Absorbing Commonsense Knowledge from LLMs for Improving Social Fairness in Pre-trained Language Models

Anonymous ACL submission

Abstract

Pre-trained Language models (PLMs) are trained on inherently socially biased sources, inevitably causing undesirable application impacts. Current debiasing paradigm involves identifying bias from external corpora, which have limited quality, diversity, or equivalence among different groups, potentially impacting bias location and debiasing effectiveness. In light of this, we advance fairness in PLMs by absorbing coherent, balanced, and semantically informative social Commonsense Knowledge (CK-Debias) automatically generated from large language models (LLMs). Our study addresses the demographic CK generation from LLM and explores strategies to optimize CK utilization. This is achieved by employing causal analysis to align knowledge for estimating bias space and identifying the most biased prompts to enhance bias avoidance capability. Experiment results on public datasets and intrinsic and extrinsic metrics show that CK-Debias can significantly reduce multiple social biases across various PLMs while keeping their language expressiveness intact.

1 Introduction

Lightweight pre-trained language models (Devlin et al., 2019; Liu et al., 2019) have made unprecedented progress across a broad spectrum of tasks, ranging from language understanding (Meng et al., 2022), document classification (Bhardwaj et al., 2021), to multitasks text generation , which are more suitable for deployment on resource-constrained devices compared to LLM. However, the prevalence of out-of-distribution issues (Lu et al., 2022) or inherent stereotypical remarks in the training corpus may inadvertently reinforce biased or stereotypical representations (Caliskan et al., 2017), leading to potential unfairness across diverse demographic groups. In specialized domains like law, medicine, or human resources (Jatobá et al., 2019), ensuring the neutrality and fairness of their encoded representations becomes crucial.

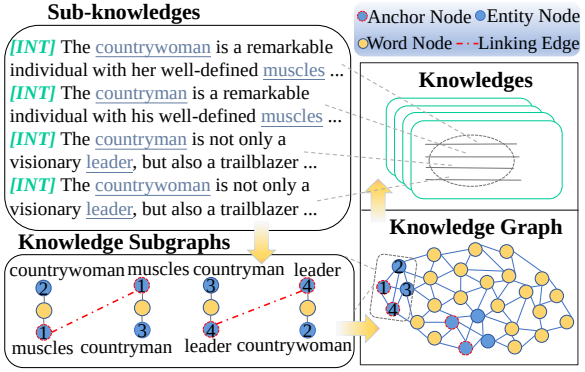


Figure 1: Our rationale to generate social CK. Each sub-knowledge has a knowledge subgraph that includes Anchor/Entity Node (i.e., specified attribute/target words). Subgraphs with the same Anchor Node can be associated by the Linking Edge to form a larger subgraph.

Recent task-agnostic contextualized debiasing works (Kaneko and Bollegala, 2021; Cheng et al., 2021; He et al., 2022) are devoted to designing specific loss functions to fine-tune PLMs toward mitigating inherent biases. Despite the remarkable success, they all involve drawing sentences from external corpora to identify and mitigate biases, aiming to include sufficient diversity across demographics. Moreover, some of them (Ghanbarzadeh et al., 2023; Zhou et al., 2023) try to achieve equivalence among opposite demographics via balanced counterfactuals, which might yield incoherent or noisy knowledge when multiple entities are referenced. However, collecting high-quality corpora is usually costly, and noisy knowledge is easily introduced (Zheng et al., 2023), resulting in insufficient or inaccurate bias mitigation. While certain study (Guo et al., 2022) tries to generate prompts as a replacement for external corpora, it also relies on the additional Wikipedia in search space, and its short prompts often fail to consider syntax or context, which leads to gaps when used in semantic downstream tasks.

Recent foundation models like ChatGPT (Chiang et al., 2023), LLaMa (Touvron et al., 2023), and Gemini (Team et al., 2023) have excelled in widespread applications, and extensive endeavors are embarked upon to rectify inherent biases (Gallegos et al., 2023) and elevate the commonsense research (West et al., 2022; Plenz et al., 2023). The LLMs have exhibited notable knowledgeable ability (Bian et al., 2023) as knowledge bases to generate CK accurately, its general success draws our insights to leverage them for debiasing lightweight PLMs. Arguably, we attribute PLMs' bias to limited social commonsense and bias avoidance capacity, since the debiasing guidance for lightweight PLMs is generally notably smaller than LLMs.

In this paper, we refrain from using existing external corpora as previous studies but resort to LLMs, and propose a novel paradigm that integrates automated commonsense knowledge sourced from LLMs to improve debiasing performance. The rationale is depicted in Fig. 1, which can be interpreted from a fair tuple of $(attr_1, target, attr_2)$ like muscles $(target)$ that can be owned by both countryman $(attr_1)$ and countrywoman $(attr_2)$. An important observation in our paradigm is that not all knowledge extracted from LLMs is directly applicable for debiasing, as it might lead to negative knowledge transfer (Zheng et al., 2023). Hence, we employ causal analysis to distinguish knowledge aligned with PLM's and integrate it to advance social fairness. For the remaining unaligned knowledge, we introduce a strategic bias location and mitigation process, which identifies the most biased prompts to refine PLM's debiasing capability. To mitigate the impact of the fine-tuning procedure on the model's expressive capabilities, we design a specialized loss function that can maintain model parameters as stable as possible. Our contributions are three-fold:

- Differing from reliance on existing external corpora, we are the pioneers in leveraging LLM-generated commonsense knowledge to supply rich and high-quality semantic resources for debiasing lightweight PLMs.
- We apply a structure causal model (SCM) to analyze the limitations of traditional debiasing methods, and an improved causal graph is employed to effectively harness LLM-generated knowledge.
- CK-Debias can effectively alleviate various types of biases, demonstrating superior performance across multiple PLMs, yielding superior performance in both intrinsic and extrinsic evaluations, while maintaining intact model expressiveness.

The code of CK-Debias is anonymously available at `https://anonymous.4open.science/r/CK-Debias-49B4/`.

2 Related Works

Language models, developed with data often imbued with inherent biases, can inadvertently introduce biases into their applications, thereby spurring a growing body of research aimed at mitigating biases. The earliest efforts mainly focus on debiasing static word embeddings such as Glove (Pennington et al., 2014) and word2vec (Mikolov et al., 2013) via projection-based (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019) or adversarial methods (Elazar and Goldberg, 2018; Xie et al., 2017).

The in-depth research inspired the follow-up studies debias pre-trained contextualized embeddings, as the widespread use of BERT (Devlin et al., 2019) and their variants (Lan et al., 2020; Liu et al., 2019). Based on whether they directly combine with associated downstream tasks, the external corpora-based methods can be divided into: (1) *Task-Agnostic* methods constitute the majority: **Sent-Debias** (Liang et al., 2020) and **FairFil** (Cheng et al., 2021) are *post-hoc* methods that keep the PLM parameters untouched, and **ADEPT** (Yang et al., 2023) proposes a novel training criterion that only trains the continuous prompt parameters but keeps the base model frozen; **Auto-Debias** (Guo et al., 2022), **Context-Debias** (Kaneko and Bollegala, 2021), and **MABEL** (He et al., 2022) remove biases in PLM via *fine-tuning* using various well-designed bias-neutralizing loss functions. (2) *Task-Aware* approaches, emerging recently, aim to prevent bias recurrence when applying debiased models in practical applications. Recent innovations like **Causal-Debias** (Zhou et al., 2023) unifies the debiasing procedure with downstream fine-tuning via causal invariant learning. Similarly, **Gender-tuning** (Ghanbarzadeh et al., 2023) deploys a debiasing tool for any PLM that works with original fine-tuning.

Despite the notable success of debiasing, their efficacy largely relies on the quality, quantity, and diversity of the corpora used, such as WikiText-2 (Merity et al., 2017), Standford Sentimente Treebank (Socher et al., 2013), Reddit, etc. They indiscriminately use the matched sentences to locate bias, and the simple substitution can cause a nega-

tive transfer when multiple entities are referenced. To achieve a more balanced bias distribution across groups, they typically balance sentence matching between groups, often at the expense of excluding redundant sentences. In light of LLMs' recent extensive adoption and accomplishments in diverse NLP tasks (Wang et al., 2023), we are motivated to create prompts that guide LLMs towards producing high-quality, abundant, and semantically rich social commonsense knowledge. The ultimate goal is to leverage generated social CK to contribute to our debiasing efforts in PLMs.

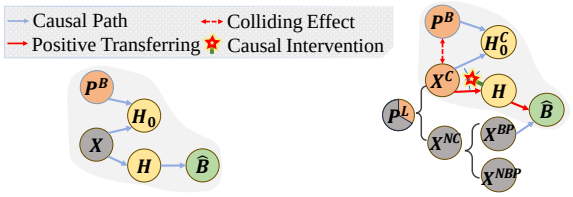
3 Methodology

In this section, we answer how to generate knowledge and explore ways to optimize the usage of generated knowledge (i.e. positive transfer and bias avoidance), as depicted in Fig. 3. Note that CK-Debias is generic to various biases or PLMs, with gender bias serving as just our example.

3.1 Demographic Commonsense Knowledge Generation from LLMs

Let $\mathcal{W}_a = \{(a_1^{(1)}, a_2^{(1)}, \cdots, a_d^{(1)}), (a_1^{(2)}, a_2^{(2)}, \cdots, a_d^{(2)}), \cdots\}$ denotes *attribute* words composed of multiple d-tuple and $\mathcal{W}_t = \{v_1, v_2, \cdots\}$ denotes *target* words, respectively. In the case of binary gender ($d = 2$), attribute words are gender-specific pairs: (*she*, *he*), (*woman*, *man*), (*mother*, *father*), target words consist of gender-neutral words (e.g., *nurse, engineer, professor*). For prompting LLM, we use two well-designed system prompts to automatically generate a pair of sentences in two steps (details *cf.* Fig. 7 in Appendix A). The generated CK sentences, containing $(a_i, \underline{v_t})$, $(a_j, \underline{v_t})$, shape a *bundle* sample with identical target words $\underline{v_t}$ over d-tuple attribute words. Its rationale can be interpreted as a fair triplet (a_i, $\underline{v_t}$, a_j), indicating that each target word $\underline{v_t}$ can establish an association with both pairwise attribute words a_i and a_j, as illustrated in Fig. 1. For brevity, we denote $x^{(n)} = \{x_1^{(n)}, x_2^{(n)}, \cdots, x_d^{(n)}\}$ as the n-th bundle sample below. We merely alter specific attribute words, aiming to retain consistency across other components of each bundle to keep semantic similarity. Additionally, we strive for uniformity in quantity and length, promoting fairness across various demographic groups. Examples of $(a_i, \underline{v_t})$, $(a_j, \underline{v_t})$, and their corresponding generated bundle samples are provided in Appendix A.

3.2 Debiasing the PLMs via Generated Commonsense Knowledge



(a) Conventional Fine-Tuning (b) Positive Transfer

Figure 2: The comparison of SCM between conventional methods and our CK-Debias.

We employ SCM to depict the causal association among data, models, and hidden factors. Subsequently, we apply this SCM to emphasize the challenges posed by conventional debiasing methods that rely on external corpora. Finally, we introduce an enhanced causal graph that harnesses commonsense knowledge to alleviate bias in PLMs.

Causality between data, models, and hiddens. As shown in Fig. 2, we denote the pre-trained data as P; the external corpora as X; the hidden of X extracted by the initial pre-trained model and fine-tuned model as H_0 and H, respectively; the bias magnitude predicted by H on external corpora as $\hat{B}$. The causal associations are: (1) $X \rightarrow H \rightarrow \hat{B}$: $X \rightarrow H$ denotes the hidden H, which is derived by PLMs from the matched sentence found in external corpora, and $H \rightarrow \hat{B}$ is the computed distance to measure bias magnitude $\hat{B}$ according to the hidden H; (2)$X \rightarrow H_0 \leftarrow P$: initial hidden H_0 is determined by both pre-trained data P and input external data X. The collider H_0 is the joint outcome of the independent causes P and X. According to casual theory (Neal, 2020), once the common effect H_0 is observed, its causes P and X become dependent, so in our scenarios, the colliding effect between pre-trained data and external corpora is preserved during the fine-tuning based debiasing process.

Conventional debiasing methods in Fig. 2 (a) rely on X sourced from external corpora, which may contain noisy data. If directly applying X to locate biases without distinction, the retained noise knowledge may result in inaccuracies of bias identification and hinder negative knowledge transfer (Zheng et al., 2023). We attribute this issue to the missing colliding effect between the external corpora X and pre-trained data P, which can also be viewed as a deficit in the alignment of their hidden spaces. Moreover, traditional methods partially mitigate bias by simply substituting attribute slots to achieve balanced counterfactual augmentations, but this

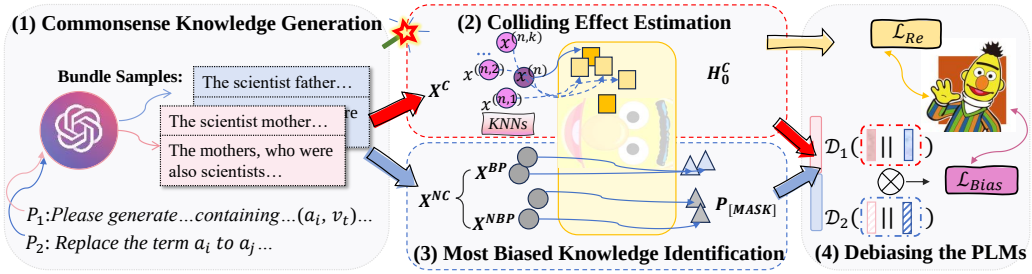


Figure 3: Overview of CK-Debias. **(1)** Obtain *bundle* samples by prompting LLM via well-designed system prompts based on pairwise *attribute* and *target* words. **(2)** Find the aligned knowledge via colliding effect estimation to achieve positive transfer. **(3)** Identify the most biased knowledge from remaining unaligned knowledge to enhance bias avoidance. **(4)** Debiasing PLMs using knowledge from (2) and (3), while keeping the PLMs' expressiveness.

can yield incoherent or contradictory text when multiple entities are referenced within a sentence. Additionally, when X is more biased towards one particular bias, debiasing unrelated bias may engender complications (Liang et al., 2020). For those issues, we employ LLM to generate bias-specific commonsense knowledge within semantic information (*cf.* Section 3.1), which is more conducive to locating bias and mitigating bias subsequently.

Our CK-Debias. We observe that the knowledge extracted from LLM cannot be directly applied to mitigate biases since the pre-training data for lightweight PLMs is typically much less than that of LLM. To achieve a positive transfer of generated commonsense knowledge and to enhance the debiasing impact of PLMs' contextualized embeddings, it is crucial to establish an alignment between the hidden space of PLMs and LLM. As shown in Fig. 2 (b), we split the generated bundle knowledge into two nodes X^C and X^{NC}. X^C represents the bundle samples where we calculate colliding effects, and their knowledge should align with PLMs to enhance their fairness. X^{NC} signifies samples exempt from collision effect calculation, presenting biased knowledge due to negative transfer – a phenomenon PLMs should strive to evade. To maximize its utilization, X^{NC} are used to find the most biased prompts X^{BP} that induce bias in PLMs, enabling PLMs to avoid bias when assimilating commonsense knowledge impartially. In summary, the fine-tuned PLMs assimilate LLM's commonsense knowledge by utilizing colliding effects ($P \leftrightarrow X^C$) while striving to avoid bias via most biased prompts ($X^{BP} \rightarrow \hat{B}$). When conditioning on H_0^C, the final bias magnitude depends on the degree of assimilating aligned knowledge from causal paths $P \leftrightarrow X^C \rightarrow H \rightarrow \hat{B}$ (positive transfer), and avoiding most biased knowledge $X^{NC} \rightarrow H \rightarrow \hat{B}$ (bias avoidance), respectively.

3.3 Estimating Colliding Effect and Finding Most Biased Knowledge

Colliding Effect Estimation. Considering when predicting $\hat{B}^{(n)}$, we obtain the hidden state from initial model $h_0^{(n)} = \mathcal{M}_0(x^{(n)})$. Controlling $H_0 = h_0^{(n)}$ means the input X in the causal graph represents all samples whose hidden feature is $h_0^{(n)}$. Nevertheless, the sole satisfying candidate for this condition is $x^{(n)}$ due to the sparsity inherent in high-dimensional spaces. Alternatively, if we slightly loosen this constraint, the colliding effect is unlikely to vanish instantly. Hence, we choose to approximate the colliding effect Ψ between P and X^C with the joint prediction of K-Nearest-Neighbor (KNN) samples. When conditioning on collider H_0, Ψ can be calculated as:

$$\Psi = \sum_{n=1}^{N} \psi^{(n)} \approx \tag{1}$$
$$\sum_{n=1}^{N} \sum_{k=0}^{k_n} B\left(\mathcal{M}_H(X^C = x^{(n,k)})\right) S_H\left(x^{(n)}, x^{(n,k)}\right)$$

where N is the total number of bundle samples from aligned knowledge X^C, k_n is the number of KNNs of n-th bundle sample for estimating $\hat{B}^{(n)}$. $x^{(n,k)}$ is the k-th nearest neighbor of $x^{(n)} \in \mathcal{X}^N$, whose similarity is greater than or equal to the preset threshold θ. $S_H(\cdot,\cdot)$ is the similarity function between $x^{(n)}$ and $x^{(n,k)}$ (abbreviate as $S_{n,k}$ for brevity), and $\sum_{k=0}^{k_n} S_H\left(x^{(n)}, x^{(n,k)}\right) = 1$. Given that $x^{(n)}$ exhibits the highest similarity with itself, we set $x^{(n,0)} = x^{(n)}$ as the anchor bundle sample when $k = 0$. $B\left(\mathcal{M}_H(X^C = x^{(n,k)})\right)$ represents the bias magnitude prediction of $\hat{B}^{(n)}$ when $x^{(n,k)}$ is the model $\mathcal{M}_H$'s input. Eq. (1) shows that the total causal effect Ψ is the sum of N aligned bundle sample's causal effect $\psi^{(n)}$, and each $\psi^{(n)}$ can be approximated by the weighted sum of the bias

prediction when the model input is the anchor bundle sample $x^{(n)}$ and its KNNs. The rationale is matching KNNs using the parameters in PLM is essentially a process of aligning the commonsense extracted from LLM with PLM.

Most Biased Knowledge Identification. For the remaining unaligned knowledge X^{NC}, we find knowledge with highly biased prompts among them, enabling the location of bias in PLMs. For each bundle sentence $x^{(n)} \in \mathcal{X}^{NC}$, the identical component in $x^{(n)}$ is regarded as the prompt , namely excluding the attribute and target words utilized in generating the CK sentence, as well as the personal pronouns altered by the LLM. The most biased prompt has the highest disagreement when $\mathcal{M}$ predicts target words v_t (replaced by `[MASK]` placeholder in advance) over each bundle sample $x^{(n)}$ regarding different demographic groups a_i, a_j. So, when traversing through $\mathcal{X}^{NC}$, if one bundle sample contains the most biased prompt, we regard it as the most biased knowledge $\mathcal{X}^{BP} \subset \mathcal{X}^{NC}$, which is the subset of unaligned knowledge. In practice, we compute Jensen–Shannon divergence (JSD) score between distributions $p(\texttt{[MASK]} = v_t|\mathcal{M}, x_{BP}^{(n)}), v_t \in \mathcal{W}_t$ on each group a_i, a_j to measure the disagreement. The obtained most biased knowledge can locate bias and then subsequently enable PLMs to avoid bias, thus also maximizing the usage of the extracted knowledge. More details about the most biased knowledge *cf.* Appendix F.

3.4 Training Objective

After obtaining the aligned knowledge X^C and the unaligned knowledge with most biased prompt X^{BP}, we proceed our CK-Debias as follows: given a pre-trained model $\mathcal{M}_0$ with initial hidden H_0, we aim to fine-tune $\mathcal{M}$ to attain the optimal hidden H with minimal biases $\hat{B}$. The overall debiasing objective is as follows:

$$\mathcal{L}_{Bias} = (1-\alpha) \underbrace{\sum_{x^{(n)} \in \mathcal{X}^C} \sum_{k=0}^{k_n} \mathcal{D}_1\big(X^C = x^{(n,k)}\big) S_{n,k}}_{\mathcal{L}_C} + \alpha \underbrace{\sum_{x^{(n)} \in \mathcal{X}^{BP}} \mathcal{D}_2\big(X^{BP} = x^{(n)}\big)}_{\mathcal{L}_{BP}} \quad (2)$$

where the first term $\mathcal{L}_C$ is a rewrite of colliding effect Ψ estimated from X^C. To integrate the aligned knowledge X^C into PLMs, we distinguish the strength of knowledge preservation for each bundle sample by selecting the KNNs – $x^{(n,k)}$ for the anchor sample $x^{(n)}$. $\mathcal{D}_1$ quantifies the relative JSD between bundle samples with pairwise attribute words and those with neutral words in a high-dimensional space, which is defined as:

$$\mathcal{D}_1\left(x^{(n)}\right) = \sum_{i,j \in \{1,\dots,d\}, i<j} \left\{ JS\left(R^{x_i^{(n)}} \| R^{x_j^{(n)}} \right) \right\}$$

where $R^{x_i^{(n)}} = Distance(E^{target} | e^{x_i^{(n)}})$ measures the distance from sentence $x_i^{(n)}$ containing attribute words $a(i)$ to sentences containing all target words, and $E^{target} = [e^{x_{v_1}}, e^{x_{v_2}}, \cdots]$. $\mathcal{L}_{BP}$ mitigates bias derived from the obtained most biased knowledge, and $\mathcal{D}_2$ is defined as:

$$\mathcal{D}_2\left(x^{(n)}\right) = \sum_{i,j \in \{1,\dots,d\}, i<j} \left\{ JS\left(P^{x_i^{(n)}} \| P^{x_j^{(n)}} \right) \right\}$$

$$P^{x_i^{(n)}} = p(\texttt{[MASK]} = v_t | \mathcal{M}, x_{BP}^{(n)}), v_t \in \mathcal{W}_t$$

where $\mathcal{D}_2$ computes the JSD scores to minimize the disagreement between the predicted `[MASK]` token, which means a fair NLP system should yield scores independent of the selection of the attribute concepts. Two distance losses are linearly interpolated by a tunable coefficient α.

As fine-tuning with full parameter modifications can potentially harm the expressiveness of PLM, we add an auxiliary representation loss $\mathcal{L}_{Re}$ to preserve the inherent language modeling capability, which is defined as:

$$\mathcal{L}_{Re} = MSE(\mathcal{M}_H(\cdot) || \mathcal{M}'_H(\cdot)) \quad (3)$$

where $\mathcal{L}_{Re}$ measures disparity between the original model's hidden states $\mathcal{M}_H$ and the debiased model's hidden states $\mathcal{M}'_H$ via Mean Squared Error (MSE), striving to minimally alter the PLM's parameters. The overall training loss is as follows:

$$\mathcal{L} = \mathcal{L}_{Bias} + \lambda \cdot \mathcal{L}_{Re},$$

wherein $\mathcal{L}_{Re}$ is tempered by the hyper-parameter λ. Detailed formula derivation, hyper-parameter configuration, and algorithm process are presented in Appendix B.

4 Experiments

Benchmarks. We compare CK-Debias with benchmarks based on external corpora: *Task-Agnostic* models including: **Context-Debias** (Kaneko and Bollegala, 2021), **Auto-Debias** (Guo et al., 2022),

FairFil (Cheng et al., 2021), and **MABEL** (He et al., 2022); and *Task-Aware* methods including **Causal-Debias** (Zhou et al., 2023) and **GenderTuning** (Ghanbarzadeh et al., 2023). In *Task-Agnostic* methods, the debiasing stage is independent of fine-tuning downstream tasks, and CK-Debias belongs to it. *Task-Aware* methods directly combining downstream tasks for debiasing.

LLM and PLMs. We utilize GPT-3.5-turbo API as the source LM for CK generation. In practice, we establish multiple threads to enhance the efficiency of knowledge generation, which query GPT in parallel and meanwhile store the results. Three masked PLMs are as the backbones: BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and RoBERTa (Liu et al., 2019). Following (Guo et al., 2022), we implement them using Huggingface Transformers library (Wolf et al., 2020).

Bias Word Lists. We generate commonsense knowledge sentences by prompting GPT-3.5 based on human-created word lists. Following prior studies, we choose the gender/racial/religion word lists from (Kaneko and Bollegala, 2021), (Manzini et al., 2019), and (Liang et al., 2020) respectively – *cf.* Appendix C for details.

Evaluating Metrics. Given the diverse ways in which bias can be embedded in language, we quantify biases in PLM embeddings against a diverse set of intrinsic and extrinsic indicators, including *intrinsic* metrics with **SEAT** (May et al., 2019), **CrowS-Pairs** (Nangia et al., 2020) and **StereoSet** (Nadeem et al., 2020), and *extrinsic* metrics with **WinoBias** (Zhao et al., 2018). Following Guo et al. (2022); Liang et al. (2020), we apply all metrics to measure gender bias, use SEAT to measure racial bias, and use mean average cosine similarity (MAC) (Manzini et al., 2019), a modified SEAT version to measure multi-class religion bias. Specifically, we apply SEAT 6, 6b, 7, 7b, 8, and 8b tests to measure gender bias, and use SEAT 3, 3b, 4, 5, 5b tests for racial bias evaluation. The measure of bias in the SEAT is indicated by its effect size – the closer to 0, the less biased the model is. Following (He et al., 2022), we first fine-tune the model on OntoNotes 5.0 dataset (Hovy et al., 2006), and then evaluate on the coreference resolution task WinoBias, which assesses a system's ability to accurately associate a gendered pronoun to occupations in both pro- and anti-stereotypical scenarios. Coreference is deduced via syntax cues in Type 1 sentences or trickier semantic cues in Type 2. Detailed metrics are provided in Appendix D.

Other Details. To verify debiased PLMs whether still preserve general language understanding, we examine them on six GLUE benchmarks (Wang et al., 2019), including **SST-2**, **CoLA**, **QNLI**, **RTE**, **WNLI**, and **QQP** tasks. We trained CK-Debias in 4 epochs with learning rate $5 \times e^{-5}$ on a single GeForce RTX 3090 GPU, and all results are averaged over 4 runs. Due to space constraints, we include gender results in the main text and provide details on race and religion cases in Appendix G.

4.1 Results on Intrinsic and Extrinsic Metrics

Intrinsic Metrics. As indicated by the remarkable ICAT metric score in Table 1, our CK-Debias strikes a favorable balance between language expressiveness and fairness. Notably, CK-Debias even exhibits a slight improvement in LM metrics compared to the original BERT model, with the score rising from 84.17 to 85.42. For the SEAT value, CK-Debias achieves the best score, and improves 0.075 compared to the SOTA model Auto-Debias. Additionally, CK-Debias outperforms others in CrowS-Pairs with the best score of 50.45 (Stereo score: 51.55, Anti-Stereo score: 49.3).

While CK-Debias does not rank top in terms of the SS value in StereoSet, we note that this metric should be considered alongside LM, rather than evaluated in isolation. For instance, FairFil achieves the highest SS, yet its language modeling capability, as indicated by the lower LM score, suffers a marked decline and trails other methods.

Extrinsic Metrics. CK-Debias and other models achieve similar average F1 scores on OntoNotes, suggesting indistinguishable coreference resolution capabilities. For the evaluation on WinoBias, CK-Debias has a notable improvement compared to all backbones. Specifically on BERT, CK-Debias shows noteworthy advancements in anti- and pro-stereotypical tasks, with an average increase of 2.37% and 0.61% on Type 1 and Type 2 sentences, indicating CK-Debias effectively mitigates the stereotypical token-level associations between occupations and gender. Meanwhile, CK-Debias exhibits the most substantial improvement in fairness, notably lowering average TPR-1 and TPR-2 (a reduction in true positive rates by 6.08). Compared to Auto-Debias, CK-Debias is more effective in mitigating gender bias, with an average improvement of 1.99%, 0.5% across Type 1 and Type 2 sentences, respectively, and notably lowers TPR-1 and TPR-2 by an average of 3.09. However, CK-Debias does not surpass MABEL on several in-

Methods	SEAT	StereoSet			CrowS-P*◇	OntoN*	1A	1P	2A	2P	TPR-1	TPR-2
		LM↑	SS ◇	ICAT ↑								
BERT	0.35	84.17	60.28	66.86	57.25	73.94	55.47	86.7	91.8	96.74	31.22	9.94
+CONTEXT-D*	0.53	85.42	59.35	69.45	58.01	73.76	59.81	84.21	83.63	92.97	23.4	9.62
+FAIRFIL	0.15	44.85	**50.93**	44.01	49.07	71.79	53.24	85.77	77.37	91.40	32.43	14.03
+AUTO-D*	0.14	74.08	52.88	69.81	54.92	73.84	57.04	85.88	91.21	97.54	28.84	6.33
+MABEL	0.582	84.80	56.92	73.07	50.76	73.07	**59.82**	84.21	89.39	95.1	24.39	**5.71**
+CK-D*(Ours)	**0.065**	**85.42**	55.54	**75.96**	**50.45**	**73.98**	59.78	**87.12**	**92.14**	**97.61**	**22.93**	6.07
ALBERT	0.28	90.73	63.58	66.09	56.87	39.98	27.07	**46.21**	33.96	51.69	19.14	17.72
+CONTEXT-D*	0.33	91.02	60.23	72.40	53.91	40.53	18.35	23.61	16.67	33.26	5.26	16.6
+AUTO-D*	0.18	88.43	61.76	67.62	47.86	40.32	22.35	24.83	27.47	36.23	8.61	13.21
+CK-D*(Ours)	**0.15**	**91.32**	**58.93**	**74.93**	**48.07**	**40.79**	**27.21**	43.89	**37.12**	**52.58**	**4.56**	**11.82**
RoBERT	0.67	71.75	53.65	66.50	54.96	40.61	22.02	35.34	9.62	13.21	13.32	**3.59**
+CONTEXT-D*	1.09	70.85	54.74	64.13	59.48	40.67	26.7	37.37	15.38	19.59	10.68	4.21
+AUTO-D*	0.20	69.85	54.21	63.13	49.77	40.53	23.62	**37.74**	17.54	21.71	12.25	5.87
+CK-D*(Ours)	**0.15**	**72.63**	**52.93**	**68.37**	**50.21**	**40.71**	**29.24**	36.54	**19.12**	**24.34**	**9.12**	6.23

Table 1: Gender debiasing results on intrinsic and extrinsic metrics. *: abbreviations for a model or metric. ◇: the closer to 50, the better. **OntoN***, **1A**, **1P**, **2A**, **2P**: the larger, the better. **TPR-1**, **TPR-2**: the smaller, the better.

dicators. This may stem from that MABEL exploits supervised entailment pairs including gendered terms derived from natural language inference data, which involve learned linguistic reasoning abilities crucial for gender-specific coreference resolution tasks. Conversely, our unsupervised sentences have limited inference ability, posing challenges for proficient reasoning in all WinoBias tasks.

4.2 Ablation Study

To verify the effectiveness of CK-Debias, we consider the following ablated version:

- (V1) w/o $\mathcal{L}_{BP}$: Removing the most biased knowledge X^{BP} and corresponding loss L_{BP};
- (V2) w/o $\mathcal{L}_{C}$: Removing the aligned knowledge X^{C} and corresponding loss $\mathcal{L}_{C}$;
- (V3) w/o $\mathcal{L}_{Re}$: Removing the designed representation preserving loss $\mathcal{L}_{Re}$;
- (V4) Rand-1: Replacing KNNs for colliding effect estimation with randomly selected samples;
- (V5) Rand-2: Replacing the most biased prompt with random samples from unaligned knowledge.

As illustrated in Figure 4, all variants are inferior to the full model CK-Debias. The notable performance drop without $\mathcal{L}_{BP}$ suggests that keeping non-colliders (i.e., unaligned knowledge) is beneficial for mitigating model bias. Compared to the full model CK-Debias, removing $\mathcal{L}_{C}$ has a greater decline (both in ICAT and Acc.) than V1, indicating estimating the colliding effect for aligned knowledge acquisition is crucial to ensure positive transfer. The removal of $\mathcal{L}_{Re}$ obviously weakens SST-2 accuracy and the ICAT value, indicating its role in preserving language modeling ability. For Rand-1 and Rand-2 variants, the ICAT score shows

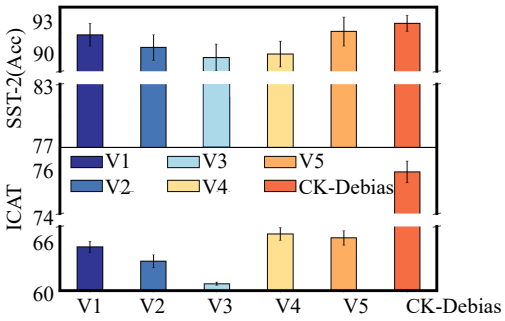
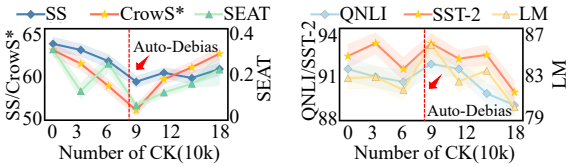


Figure 4: Ablation results. The higher ICAT score and higher SST-2 accuracy are better.

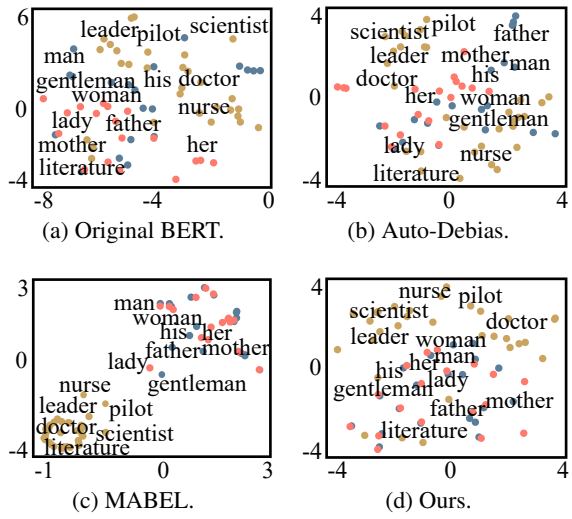


(a) Debiasing effectiveness. (b) Language understanding.

Figure 5: The number impact of the CK sentences.

their performance is inferior to CK-Debias model. Rand-1 results demonstrate the KNNs are crucial for estimating the collision effect, which intentionally transfers aligned commonsense knowledge regarding social diversity from LLM to PLM, ensuring semantic richness. Meanwhile, Rand-2 results emphasize that acquiring the most biased prompt helps the LM mitigate biases by learning to avoid absorbing biased CK.

As the number of CK sentences increases in Fig. 5, CK-Debias improves in expressiveness and fairness, as reflected by the debiasing effectiveness (SS, CrowS-Pairs, SEAT metrics) and language understanding (LM metric and SST-2, QNLI tasks). However, this improvement diminishes beyond 90k sentences, indicating an optimal quantity for debi-



(a) Original BERT. (b) Auto-Debias. (c) MABEL. (d) Ours.

Figure 6: t-SNE plots on BERT.

asing. When comparing all indicators with an equal number of sentences, CK-Debias significantly outperforms Auto-Debias. This difference may stem from the short biased prompts in Auto-Debias lacking syntax or context, highlighting the commonsense knowledge extracted from LLM is a valuable resource, as it provides semantically rich information across diverse demographic groups.

4.3 Results on Language Understanding

Table 2 reports three GLUE results on debiased models (more GLUE results *cf.* Appendix H). Notably, CK-Debias shows slightly superior performance compared to Auto-Debias across CoLA and SST-2 tasks. Our unimpaired downstream task performance highlights that our designed L_{Re} loss effectively tackles the widespread issue of declining language understanding ability found in most debiasing models (He et al., 2022; Liang et al., 2020). However, in the BERT backbone, CK-Debias has a drop result compared to the task-aware SOTA model Causal-Debias over the QNLI task. We caution that Causal-Debias integrates the debiasing process with fine-tuning downstream tasks, while ours operates upstream in a task-agnostic manner. This distinction might pose greater challenges compared to task-aware methods. Surprisingly, our indicators for each task do not decrease compared to the original BERT model, and the Accuracy on SST-2 even shows a significant improvement.

The t-SNE visualization in Fig. 6 explores the debiasing effects and model expressiveness by examining the words' correlation. In Fig. (6d), CK-Debias successfully preserves relative distances between words while pulling attribute words closer to each other. In contrast, Fig. (6b) shows that Auto-Debias clusters male and female words separately, an undesirable behavior indicating that concepts with opposing gender directions are pushed far apart in the hidden space, even when they have significant contextual similarities, thereby introducing biases. In Fig. (6c), MABEL separates target words and gender words, yet the distance between attribute terms and target words is significantly maximal compared to the gaps among attribute words. This may lead to substantial damage to model expressiveness.

Methods	SST-2	CoLA	QNLI
BERT	92.7	57.6	91.3
+AUTO-DEBIAS	92.1	52.1	91.1
+GENDER-TUNING	92.1	56.6	91.3
+CAUSAL-DEBIAS	92.9	58.1	**91.6**
+CK-DEBIAS (ours)	**93.0**	**60.07**	91.4
ALBERT	92.6	58.5	91.3
+AUTO-DEBIAS	94.1	58.3	92.1
+GENDER-TUNING	91.7	58.4	92.1
+CAUSAL-DEBIAS	92.9	57.1	91.6
+CK-DEBIAS (ours)	**94.3**	**58.7**	**92.9**

Table 2: GLUE results over benchmarks.

5 Conclusion

In this paper, we offer a flexible, universally applicable solution CK-Debias capable of debiasing lightweight PLMs by harnessing rich, contextually relevant commonsense knowledge sourced from LLM, unlike existing methods reliant on crafted external corpora. CK-Debias roots in SCM to reveal the limitations of traditional task-agnostic debiasing methods, such as negative knowledge transfer and inaccurate bias identification. This analysis laid the groundwork for our improved causal graph, optimizing the utilization of LLM-generated knowledge by distinguishing aligned knowledge beneficial for positive transfer and unaligned knowledge for strategic bias identification and mitigation. Extensive evaluations show CK-Debias's efficacy in mitigating diverse biases across various PLM architectures, and achieving superior performance in both intrinsic and extrinsic assessments while preserving model expressiveness. Our paper promotes the NLP fairness fields by utilizing LLM-generated knowledge strategically for effective debiasing PLMs. We aim for this study to offer insights into mitigating biases for building fair and accountable NLP systems, hoping to inspire further exploration in other fields using knowledgeable LLM to address practical problems.

6 Limitations

Considering limitations of our debiasing work CK-Debias, we draw inspiration from previous debiasing efforts (Cheng et al., 2021; He et al., 2022; Ghanbarzadeh et al., 2023) and utilize human-collected word lists related to gender, race, and religion to extract commonsense knowledge containing demographic information from LLM. Obviously, human-collected word lists are insufficient to cover all demographic groups related to specific biases. We believe a potential improvement is to use these word lists as a foundation, encouraging the linguistically capable LLM to generate semantically similar words or inspiring the generation of intersectional words (Abbasi et al., 2021) through lexical associations to enrich the existing word lists.

We also noticed that although our work CK-Debias achieves satisfactory results on StereoSet and CrowS-Pairs, there is a weak correlation between their stereotype scores. Taking MABEL (He et al., 2022) as an example, which utilizes SNLI entailment data for training, its stereotype score is the worst on StereoSet, but it is one of the best in CrowS-Pairs. Given that CrowS-Pairs comprises only 266 example pairs, significantly fewer than StereoSet's $2,313$ example pairs, it often serves as a more ambiguous metric. This inconsistency raises concerns about the lack of universality and consistency within the existing evaluations, presenting a fundamental challenge in this field.

Our experimental setup relies on a crucial assumption: the pre-training data size of the LLM is significantly larger, potentially covering the pre-training data of the majority of lightweight PLMs. However, given the unavailability of both pre-training datasets, we employ causal collision effects as a soft constraint to filter out the jumbled data extracted from LLM, aiming to mitigate negative transfer (Zheng et al., 2023). All knowledge acquired from LLM is first directly incorporated into PLM, establishing a connection between the pre-training data of the two models to distinguish data with collision effects for positive transfer. Nonetheless, this approach may not precisely assess the extent of negative transfer attenuation. Therefore, optimizing the alignment between the two models is a potential enhancement for future work.

Moreover, the knowledge we automatically extract is entirely reliant on LLM. We enhance the prompt quality by specifying a range of answer lengths and emphasizing logical, creative, and diverse responses. While these prompt improvements ensure that generated sentences surpass simple structures, enhancing the overall quality of extracted sentences, there is still considerable potential for further enhancement. Currently, we lack a guided evaluation of the LLM's responses, relying solely on mechanical sentence generation without fully harnessing the potential of the LLM as a robust corpus. Additionally, despite using KNN as a soft constraint to filter out sentences unaligned with PLMs, we cannot guarantee the complete absence of bias in the generated sentences.

7 Ethics Statement

For the ethical considerations, it is essential to underscore that our primary contribution centers around methodology. The bias word lists and evaluation metrics utilized in our study are consistent with prior research (Cheng et al., 2021; Zhou et al., 2023). However, owing to their availability constraints, our examination of social biases is confined to binary gender, race, and religion. This simplification might inadvertently perpetuate or reinforce other stereotypes. Binary remains a common challenge in most debiasing methods, and we acknowledge the limitations concerning individuals who identify with third genders, such as transgender, non-binary, etc. We acknowledge the diversity of gender, but due to the limitations of existing word lists and comparison benchmarks, we are constrained to a binary gender. Hence, our research highlights the need for future research to delve into collecting additional attributes regarding more bias diversity or conduct cross-analyses of intersectional biases. During the debiasing process, researchers should pay more attention to alleviating the potential risk of unintended re-propagation.

Another ethical dimension pertains to the fact that current debiasing methods predominantly rely on the English system or high-resource languages, which may inadvertently overlook biases present in various cultural and regional contexts. The application of debiasing methods necessitates an ongoing and thorough evaluation of potential ethical issues to maintain the rationality, impartiality, and social value of research.

Our commitment to ethical practices includes ongoing reflection and consideration of the broader social implications of our work. We are dedicated to fostering inclusivity, diversity, and fairness in AI research and applications.

References

Ahmed Abbasi, David Dobolyi, John P Lalor, Richard G Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3748–3758.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

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 *NIPS*, volume 29.

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

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *ICLR*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. *arXiv preprint arXiv:2307.10522*.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *ACL*, pages 1012–1023.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. Mabel: Attenuating gender bias using textual entailment data. *arXiv:2210.14975*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Mariana Jatobá, Juliana Santos, Ives Gutierriz, Daniela Moscon, Paula Odete Fernandes, and João Paulo Teixeira. 2019. Evolution of artificial intelligence research in human resources. *Procedia Computer Science*, 164:137–142.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv:1906.00742*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *EACL*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *ACL*, pages 5502–5515.

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:1907.11692*.

Jinghui Lu, Linyi Yang, Brian Mac Namee, and Yue Zhang. 2022. A rationale-centric framework for human-in-the-loop machine learning. *arXiv preprint arXiv:2203.12918*.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *NAACL*, pages 615–621.

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *NAACL*, pages 622–628.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *ICLR*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *EMNLP*, pages 1953–1967.

Brady Neal. 2020. Introduction to causal inference. *Course Lecture Notes (draft)*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Moritz Plenz, Juri Opitz, Philipp Heinisch, Philipp Cimiano, and Anette Frank. 2023. Similarity-weighted construction of contextualized commonsense knowledge graphs for knowledge-intense argumentation tasks. *arXiv preprint arXiv:2305.08495*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Jiaan Wang, Jianfeng Qu, Yunlong Liang, Zhixu Li, An Liu, Guanfeng Liu, and Xin Zheng. 2023. Snowman: A million-scale chinese commonsense knowledge graph distilled from foundation model. *arXiv preprint arXiv:2306.10241*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. *NeurIPS*, 30.

Ke Yang, Charles Yu, Yi R Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10780–10788.

Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhiming Ma. 2021. Improved ood generalization via adversarial training and pretraing. In *International Conference on Machine Learning*, pages 11987–11997. PMLR.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*, pages 15–20.

Junhao Zheng, Qianli Ma, Shengjie Qiu, Yue Wu, Peitian Ma, Junlong Liu, Huawen Feng, Xichen Shang, and Haibin Chen. 2023. Preserving commonsense knowledge from pre-trained language models via causal inference. *arXiv preprint arXiv:2306.10790*.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241.

A Designed Prompts and Instances of Generated Commonsense Knowledge Sentences

Initially, we designed a prompt that generates commonsense knowledge at once, as shown below:
P: Please use the word pair (a_i, v_t) *to generate five commonsense sentences simultaneously, then substitute the word* a_j *with* a_i *to generate the corresponding five sentences. The word count of each sentence should not exceed* 25.

In practice, we observed that with a more intricate prompt, the LLM tends to focus on a specific segment of the prompt. The resulting single commonsense sentence includes either three words from the triple (a_i, a_j, v_t) or incorporates only the sole word from the triple. This phenomenon became notably apparent when generating a large number of bundle samples simultaneously. Furthermore, the generation of a substantial quantity of bundle samples with similar components is characterized by high instability. However, we anticipate the content in bundle samples to exhibit maximum consistency, excluding placeholders (attribute/target words). In this manner, we aim to mitigate the impact on length inconsistency, other component words (excluding placeholders), and semantics, among bundle samples. Consequently, it narrows relative distance between attribute words and target words, thereby preserving fairness.

Hence, we opt for a two-step prompting method. By introducing two simplified, comprehensible, and clear prompts P_1 and P_2 (depicted in Figure 7), we prevent the LLM from deviating from instructions and excessive imagination, ensuring better alignment with our specified requirements. $a_i \in \{a_1, a_2, \cdots, a_d\}$, and a_j used for replacement comes from other $d-1$ elements, and $v_t \in \mathcal{W}_t$. The generated CK sentences, containing $(a_i, \underline{v_t})$, $(a_j, \underline{v_t})$, shape a *bundle* sample with identical target words $\underline{v_t}$ over d-tuple attribute words. The prompts are enhanced by imposing the answer length and emphasizing logical, creative, and diverse responses, ensuring that the generated sentences go beyond simple structures and improve the overall quality of the extracted sentences.

In Table 3, we present the resulting bundle samples $x^{(n)}$ produced by GPT-3.5-turbo through our crafted prompts P_1 and P_2, which contain our specifilized pairwise attribute and target words $(a_i, \underline{v_t})$, $(a_j, \underline{v_t})$. Also, their corresponding KNNs are provided in Table 3.

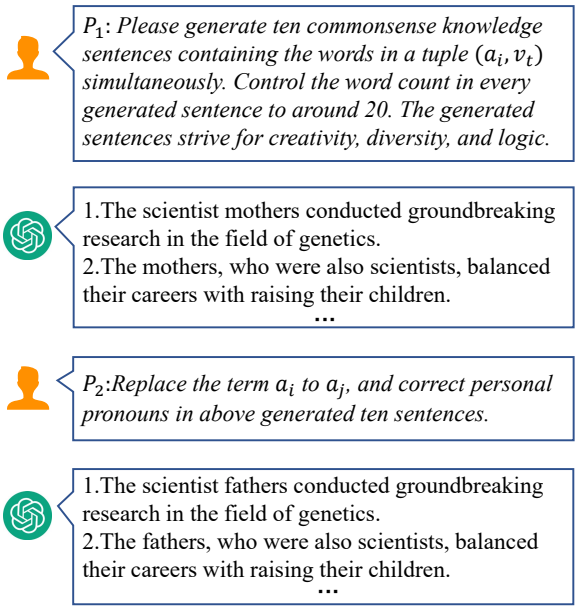


Figure 7: A two-step prompting for generating commonsense knowledge.

B More Derivation Details and Algorithm Process

$R_i^{x_i^{(n)}}$ measures the relative distance from sentence $x_i^{(n)}$ with attribute words $a(i)$ to those sentences with all target words in $\mathcal{W}_t$, which is defined as:

$$\begin{aligned} R^{x_i^{(n)}} &= Distance(E^{target}|e^{x_i^{(n)}}) \\ &= [p_{v_1|a(i)}, p_{v_2|a(i)}, ...] \end{aligned} \tag{4}$$

$$p_{v_j|a(i)} = \frac{\exp(-\frac{||e^{a(i)}-e^{v_j}||^2}{2\rho^2})}{\sum_{v_k \in \mathcal{W}_t}\{\exp(-\frac{||e^{a(i)}-e^{v_k}||^2}{2\rho^2})\}} \tag{5}$$

where each $p_{v_j|a(i)}$ in $R_i^{x_i^{(n)}}$ serves as a normalization factor, quantifying the extent to which information from attribute $a(i)$ can be recovered from target word v_j in $\mathcal{M}_H$, and ρ is a hyperparameter.

The Jensen–Shannon divergence used in $\mathcal{D}_1$ and $\mathcal{D}_2$ of Section 3.4 is defined as:

$$\begin{aligned} &JS\left(p_1, p_2, \ldots, p_m\right) \\ =&\frac{1}{m}\sum_i KLD\left(p_i \| \frac{p_1+p_2+\ldots+p_m}{m}\right) \end{aligned} \tag{6}$$

where p_1, p_2, ..., p_m are distributions, the Kullback–Leibler divergence (KLD) between two distributions p_i, p_j is calculated as $KLD\left(p_i \| p_j\right) = \sum_{v \in \mathcal{V}} p_i(v) \log\left(\frac{p_i(v)}{p_j(v)}\right)$, and v is the vocabulary from the PLM. Moreover, the detailed algorithm process of CK-Debias is provided in Algorithm 1.

	(fraternal, soldiers), (sororal, soldiers)	(african people, beauty), (caucasian people, beauty)
Bundle samples	P_1: In times of war, **soldiers** rely on their **fraternal** instincts to protect and defend their comrades, displaying unwavering bravery and selflessness on the front lines. P_2: In times of war, **soldiers** rely on their **sororal** instincts to protect and defend their comrades, displaying unwavering bravery and selflessness on the front lines.	P_1: **African people** appreciate the **beauty** of traditional attire, and romantic occasions often involve dressing in traditional garments that symbolize cultural pride and heritage. P_2: **Caucasian people** appreciate the **beauty** of traditional attire, and romantic occasions often involve dressing in traditional garments that symbolize cultural pride and heritage.
KNNs	P_1: During times of war, **soldiers** display unwavering loyalty to their **countrymen**, willingly sacrificing their own safety and comfort, while remaining submissive to the chain of command. P_2: During times of war, **soldiers** display unwavering loyalty to their **countrywomen**, willingly sacrificing their own safety and comfort, while remaining submissive to the chain of command.	P_1: **Sewing** has become a means of social cohesion for **african people**, as they come together to sew quilts and blankets for those in need, promoting solidarity and compassion. P_2: **Sewing** has become a means of social cohesion for **white people**, as they come together to sew quilts and blankets for those in need, promoting solidarity and compassion.
	P_1: Throughout history, **men** have been drawn to the allure of becoming **warriors**, seeking glory on the battlefield and proving their valor in the face of adversity. P_2: Throughout history, **women** have been drawn to the allure of becoming **warriors**, seeking glory on the battlefield and proving their valor in the face of adversity.	P_1: The **owner** of a cultural exchange program facilitates meaningful interactions between **african people** and individuals from different backgrounds, fostering understanding and appreciation. P_2: The **owner** of a cultural exchange program facilitates meaningful interactions between **caucasian people** and individuals from different backgrounds, fostering understanding and appreciation.
	P_1: **Men** who excel as **snipers** possess a unique blend of patience, discipline, and tactical thinking, enabling them to adapt to various scenarios and overcome challenges. P_2: **Women** who excel as **snipers** possess a unique blend of patience, discipline, and tactical thinking, enabling them to adapt to various scenarios and overcome challenges.	P_1: The sergeant, a **leader** with vision, understands that diversity strengthens the military and actively promotes the inclusion of **black people** in all aspects of military life. P_2: The sergeant, a **leader** with vision, understands that diversity strengthens the military and actively promotes the inclusion of **white people** in all aspects of military life.
	P_1: Throughout history, great **leaders** have inspired their **countrymen** to be both patriotic and submissive, fostering a sense of collective responsibility towards the well-being of their nation. P_2: Throughout history, great **leaders** have inspired their **countrywomen** to be both patriotic and submissive, fostering a sense of collective responsibility towards the well-being of their nation.	P_1: In the world of aviation, **black people** have excelled as **pilots** and skippers, breaking barriers and inspiring future generations. P_2: In the world of aviation, **white people** have excelled as **pilots** and skippers, breaking barriers and inspiring future generations.
	P_1: In times of crisis, **heros** emerge from all walks of life, including doctors, **firefighters**, and police officers, who selflessly put their lives on the line to protect and serve their communities. P_2: In times of crisis, **heroines** emerge from all walks of life, including doctors, **firefighters**, and police officers, who selflessly put their lives on the line to protect and serve their communities.	P_1: The sergeant, a role model for all **soldiers**, recognizes the unique challenges faced by **black people** and provides support and guidance to help them overcome obstacles. P_2: The sergeant, a role model for all **soldiers**, recognizes the unique challenges faced by **white people** and provides support and guidance to help them overcome obstacles.

Table 3: The instances of bundle samples and their KNNs over gender, and race cases. The 2-th row represents the bundle sample, while the rest rows represent their corresponding KNNs.

Bias Type	Test	Demographic-specific words	Stereotype words
Racial	SEAT-3	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-3b	European-American/African American terms	Pleasant vs. Unpleasant
	SEAT-4	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-5	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-5b	European-American/African American terms	Pleasant vs. Unpleasant
Gender	SEAT-6	Male vs. Female names	Career vs. Family
	SEAT-6b	Male vs. Female terms	Career vs. Family
	SEAT-7	Male vs. Female terms	Math vs. Arts
	SEAT-7b	Male vs. Female names	Math vs. Arts
	SEAT-8	Male vs. Female names	Science vs. Arts
	SEAT-9b	Male vs. Female terms	Science vs. Arts

Table 4: The SEAT test details extended from (Caliskan et al., 2017).

Binary Gender	Binary Race	Multiclass Religion
countrywoman, countryman	africa, europe	muslim, jewish, christian
heroine, hero	black, white	muslims, jews, christians
mothers, fathers	africa, america	quran, torah, bible
her, him	black people, white people	mosque, synagogue, church
hostess, host	african, american	imam, rabbi, priest

Table 5: Examples of word pairs to estimate the three types of biases.

Detailed experimental setup. We trained the gender debiasing process in 6 hours (4 epochs), 2 hours for the racial case, and 3 hours for the religion case.

Here, we provide the sources for the results in Table 1. For the reported SEAT scores, MABEL is derived from its original paper (He et al., 2022). Results for Context-Debias and Auto-Debias on Bert, Albert, and Roberta, as well as FairFil on Bert, are provided by Auto-Debias (Guo et al., 2022). For the reported StereoSet results, Context-Debias, FairFil, and MABEL on Bert are provided by MABEL (He et al., 2022). Results for Auto-Debias on Bert, Albert, and Roberta backbones, as well as Context-Debias on Albert and Roberta, are obtained through our testing. For CrowS-Pairs, Context-Debias, FairFil, and MABEL on Bert are provided by MABEL (He et al., 2022). Results for Auto-Debias on Bert, Albert, and Roberta are sourced from its original paper (Guo et al., 2022), while results for Context-Debias on Albert and Roberta are obtained via our testing based on the parameters provided in their paper. All extrinsic metrics are obtained through our testing. Note that all Bert results are based on the `bert-base-uncased` version, thus differing from the results reported in MABEL (`bert-base-cased`).

C Bias Words List

We used the gender/race/religion attribute and target words lists proposed in (Kaneko and Bollegala, 2021), (Manzini et al., 2019), and (Liang et al., 2020), respectively, which is widely used in debiasing studies (Guo et al., 2022; Yang et al., 2023). Examples of word pairs are provided in Table 5.

D Metrics Details

The WEAT metric measures the bias by comparing two sets of attribute words W_a (i.e., M and F) and two sets of target words W_t (i.e., A and B). In the case of gender, M denotes masculine words like "he", and F denotes feminine words like "she". Meanwhile, A and B are gender-neutral words (e.g., career or adjectives) whose embeddings should be equivalent between M and F. Formally, bias degree of each word w is defined as:

$$s(w, A, B) = \frac{1}{|A|}\sum_{a\in A}\cos(w, a) - \frac{1}{|B|}\sum_{b\in B}\cos(w, b), \quad (7)$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity. Based on Equation (7), the WEAT effect size is:

$$d_{\text{WEAT}} = \frac{\mu(\{s(m, A, B)\}_{m\in M}) - \mu(\{s(f, A, B)\}_{f\in F})}{\sigma(\{s(t, A, B)\}_{t\in A\cup B})}, \quad (8)$$

where μ and σ denote the mean and standard deviation, respectively. The SEAT metric generalizes the WEAT via replacing the word embeddings with a few simple sentence templates (e.g., "This is the <word>"). We can conclude from Equation (8) that the absolute SEAT effect size closer to 0 means lower biases. We list more details about the SEAT tests that are used in our experiments in Table 4, which are adapted from (Caliskan et al., 2017).

CrowS-Pairs contains sentence pairs regarding stereotype/anti-stereotype but with semantics closest to each other, and its score closer to 50% is less stereotypical, indicating that the model assigns an equal probability to male and female sentences.

StereoSet assesses a language model's expressiveness and biases using cloze tests, selecting stereotypical, anti-stereotypical, and unrelated words, including three metrics: Language Modeling Score (LM), indicating expressiveness by word relevancy frequency (higher scores signify better performance); Stereotype Score (SS), measuring bias by the frequency of selecting stereotypical words (scores near 50 indicate less bias). The Idealized Context Association Test (ICAT) combines LM and SS, providing a comprehensive metric where a perfect score of 100 denotes high expressiveness with minimal bias.

WinoBias (Zhao et al., 2018) assesses intra-sentence coreference resolution by examining a system's ability to accurately link a gendered pronoun to an occupation within both pro- and anti-stereotypical contexts. Coreference resolution involves identifying connections based on syntactic

cues in Type 1 sentences and more complex semantic cues in Type 2 sentences. Our approach involves initial model fine-tuning on the OntoNotes 5.0 dataset (Hovy et al., 2006) followed by evaluation using WinoBias benchmark. We present average F1-scores on OntoNotes and for pro- and anti-stereotypical instances, along with the true positive rate difference in average F1-scores across Type 1 and Type 2 examples. The metrics **1** = Type 1; **2** = Type 2, **A** = Anti-stereotypical; **P** = Pro-stereotypical; **TPR** = Ture Positive Rate.

E How to Apply KNNs to Estimate the Causal Effect?

Our rationale for aligning the commonsense extracted from LLM with PLM involves matching KNNs based on the PLM's parameters. In the pre-training stage of the BERT model, `[CLS]` token embedding is employed as an aggregate representation of the entire sentence, which is expected to distinguish between different sentences and to capture crucial high-level semantics. Therefore, for the KNNs process, we first acquire `[CLS]` token embedding of all generated commonsense knowledge sentence samples from pre-trained BERT model. Subsequently, we calculate the similarity between each pair of samples based on their `[CLS]` token embedding. Finally, we identify the K nearest neighbors and obtain the corresponding K similarity scores $\{S_k\}_{k=1}^{K}$ for each CK sample.

Pre-defined threshold. However, if the similarity score S_k of a specific anchor sample does not exceed a pre-defined threshold θ, we filter the k-th neighbor of this sample, as this indicates a perceived absence of collision effects with the pre-training data of the BERT model. In contrast, we retain the remaining neighbors, as they are deemed to exhibit a significant correlation with the BERT pre-training data. This filtering means the greater the similarity scores $\{S_k\}_{k=1}^{K}$, the more similar samples satisfy the threshold, and the more aligned commonsense knowledge shared by BERT models and LLM (e.g., ChatGPT). In the end, we estimate the colliding effect Φ between BERT's pre-trained data P and aligned CK X_C with the joint prediction of KNN samples. In practical implementation, we set K to 5, and θ serves as a hyper-parameter (we experimented with values $\{215, 220, 225, 227\}$), guiding the partition between X_C with colliding effect and X_{NC} without colliding effect.

Accelerating similarity calculation. Due to the large volume of CK sentences extracted from LLM, the traditional method (i.e., cosine similarity or Euclidean distance) for calculating the similarity matrix can be time-consuming. To tackle this issue, we utilize the Faiss (Facebook AI Similarity Search), a library designed for efficient similarity search and clustering of dense vectors. Faiss is instrumental in accelerating the process and managing large-scale vector datasets effectively, which retrieves top-K similar vectors from large-scale vector datasets by constructing an index for the base vector data. In our experiment, the top-K matrix computation for 100k sentence embeddings is completed in just six minutes.

F Example of the Most Biased Prompts

We find the most biased prompts X^{BP} from X^{NC}, and devise the specific loss function to enhance the bias avoidance capability of PLMs, which resemble adversarial training (Yi et al., 2021). In our practical implementation, we first compute the disagreement of target words v_t (`[MASK]`) over each bundle sample $x^{(n)}$ regarding different demographic groups a_i, a_j, and then sort all disagreements (i.e. probability) in unaligned knowledge sentences in descending order. Finally, we select the top 0.55 of the prompts as the most biased prompts. Some most biased prompts X^{BP} are provided in Table 9. Note that for enhancing the bias avoidance capacity, we initially tried to directly extract biased knowledge from LLMs, given that LLMs prevent harmful text generation, we shifted to this alternative strategy, which also better utilizes the generated commonsense knowledge. It is worth noting that we cannot observe bias in the semantics from the provided examples, as the bias is implicit and measured by probability within the model itself.

G The Results of Race and Religion Cases

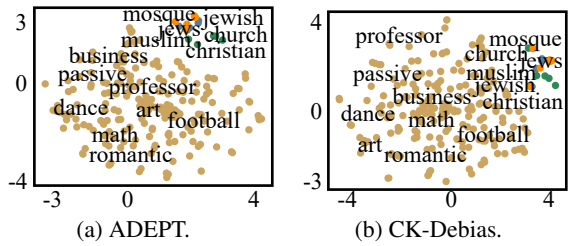
We consistently achieve debiasing results and maintain language modeling capability in both racial and religious cases, similar to the gender case discussed in our main text (*cf.* Section 4).

For the racial case, we employ backbones including BERT, and ALBERT to examine our CK-Debias debiasing framework. Following prior works (Guo et al., 2022; Zhou et al., 2023), we report the degree of racial bias in the debiased models across SEAT 3, 3b, 4, 5, and 5b. The racial SEAT test examines associations between European-American/African-American

names/terms and stereotype words (pleasant vs. unpleasant). The results in Table 6 show that the debiasing effect of CK-Debias surpasses that of Auto-Debias, yielding superior results. Specifically, in the BERT backbone, CK-Debias successfully mitigates racial bias in 4 out of 5 SEAT sub-tests, resulting in a notable decrease in the average score from 0.23 to 0.16. Remarkably, CK-Debias significantly reduces bias across all SEAT in ALBERT.

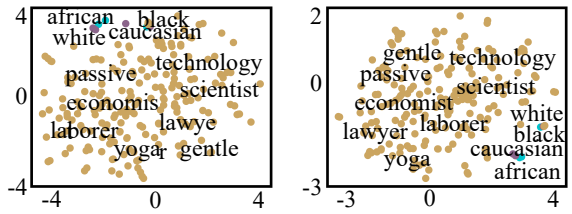
For the religious case, following previous works (Liang et al., 2020; Yang et al., 2023), we perform CK-Debias in BERT backbone and use the MAC score to evaluate the degree of religious bias in the debiased model. The results reported in Table 7 demonstrate the effectiveness of CK-Debias in mitigating religious bias, as the MAC score has an increase of 0.04 compared to Sent-Debias.

We also explore the debiasing effects and model expressiveness by examining the correlations in religious/racial vocabulary, as shown by the t-SNE visualization in Fig. 8 and Fig. 9, respectively. From



(a) ADEPT. (b) CK-Debias.

Figure 8: t-SNE visualization in the religion domain.



(a) Auto-Debias. (b) CK-Debias.

Figure 9: t-SNE visualization in the racial domain.

Fig. (8b), we can conclude that CK-Debias effectively preserves the relative distances between words while narrowing the gap between the triple attribute words represented by Islam, Judaism, and Christianity. In contrast, as depicted in Fig. (8a), ADEPT merely minimizes the distance between Islam and Judaism. Furthermore, from Fig. (9a) and Fig. (9b), we observe that CK-Debias effectively reduces the distance between oppositional attribute words (e.g., Black and White people), surpassing the performance of Auto-Debias.

H More GLUE Results

Due to space constraint, we only report three GLUE tasks in Section 4, including sentiment classification task **SST-2**, grammatical acceptability judgment task **CoLA**, question-answering task **QNLI**. The additional GLUE tasks are reported in Table 8, including **RTE** (Recognizing Textual Entailment) task, determining whether the hypothesis sentence can be inferred from the premise sentence or not; **WNLI** (Winograd Schema Challenge - Pronoun Disambiguation), resolving pronoun references in sentences by understanding the context; **QQP** (Quora Question Pairs), determining whether question pairs are semantically equivalent or not. We can observe CK-Debias outperforms other methods in most cases, indicating the preservation of language understanding.

Methods	SEAT-3	SEAT-3b	SEAT-4
BERT	-0.10	0.37	0.21
+AUTO-DEBIAS	0.25	0.19	0.12
+CK-DEBIAS (ours)	**0.17**	**0.18**	**0.07**
ALBERT	0.60	0.29	0.53
+AUTO-DEBIAS	**0.10**	**0.12**	0.19
+CK-DEBIAS (ours)	0.12	0.14	**0.08**
	SEAT-5	**SEAT-5b**	**Avg.**
BERT	0.16	0.34	0.23
+AUTO-DEBIAS	**0.15**	0.17	0.18
+CK-DEBIAS (ours)	0.24	**0.14**	**0.16**
ALBERT	0.40	0.46	0.46
+AUTO-DEBIAS	0.26	0.19	0.17
+CK-DEBIAS (ours)	**0.21**	**0.14**	**0.14**

Table 6: Race debiasing performance on SEAT.

Model	MAC
BERT	0.035
+Sent-Debias	0.37
+CK-Debias(ours)	**0.41**

Table 7: Religion debiasing results on MAC metric (ranging from 0 to 2, closer to 1 indicates lower bias).

Methods	RTE	WNLI	QQP
BERT	58.1	55.1	90.2
+AUTO-DEBIAS	60.2	56.1	91.1
+CAUSAL-DEBIAS	62.5	55.4	**91.5**
+CK-DEBIAS (ours)	**64.3**	**57.7**	90.8
ALBERT	74.4	55.2	91.1
+AUTO-DEBIAS	75.1	58.3	**92.1**
+CAUSAL-DEBIAS	74.6	57.6	91.5
+CK-DEBIAS (ours)	**75.6**	**58.8**	91.8

Table 8: More GLUE results over benchmarks.

Attribute words	Most Biased Prompts Examples
(lasses, lads)	During a neighborhood watch meeting, the **A** discussed ways to improve the relationship between the community and the [MASK], aiming for a safer and more harmonious environment.
(businesswoman, businessman)	The visionary **A** and [MASK] recognized the potential of renewable energy sources early on, investing heavily in solar and wind power projects that not only generated substantial profits but also contributed to a greener future.
(motherhood, fatherhood)	The joy of **A** can be compared to the satisfaction a [MASK] feels when solving a difficult equation, as both experiences require perseverance and a sense of accomplishment.
(woman, man)	When traveling, a **A** should pack versatile [MASK] that can be easily mixed and matched, allowing him to create various outfits with minimal luggage, ensuring both practicality and style.
(cow, bull)	The [MASK]'s acrobatic flips and jumps mesmerized the **A**, momentarily distracting it from its aggressive nature.

Table 9: Examples of most biased prompts X^{BP} in gender case. **A** represents one element of gender attribute words from the first column.

Algorithm 1 CK-Debias.

Require: Pre-trained language model $\mathcal{M}$, n-steps prompts P_i , attribute word tuples $\mathcal{W}_a = \{(a_1^{(1)}, a_2^{(1)}, \cdots, (a_d^{(1)}), a_1^{(2)}, a_2^{(2)}, \cdots, a_d^{(2)}), \cdots\}$, target words $\mathcal{W}_t = \{v_1, v_2, \cdots\}$.
Ensure: Debiased Language Model $\mathcal{M}'$;
Bias-related sentences from LLM: $X^{(1)}$, containing $(a_i^{(1)}, \underline{v_t})$;
Corresponding bias-related sentences with specific attribute words replacement from LLM: $X^{(2)}$, containing $(a_j^{(2)}, \underline{v_t})$;
Compute the KNNs for each sample $x^{(i)}$ with $\mathcal{M}$;
for $x^{(i,j)}$ in $neighbor$ **do**
    if $similarity(x^{(i)}, x^{(i,j)}) \geq \theta$ **then**
        append $x^{(i,j)}$ to X_C;
    else
        continue
    end if
end for
Divide the data X into X_C with Colliding Effect and X_{NC} without Colliding Effect
Search for the most biased prompt $X_{BP} \leftarrow \text{Top}_{ratio}\left\{\text{JSD}(p^{x_i^{(n)}} || p^{x_j^{(n)}}), n \in \{1, 2, \ldots\}\right\}$
for Epoch in 1, 2, $\cdots$ **do**
    Compute debiasing loss L_{Bias} with data X_C and data X_{BP} according to Eq. (2);
    Compute representation loss $\mathcal{L}_{Re}$ according to Eq. (3);
    Compute overall training loss $\mathcal{L}$ according to Eq. (4);
    Compute gradient;
    Debias language model.
end for
return Debiased Language Model $\mathcal{M}'$.