zhao et al. (2019); Garg et al. (2019); Touvron et al. (2023a) mitigated the bias by changing the pre-training data or the underlying word embeddings, and then by retraining the model. Bordia and Bowman (2019) mitigated the bias in a word level and also required the retraining. The Counterfactual Data Augmentation (CDA) method, proposed by Zimmermann and Hoffmann (2022), and the Data Interventions (DI) method, proposed by Thakur et al. (2023), both used designed examples to mitigate bias in the models. Liang et al. (2020) proposed the SENT-DEBIAS method, which captures the bias subspace of sentence representations by using templates. However, all these methods still suffer from high training cost or unsatisfactory bias control.

3 Proposed Method

As shown in Figure 1, we propose the utilization of structured knowledge to mitigate the bias within LLMs. Our method consists of two main steps: first acquiring textual data for second pre-training from the structured knowledge, and then pre-training the pre-trained LLM using the acquired data, before fine-tuning it for specific downstream tasks.

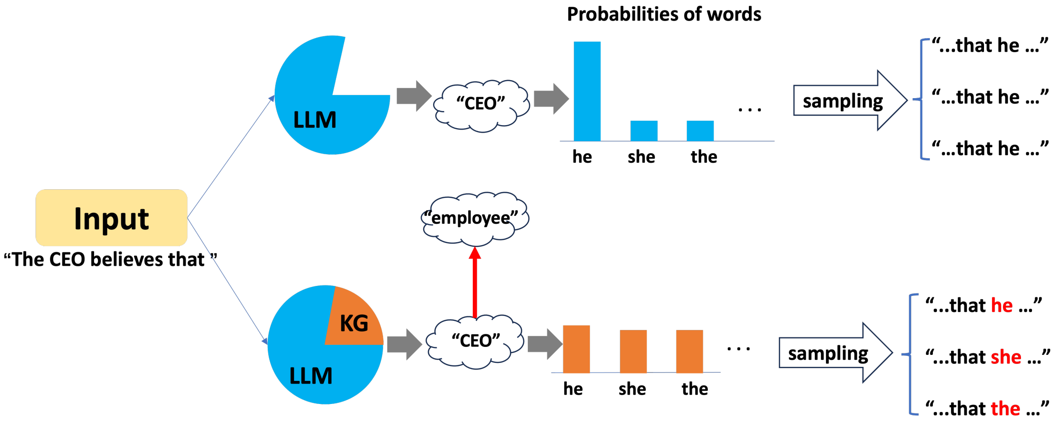


Figure 1: Proposed model framework. **LLM** represents a pre-trained large language model. **KG** represents the structured knowledge.

3.1 Acquiring Training Textual Data from Structured Knowledge

Previous research (Brunet et al., 2019; Dev et al., 2020; Shaikh et al., 2023; Touvron et al., 2023a,b) has suggested that the types of bias in LLMs are linked to human-related attributes such as gender, occupation, religion, and so on. Hence, in our method, we utilize structured knowledge related to humans for debiasing.

We first collect human-related nouns, such as "CEO" and "artist". Then, we obtain a set of structured knowledge pieces for humans, where a structured knowledge piece is between two nouns within the 'is-a' relation, by utilizing the collected nouns; for example, "human-related noun is-a X" and "Y is-a human-related noun", where X and Y are a hypernym and hyponym of the noun, respectively. Since we cannot incorporate the acquired structured knowledge directly into LLMs, to use the structured knowledge in LLMs, by adhering to human grammatical convention, we transform it into sentences, such that "a human-related noun is a X". This sentence construction method has been commonly used in previous work (Bosselut et al., 2019; Guan et al., 2020).

3.2 Training LLMs on Acquired Textual Data

Next, we pre-train the pre-trained LLMs on these acquired textual data to incorporate the structured knowledge into the LLMs. We have contemplated the following two aspects:

1. Higher flexibility: We aspire for a knowledge incorporation method to be plug-and-play. That is, even if a new knowledge base is available in the future, the training would necessitate only adjustment in light of the new knowledge base.

2. Lower training cost: We aim that the training should not start from scratch, thereby diminishing energy consumption throughout the training process, with the broader goal of reducing carbon emission.

Here, in order to make the training process consistent with the features of the LLMs themselves, for different LLMs, we adopt whichever training method was utilized in their pre-training to incorporate the structured knowledge into the LLMs. The details of the objectives are listed in Appendix A.

4 Experiments

4.1 Knowledge Databases

To obtain human-related nouns, we used WordNet (Miller, 1995), a large lexical database in English. Within the version v3.1 of WordNet, there exists a category "noun.person", which contains various human-related nouns. After the processing, we procured 6,904 nouns related to humans, for example, forager and runner. In Appendix B, we show the details of the processing steps.

Hyponym	Relation	Hypernym
assistant	IsA	worker
janitor	IsA	employee
employee	IsA	worker
...	IsA	...

Table 2: Examples of obtained human-related structured knowledge pieces.

Subsequently, we utilized ConceptNet (Speer et al., 2017) to obtain human-related structured knowledge pieces. Table 2 showcases examples of the knowledge pieces derived from ConceptNet.

Finally, we converted the obtained structured knowledge pieces into sentences. The number of the obtained sentences for second pre-training is 33,224 in total.

4.2 LLMs

4.2.1 Autoregressive Language Models

In the experiments for autoregressive language models, we utilized GPT2, GPT-Neo (Black et al., 2021), and LLaMAs (Touvron et al., 2023a,b) as base LLMs. **GPT2** is a general model that can be applied to a variety of tasks. **GPT-Neo** is an instantiation of models similar to GPT2, applying Mesh TensorFlow to facilitate distributed processing support. **LLaMA2** is one of the pre-trained LLMs trained only on publicly available datasets. It shows a competitive performance with existing state-of-the-art (SOTA) LLMs.

Models	**No. of parameters**
Autoregressive language models	
GPT2	774M
GPT-Neo	1.3B
LLaMA2	7B
Masked language models	
BERT	340M
RoBERTa	355M

Table 3: The number of parameters of the pre-trained language models utilized in the experiments.

The number of parameters of the models is shown in Table 3. The learning rate for GPT2 was set to 2e-5, the optimizer was Adam (Kingma and Ba, 2015). The learning rate for GPT-Neo was set to 2e-5, the optimizer was Adam. The experiments with LLaMA2 used deepspeed.[2] The learning rate was set to 2e-4, the optimizer was Adam. The models with GPT2 and GPT-Neo were trained on an A6000 server, and the models with LLaMAs were trained on an A100 server.

We also show the reported scores for three large autoregressive language models to compare with: **LLaMA 65B** (Touvron et al., 2023a), **OPT 175B** (Zhang et al., 2022), an open-sourced language model with 175 billion parameters, and **GPT3** (Brown et al., 2020), a non-public language model with 175 billion parameters.

4.2.2 Masked Language Models

In the experiments for masked language models, we utilized BERT and RoBERTa (Liu et al., 2019) as base LLMs. **BERT** is a pre-trained language model that uses a deep bidirectional transformer architecture, that can capture contextual information from both left and right contexts in all layers. **RoBERTa** is a pre-trained language model that was improved upon BERT through several modifications to the training procedure.

The number of parameters of the models are listed in Table 3. The learning rate for BERT was set to 4e-5, the optimizer was Adam. The learning rate for RoBERTa was 4e-5, the optimizer was Adam. All experiments were done on an A6000 server.

4.3 Baselines

We selected several methods for mitigating bias in the models, including the SOTA model, as baselines for comparison.

SENT-DEBIAS: A sentence embedding debiasing approach proposed by Liang et al. (2020).

DI (Thakur et al., 2023): A data-based approach that utilizes few-shot data to mitigate gender bias. They claimed it is a SOTA model.

Backpack (Hewitt et al., 2023): A baseline applying multiple non-contextual sense vectors and representing a word with the sense vectors for debiasing. It is the current SOTA model.

Following the prior work (Hewitt et al., 2023; Thakur et al., 2023), we used a small version of GPT2 as the base model for training and comparing against Backpack, and used a base version of BERT as the base model for training and comparing with the remaining baselines, for fair comparison. To augment the analysis, we also introduced GPT-Neo 1.3B and the large version of RoBERTa for training and comparison purposes.

Further, we also constructed two strong baselines for comparison. One is the model pre-trained in the same way as in our model on the dataset from Wikitext-2 (Merity et al., 2017) of the same size as our dataset constructed from structured knowledge. Since Wikitext-2 consists of formally written Wikipedia articles, it has been said to contain less explicit bias (Thakur et al., 2023), while it does not consist of structured knowledge, thus making it a suitably strong baseline. We call this baseline

[2] https://www.deepspeed.ai/

Models	**Race/ Color**	**Gender**	**Socioeconomic Status**	**Nationality**	**Religion**	**Age**	**Sexual Orientation**	**Physical Appearance**	**Disability**	**Avg.**
Autoregressive Language Models										
GPT2	9.5	7.6	16.9	2.8	27.1	5.2	**19.0**	10.3	16.7	10.5
Wikitext-tuning	7.0	**6.9**	18.6	15.4	22.3	23.6	22.7	24.6	28.3	13.7
Gen-tuning	8.3	9.9	18.0	1.6	18.6	4.0	20.2	13.5	**11.7**	10.1
Synonym-KG	5.8	9.8	14.0	2.2	**17.6**	9.8	22.6	18.3	18.3	9.4
Ours	**5.4**	9.5	**12.8**	**0.3**	18.6	**2.9**	20.2	**10.3**	15.0	**8.6***
GPT-Neo	12.2	9.5	17.4	4.7	18.6	9.8	19.0	11.9	21.7	12.6
Wikitext-tuning	8.5	6.9	18.6	4.1	17.6	8.6	17.9	**10.3**	20.0	10.6
Gen-tuning	5.8	11.1	19.7	1.6	19.5	13.2	**16.7**	11.9	15.0	10.5
Synonym-KG	6.2	3.1	**12.8**	6.6	15.7	7.5	21.4	18.3	25.0	9.3
Ours	**2.7**	**2.3**	14.5	**0.9**	**12.9**	**7.5**	22.6	11.9	**15.0**	**6.6***
LLaMA2	17.6	10.3	18.0	**8.5**	26.2	21.3	26.2	26.2	**26.7**	17.4
Wikitext-tuning	15.7	9.2	18.6	11.0	24.3	**16.7**	23.8	27.8	30.0	16.4
Synonym-KG	15.5	9.2	**14.5**	14.2	27.1	20.1	**21.4**	**23.0**	30.0	16.5
Ours	**7.0**	**6.9**	18.6	15.4	**22.3**	23.5	22.6	24.6	28.3	**13.6**
Large Language Models										
†LLaMA 65B	7.0	20.6	21.5	14.2	29.0	20.1	31.0	27.8	16.7	16.6
†OPT 175B	18.6	15.7	26.2	12.9	18.6	17.8	28.6	26.2	26.7	19.5
†GPT3	14.7	12.6	23.8	11.6	23.3	14.4	26.2	24.6	26.7	17.2
Masked Language Models										
BERT	7.6	8.8	12.2	1.6	26.2	14.4	23.8	19.8	**8.3**	10.5
Wikitext-tuning	7.2	9.5	23.3	12.9	25.2	20.1	17.9	15.1	15.0	7.0
Gen-tuning	4.3	**3.4**	11.6	2.8	27.1	7.5	**16.7**	**13.5**	11.7	8.0
Synonym-KG	1.0	5.7	12.2	7.9	23.3	8.6	26.2	18.3	16.7	6.6
Ours	**0.8**	4.6	**11.0**	**1.6**	**13.8**	**4.0**	30.9	18.3	25.0	**6.3**
RoBERTa	19.2	9.5	22.1	6.0	24.3	16.7	17.9	24.6	18.3	16.8
Wikitext-tuning	16.9	10.3	**17.4**	7.2	20.5	10.9	20.2	23.0	18.3	15.2
Gen-tuning	18.8	6.5	24.4	2.8	**17.6**	16.7	17.9	19.8	10.0	14.5
Synonym-KG	12.6	3.8	19.8	**0.3**	26.2	15.5	17.9	15.1	25.0	12.6
Ours	**4.1**	**1.9**	20.4	17.3	31.9	**0.6**	**5.9**	**11.9**	**1.7**	**6.5***

Table 4: Scores for different models on the CrowSPairs dataset. A lower score represents less bias. The scores denoted by † are reported in Touvron et al. (2023a). The scores marked with * mean our models outperform the original models significantly with t-test ($p < 0.05$). The original CrowSPairs scores are in Appendix D.

Wikitext-tuning.

The second model was pre-trained using sentences generated from LLaMA2. We employed a prompt to make LLaMA2 generate sentences that describe the relationship between the extracted nouns and their hypernyms. In contrast to Wikitext-2, these newly generated sentences were not utilized during the pre-training, containing the information of the extracted nouns, but might contain bias. This baseline is termed **Gen-tuning**. The specifics of the training data for Gen-tuning are presented in Appendix C.[3]

Additionally, we introduced a variant of our method. In this variant, instead of using the is-a relation, we used the synonym relation to obtain structured knowledge for the extracted nouns. Then, we transformed it into sentences using a template, such that "a human-related noun *is similar to* a X", where X is a synonym of the human-related noun. We call this variant **Synonym-KG**.

4.4 Datasets

CrowSPairs (Nangia et al., 2020) is a common dataset for evaluating bias in LLMs. It contains 1,508 instance pairs in nine categories: Race/Color, Gender, Socioeconomic Status, Nationality, Religion, Age, Sexual Orientation, Physical Appearance, and Disability. Each instance pair consists of a stereotypical and anti-stereotypical sentence.

BOLD (Dhamala et al., 2021) is a dataset that contains 23,679 instances in five domains: Religion, Profession, Gender, Race, and Policy. Every instance is a sentence extracted from Wikipedia, and the first six to nine words are extracted from the sentence as a prompt. Since the first four categories in the BOLD dataset, i.e., Religion, Profession, Gender, and Race, are related to human, we selected the prompts in these four categories for

[3] As these sentences originate from LLaMA2, they were not introduced in the experiments for LLaMA2.

Models	Religion		Profession		Gender		Race	
	Polarity↓	Neutral↑	Polarity↓	Neutral↑	Polarity↓	Neutral↑	Polarity↓	Neutral↑
GPT2	60.4	18.3	47.6	41.7	66.6	23.7	62.8	25.9
Ours	59.9	19.8	44.2	46.7	54.0	35.0	54.2	33.8
GPT-Neo	58.7	21.1	48.7	43.3	67.8	24.3	66.2	25.0
Ours	52.8	29.6	35.1	59.8	34.4	60.8	57.1	37.8
LLaMA2	63.8	16.3	47.3	44.0	56.9	33.5	58.7	30.7
Ours	48.8	34.0	31.9	62.3	30.0	64.1	35.2	57.7

Table 5: Scores for autoregressive language models in **Regard**. Lower scores in **Polarity** represents less bias and higher scores in **Neutral** represents less bias.

evaluation.

4.5 Evaluation Metrics for Debiasing

Bias Score We need to evaluate bias on CrowS-Pairs. Following Touvron et al. (2023a), we calculated the perplexity for both sentences in each pair in a zero-shot setting to measure the model preference for the stereotypical sentence. The CrowSPairs score represents the percentage of instance pairs in which the stereotypical sentence has lower perplexity than the anti-stereotypical sentence in the total number of instance pairs. Since it is not easy to understand that the scores closer to 50 indicate less bias, we instead utilized the **Bias score**, which is defined as equal to |CrowSPairs score − 50|, to replace the CrowSPairs score. A lower score represents lower bias.

Models	**Bias Score (CrowSPairs) ↓**	
	BERT-base based	*RoBERTa-large based*
Based model	8.7*	16.8*
SENT-DEBIAS	7.3	-
DI	5.2	16.6*
Wikitext-tuning	5.8	15.2*
Gen-tuning	7.8	14.5*
Synonym-KG	7.1	12.6*
Ours	**4.1**	**6.5**
	GPT2-small based	*GPT-Neo1.3B based*
Based model	9.4*	12.6*
Backpack	7.3*	-
Wikitext-tuning	8.1*	10.6*
Gen-tuning	6.7	10.5*
Synonym-KG	8.7*	9.3
Ours	**5.4**	**6.6**

Table 6: Comparison with the baselines. Our method outperforms the baseline methods without training from scratch, that is, DI, Wikitext-tuning, and Gen-tuning, significantly with t-test ($p < 0.05$), indicated by $*$.The original CrowSPairs scores are in Appendix D.

Regard We further evaluated bias in the experiments for the autoregressive language models by the Regard metric (Sheng et al., 2019). This metric evaluates the bias by introducing the concept of regards (positive, neutral, and negative), which is similar to sentiment. When inputting a text, this metric computes three scores for the text. We first used prompts in the BOLD dataset to generate continuous texts with the model. Then, each generated text gets three scores for the corresponding regard from the Regard metric. Finally, we calculated the average score for each regard as the Regard score. Since we need to evaluate only whether a generated text shows bias (positive or negative) or not, we summed the Regard scores for positive and negative regards as a new category, *polarity*. A lower score for polarity indicates lower bias and a higher score for neutral indicates lower bias.

4.6 Downstream Tasks

To evaluate whether the performance of the models on downstream tasks might degrade with applying our method for debiasing, we tested eight downstream tasks, spanning four different aspects for testing.

- For testing the *Common Sense Reasoning* ability, we utilized **WinoGrande** and **PIQA** (Physical Interaction: Question Answering) under a zero-shot setting. In both tasks, when given a question, a model is required to determine the correct answer from the list of answer candidates.[4]

- For testing the *Sentiment Analysis* ability, we utilized **SST** (Stanford Sentiment Treebank) under fine-tuning. Models need to generate the correct sentiment polarity (negative or positive) for the input text.

- For testing the *Natural Language Inference* ability, we utilized **MultiNLI** (Multi-Genre

[4] The software used for the these two tasks is from https://github.com/EleutherAI/lm-evaluation-harness (Gao et al., 2021).

Models	WinoGrande	PIQA	SST	MultiNLI	RTE	QNLI	WNLI	QQP	Avg.
Autoregressive language models									
GPT2	55.3	70.3	94.3	86.3	74.0	92.0	47.9	90.2	76.2
Ours	56.0	57.3	95.2	86.5	72.2	92.0	43.7	90.3	74.1
GPT-Neo	54.9	71.1	94.2	86.2	68.2	92.0	46.5	91.0	75.5
Ours	54.4	66.4	94.6	86.9	71.5	91.7	49.3	89.5	75.5
LLaMA2	73.0	78.5	88.0	37.1	66.4	53.3	57.7	62.6	62.3
Ours	69.0	77.4	73.6	37.7	71.1	54.6	62.0	62.8	61.2
Masked language models									
BERT	50.2	47.7	93.7	85.5	56.0	91.2	43.7	89.9	69.7
Ours	50.2	48.2	93.2	85.7	60.7	92.3	43.7	89.8	70.5
RoBERTa	48.8	48.8	96.0	90.5	52.7	94.6	46.5	90.7	71.1
Ours	50.8	50.2	95.9	90.5	52.7	94.2	45.1	90.5	71.2

Table 7: Scores for different models on the downstream tasks. The scores for WinoGrande, SST, MultiNLI, RTE, QNLI, WNLI, QQP are accuracy. The score for PIAQ is the length-normalized accuracy. Higher scores indicate better performances. These downstream tasks were configured under a fine-tuning setting, with the exception of a zero-shot setting applied to WinoGrande and PIQA. Due to constraints in computing resources, a 3-shot setting was employed for LLaMA2 instead of fine-tuning. Our models obtained scores close to those of the original LLMs. T-test shows there are no significant difference between them.

Natural Language Inference), **RTE** (Recognizing Textual Entailment), **QNLI** (Question-answering Natural Language Inference), and **WNLI** (Winograd Natural Language Inference) under fine-tuning. MultiNLI and RTE require models to generate the entailment relationship (entailment or contradiction) between two sentences. QNLI requires models to discern whether a question and a text contain the correct answer. WNLI requires models to discern if a text whose pronouns were replaced with their referents is entailed by the original text.

- For testing the *Paraphrase Identification* ability, we utilized **QQP** (Quora Question Pairs) under fine-tuning. In this task, models need to discern whether two questions have the same meaning.

5 Results

5.1 Debiasing

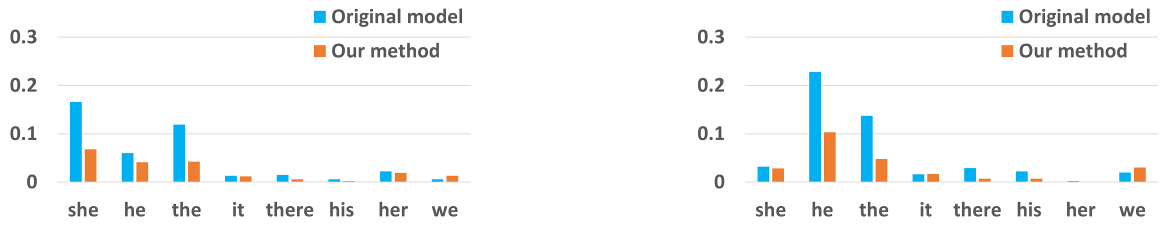
Table 4 presents the experimental results of various models on the CrowSPairs dataset. Our models exhibit lower bias in the majority of bias categories, that indicates the effectiveness of our method. Notably, LLaMA, only with 7 billion parameters, showcases lower bias than other LLMs with more parameters.

In contrast, the strong baseline **Wikitext-tuning** shows only a limited ability to mitigate bias, even though its pre-training data contain text with less bias. While **Gen-tuning** contains the hypernym information, its capacity for debiasing is lower than our method. It might be because the generated sentences for training it do not necessarily provide correct structured knowledge and contain bias. For instance, the actually generated sentence for "CEO", "*The CEO serving as the highest-level leader and managers, supervisors, and individual employees reporting to them.*", contains bias related to its superiority. This might prevent the models from accurately enhancing its representation.

The variant **Synonym-KG**, which used synonym information, also showed the effectiveness in mitigating bias. This might highlight the influence of structured knowledge, which contributes to the bias mitigation. However, its Bias scores are still higher than our method. This might indicate that the hypernym information is more useful for debiasing than the synonym information.

However, we found that our method shows a weaker control in some categories, e.g., Age, Sexual Orientation, and Physical Appearance, than the others (e.g., Race/Color and Gender). To investigate the reason, we calculated the ratio of knowledge for the categories, as shown in Table 8.[5] The weaker control is caused by the lack of the

[5] We trained a classification model to classify sentences obtained from the structured knowledge into one of the nine bias categories in the CrowSPairs dataset. More details are in Appendix F.



(a) X = nurse

(b) X = CEO

Figure 2: The effect of our method on the conditional probability distribution for the prompt "When the X walked into the room," (X=nurse, CEO). There is a smaller gap between the probability for "she" and "he" after training with our method.

Category	**Ratio(%)**
Race/Color	22.34
Gender	48.83
Socioeconomic Status	13.60
Nationality	5.91
Religion	3.39
Age	1.62
Sexual Orientation	0.32
Physical Appearance	0.77
Disability	3.2

Table 8: Ratio of knowledge for different bias categories.

structured knowledge related to these categories. This indicates that more structured knowledge is required when mitigating specific kinds of bias.

Table 5 shows the Regard scores for autoregressive language models. It is evident that models trained with structured knowledge exhibit lower polarized regard scores across various categories than the original models. This demonstrates that our method enables the models to use more general concepts from the hypernyms of a word when generating text, thus preventing excessive bias towards specific polarized content related to the word and favoring the generation of neutral content. The higher scores for the neutral regard further support this point.

As shown in Table 6, the Bias scores of our model are lower than other debiasing methods, especially in cases of larger-scale models with more bias, indicating that our method outperforms the previous methods in effective debiasing.

In order to conduct a more thorough comparison, our approach was also evaluated alongside baseline methods using the StereoSet dataset (Nadeem et al., 2021), a natural English dataset designed for assessing stereotypical bias. The experimental results, presented in Appendix E, illustrate that our method effectively mitigates bias while it slightly decreases the language modeling performance in the Autoregressive Language Models.

5.2 Downstream Tasks

Table 7 shows the results for the models on downstream tasks. Our models exhibit close performances to the original models across various downstream tasks. This demonstrates that using our method can ensure preservation of the generalization ability of the original models. Since the hypernyms of a word contain more general concepts for the word, after training with our method, although the representation of the word is adjusted, it still keeps the basic information. Thus, it can ensure that the trained models maintain their generalization abilities acquired during pre-training without leading to dramatic degradation in the performance of the downstream tasks.

5.3 Case Study

Figure 2 shows the conditional probability distribution among words after inputting a specific prompt "When the X walked into the room," to both GPT2-Neo trained with our method and the original GPT2-Neo. We used different nouns, "nurse" and "CEO", to replace "X". The gap of the probabilities between "she" and "he" is smaller after applying our method than the original model in both cases. This clearly shows the effect of our method in mitigating gender-related bias for these cases.

6 Conclusion

We proposed a simple method that utilizes structured knowledge to mitigate the issue of bias within LLMs. Our method trains the LLMs by incorporating information of hypernyms into the representations for the words to mitigate the bias. Experimental results from both autoregressive and masked language models demonstrated that our method effectively controls the inherent bias in LLMs without compromising the performance in downstream tasks. Comparative studies with other debiasing techniques showed that our method achieves a better debiasing performance. Since our method does not require models to be trained from scratch, it boasts the advantages of low cost and scalability.

7 Limitations

Since we used only limited human-related nouns to extract knowledge from the existing knowledge base, the knowledge is not comprehensive. This causes a weaker control in certain bias categories. Developing a method for acquiring more comprehensive structured knowledge or data augmentation will be the focus of future research.

References

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. 58:2.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357. Curran Associates, Inc.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation. page 8.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

John Hewitt, John Thickstun, Christopher Manning, and Percy Liang. 2023. Backpack language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9103–9125, Toronto, Canada. Association for Computational Linguistics.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in BERT - measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199, Seattle, Washington. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 2611–2624. Curran Associates, Inc.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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, Online. Association for Computational Linguistics.

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, and Dirk Hovy. 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*, pages 2398–2406, Online. Association for Computational Linguistics.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 446–457.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford and Jeffrey Wu. 2019. Rewon child, david luan, dario amodei, and ilya sutskever. 2019. *Language models are unsupervised multitask learners. OpenAI blog*, 1(8):9.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Victor Zimmermann and Maja Hoffmann. 2022. Absinth: A small world approach to word sense induction. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 121–128, Potsdam, Germany. KONVENS 2022 Organizers.

A Objectives for LLMs

A.1 Autoregressive Language Models

For autoregressive language models, such as GPT2 (Radford and Wu, 2019), we utilize the next-token prediction objective to train the model. The objective is calculated as follows:

$$\mathcal{L} = -\sum_{t=1}^{|s|} \log P(s_t | s_{<t}), \tag{1}$$

where s is a sequence of sentences constructed from the obtained structured knowledge pieces. t denotes the t-th token in the sequence.

A.2 Masked Language Models

For masked language models, such as BERT (Devlin et al., 2019), we utilize the masked token prediction objective to train the model. The objective is calculated as follows:

$$\mathcal{L} = -\sum_{\hat{s} \in m(s)} \log P(\hat{s} | s_{\setminus m(s)}), \tag{2}$$

where s is a sequence of sentences constructed from the obtained structured knowledge pieces, $m(s)$ denotes the masked tokens in s, and $s_{\setminus m(s)}$ denotes the remaining tokens in s.

B Extracting Human-related Nouns from WordNet

We processed the nouns retrieved from the "noun.person" category in the WordNet as follows:

1. We retrieved all the nouns belonging to the "noun.person" category in WordNet.

2. Since we found many person names, denoted with initial capitalization, in those nouns, we extracted the person names and their definitions in WordNet and replaced the names with the definitions, for example, "William Shakespeare" with ”dramatist".

3. We also found compound nouns consisting of multiple words. For those compound nouns, we retained only the last head noun, for example, from the noun "nationalist leader", we extracted only "leader".

4. Finally, we removed any duplicated nouns.

C Training Data for Gen-tuning

For every extracted noun, we formulated a prompt, "Use a sentence to describe the relation between X and Y." with *X* representing the noun and *Y* representing its hypernym. Subsequently, we input this prompt into the LLaMA2-7B model to obtain responses, which serve as our training data. Excluding instances where the model failed to generate a response, the total number of sentences is 33,105. Three examples of generated sentences, along with the corresponding noun-hypernym pairs, are presented in Table 9.

D CrowSPairs Scores

In Autoregressive Language Models, the perplexity is calculated by evaluating the probability of generating each token sequentially based on the preceding context. In Masked Language Models, we calculated the perplexity by evaluating the likelihood of predicting masked tokens within the given context. Table 10 shows the original CrowSPairs scores for different models in Table 4. Table 11 shows the original CrowSPairs scores in Table 6.

Noun	Hypernym	Sentence
potboy	employee	A potboy is typically an entry-level employee in a restaurant or bar who assists the cooks and servers, while an employee is a person who is hired by an organization to work in a specific job or position.
publisher	owner	The publisher and owner are closely related as the publisher is typically the entity that owns the rights to publish and distribute the content, while the owner is the individual or organization that legally owns the content itself.
comber	worker	Comber and worker are related in that a comber is a type of worker who specializes in combing or cleaning the fur of animals, such as sheep or goats, to prepare it for use in the textile industry.

Table 9: Examples of generated sentences from pairs of a hypernym and hyponym by LLaMA2.

E Results on the StereoSet Dataset

StereoSet (Nadeem et al., 2021) comprises 16,995 instances distributed across four domains: gender, profession, race, and religion. Within each instance, a stereotypical, a anti-stereotypical and a meaningless sentences are included. The SS score is employed to indicate the percentage of instances in which the stereotypical sentence exhibits lower perplexity than the anti-stereotypical sentence in the total number of instances in StereoSet. The LMS score is employed to indicate the percentage of instances in which the stereotypical sentence has lower perplexity than the meaningless sentence in the total number of instances in StereoSet. The experimental results are presented in Table 12. A SS score closer to 50 indicates a reduced level of bias. A higher LMS score indicates better language modeling performance.

F Classification Model

We trained the classification model based on the large version of RoBERTa. We used the stereotypical sentences in the CrowSPairs dataset to train this model. The nine labels are the same as the bias categories in the CrowSPairs dataset. The learning rate was set to 1e-5. We used Adam as the optimizer.

Models	Race/ Color	Gender	Socioeconomic Status	Nationality	Religion	Age	Sexual Orientation	Physical Appearance	Disability	Avg.
Autoregressive Language Models										
GPT2	59.5	57.6	66.9	47.2	77.1	55.2	69.0	60.3	66.7	60.5
Wikitext-tuning	57.0	56.9	68.6	65.4	72.3	73.6	72.7	74.6	78.3	63.7
Gen-tuning	58.3	59.9	68.0	48.4	68.6	54.0	70.2	63.5	61.7	60.1
Synonym-KG	55.8	59.8	64.0	47.8	67.6	59.8	72.6	68.3	68.3	59.4
Ours	55.4	59.5	62.8	49.7	68.6	52.9	70.2	60.3	65.0	58.6
GPT-Neo	62.2	59.5	67.4	54.7	68.6	59.8	69.0	61.9	71.7	62.6
Wikitext-tuning	58.5	56.9	68.6	54.1	67.6	58.6	67.9	60.3	70.0	60.6
Gen-tuning	55.8	61.1	69.7	51.6	69.5	63.2	66.7	61.9	65.0	60.5
Synonym-KG	56.2	53.1	62.8	56.6	65.7	57.5	71.4	68.3	75.0	59.3
Ours	52.7	52.3	64.5	49.1	62.9	57.5	72.6	61.9	65.0	56.6
LLaMA	66.9	62.6	68.0	58.5	78.1	69.0	82.1	81.0	80.0	68.2
LLaMA2	67.6	60.3	68.0	58.5	76.2	71.3	76.2	76.2	76.7	67.4
Wikitext-tuning	65.7	59.2	68.6	61.0	74.3	66.7	73.8	77.8	80.0	66.4
Synonym-KG	65.5	59.2	64.5	64.2	77.1	70.1	71.4	73.0	80.0	66.5
Ours	57.0	56.9	68.6	65.4	72.3	73.5	72.6	74.6	78.3	63.6
Large Language Models										
†LLaMA 65B	57.0	70.6	71.5	64.2	79.0	70.1	81.0	77.8	66.7	66.6
†OPT 175B	68.6	65.7	76.2	62.9	68.6	67.8	78.6	76.2	76.7	69.5
†GPT3	64.7	62.6	73.8	61.6	73.3	64.4	76.2	74.6	76.7	67.2
Masked Language Models										
BERT	57.6	58.8	62.2	48.4	76.2	64.4	73.8	69.8	58.3	60.5
Wikitext-tuning	57.2	59.5	26.7	37.1	24.8	29.9	32.1	34.9	35.0	43.0
Gen-tuning	54.3	53.4	61.6	52.8	77.1	57.5	66.7	63.5	61.7	58.0
Synonym-KG	51.0	55.7	62.2	42.1	73.3	58.6	76.2	68.3	66.7	56.6
Ours	49.2	54.6	61.0	48.4	63.8	54.0	80.9	68.3	75.0	56.3
RoBERTa	69.3	59.5	72.1	56.0	74.3	66.7	67.9	74.6	68.3	66.8
Wikitext-tuning	66.9	60.3	67.4	57.2	70.5	60.9	70.2	73.0	68.3	65.2
Gen-tuning	68.8	56.5	74.4	47.2	67.6	66.7	67.9	69.8	60.0	64.5
Synonym-KG	62.6	53.8	69.8	50.3	76.2	65.5	67.9	65.1	75.0	62.6
Ours	54.1	48.1	29.6	32.7	18.1	50.6	55.9	38.1	48.3	43.5

Table 10: Original CrowSPairs scores for different models. The scores denoted by † are reported in Touvron et al. (2023a).

Models	CrowSPairs scores ↓	
	BERT-base based	*RoBERTa-large based*
Based model	58.7	66.8
SENT-DEBIAS	42.7	-
DI	55.2	66.6
Wikitext-tuning	44.2	65.2
Gen-tuning	57.8	64.5
Synonym-KG	57.1	62.6
Ours	54.1	43.5
	GPT2-small based	*GPT-Neo1.3B based*
Based model	59.4	62.6
Backpack	57.3	-
Wikitext-tuning	58.1	60.6
Gen-tuning	56.7	60.5
Synonym-KG	58.7	59.3
Ours	55.4	56.6

Table 11: Original CrowSPairs scores for different baselines.

Models	**SS**					**LMS**				
	Race	**Gender**	**Profession**	**Religious**	**Avg.**	**Race**	**Gender**	**Profession**	**Religious**	**Avg.**
Autoregressive Language Models										
GPT2	63.2	67.1	64.8	59.5	64.2	94.0	96.5	94.6	93.7	94.5
Wikitext-tuning	60.2	65.5	66.0	60.8	63.0	92.1	95.7	93.3	91.1	93.0
Gen-tuning	60.9	73.3	67.9	68.4	65.4	90.9	94.9	92.6	89.9	92.0
Synonym-KG	59.8	70.6	66.7	65.8	64.0	89.6	93.7	90.9	92.4	90.7
Ours	58.3	69.8	65.8	55.7	62.8	89.4	94.9	91.6	91.1	91.0
GPT-Neo	63.0	67.1	64.9	67.1	64.4	95.2	95.3	94.2	96.2	94.9
Wikitext-tuning	63.3	67.5	64.8	69.6	64.2	94.8	96.1	94.1	94.9	94.7
Gen-tuning	58.3	72.2	67.9	64.6	63.9	90.1	92.5	90.7	92.4	90.7
Synonym-KG	61.1	69.4	66.2	65.8	64.2	85.7	90.6	86.7	88.6	86.8
Ours	56.4	66.3	63.5	64.6	60.6	89.0	90.2	87.4	94.9	88.7
LLaMA2	63.0	70.6	64.9	58.2	64.2	92.8	96.5	93.6	91.1	93.5
Wikitext-tuning	63.1	70.2	64.1	63.3	65.2	89.8	94.5	91.0	93.7	91.0
Synonym-KG	62.7	67.5	62.0	54.4	62.7	86.5	93.3	85.3	79.7	86.6
Ours	61.2	62.5	62.2	56.8	60.7	90.5	94.1	92.3	91.1	91.7
Masked Language Models										
BERT	62.9	65.9	63.1	64.6	63.4	92.4	94.9	92.2	92.4	92.6
Wikitext-tuning	58.1	69.9	63.3	66.3	63.3	91.5	90.2	90.4	91.1	90.9
Gen-tuning	58.8	70.2	64.0	64.8	62.8	93.8	93.7	93.3	89.9	93.4
Synonym-KG	61.2	71.0	64.0	67.1	63.7	94.1	95.3	91.0	92.4	93.0
Ours	59.6	71.8	62.5	67.1	62.4	94.2	94.1	93.0	93.7	93.7
RoBERTa	60.7	69.8	64.7	53.2	63.1	93.8	97.3	93.6	93.7	94.1
Wikitext-tuning	59.4	70.2	67.0	68.4	64.0	92.6	96.9	94.4	89.9	93.7
Gen-tuning	62.9	73.7	66.9	67.1	65.9	94.6	96.1	94.7	94.9	94.8
Synonym-KG	60.5	66.3	66.0	59.5	63.3	91.9	96.5	91.1	92.4	92.2
Ours	55.0	58.7	58.3	57.0	57.2	94.3	97.3	91.7	91.1	93.5

Table 12: Scores for different models on StereoSet. A SS score closer to 50 represents less bias. A higher LMS score represents better language modeling performance.