Self-Supervised Position Debiasing for Large Language Models

Zhongkun Liu[1], **Zheng Chen**[1], **Mengqi Zhang**[1], **Zhaochun Ren**[2], **Zhumin Chen**[1]*, **Pengjie Ren**[1]*
[1]School of Computer Science and Technology, Shandong University, China
[2] Leiden University, Leiden, the Netherlands
{liuzhongkun,chenzheng01}@mail.sdu.edu.cn,
{mengqi.zhang,chenzhumin,renpengjie}@sdu.edu.cn,
z.ren@liacs.leidenuniv.nl

Abstract

Fine-tuning has been demonstrated to be an effective method to improve the domain performance of large language models (LLMs). However, LLMs might fit the dataset bias and shortcuts for prediction, leading to poor generation performance. Previous works have proven that LLMs are prone to exhibit position bias, i.e., leveraging information positioned at the beginning or end, or specific positional cues within the input. Existing debiasing methods for LLMs require external bias knowledge or annotated non-biased samples, which is lacking for position debiasing and impractical in reality. In this work, we propose a self-supervised position debiasing (SOD) framework to mitigate position bias for LLMs. SOD leverages unsupervised responses from pre-trained LLMs for debiasing without relying on any external knowledge. To improve the quality of unsupervised responses, we propose an objective alignment (OAM) module to prune these responses. Experiments on eight datasets and five tasks show that SOD consistently outperforms existing methods in mitigating three types of position biases. Besides, SOD achieves this by sacrificing only a small performance on biased samples, which is general and effective. To facilitate the reproducibility of the results, we share the code of all methods and datasets on https://github.com/LZKSKY/SOD.

1 Introduction

Although large language models (LLMs) have demonstrated remarkable unsupervised ability in various tasks (Kojima et al., 2022), fine-tuning still overtakes it under the task-specific setting (Ding et al., 2023). However, fine-tuned LLMs might rely on the dataset biases and artifacts as shortcuts for prediction, as the fine-tuning datasets are sometimes skewed due to budget constraints (Du et al., 2022). This results in poor generalization performance when applying fine-tuned LLMs to unseen test data and these models are vulnerable to various types of adversarial attacks (Meade et al., 2022).

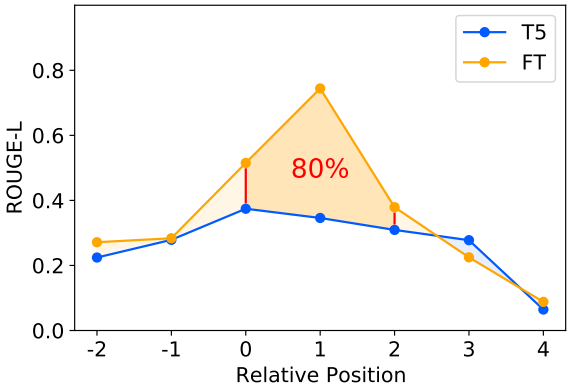


Figure 1: Question answering performance of FlanT5-large (T5) and fine-tuned FlanT5-large (FT) over different relative positions in CANARD. Relative position means the distance of grounded utterances between the last turn answer and the current turn answer.

Position bias has been demonstrated to exist across various fine-tuned LLMs (Liu et al., 2023). Specially, the well-known LLMs, e.g., GPT-3.5[1], longchat-13B[2], are skilled when the relevant information occurs at the beginning or end of the input context, while the performance significantly degrades when LLMs need to find relevant information in the middle of the context. Analysis of conversational question answering (CQA) on CANARD (Elgohary et al., 2019) dataset further confirms the existence of position bias. As shown in Fig. 1, 80% of the performance improvement after fine-tuning is attributed to fitting bias on relative positions 0-2. This encourages researchers to engage in position debiasing (Meade et al., 2022).

Early works mainly focus on mitigating position bias on extractive tasks before the emergence of LLMs (Ko et al., 2020; Karimi Mahabadi et al.,

* Corresponding authors.

[1]https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates
[2]https://lmsys.org/blog/2023-06-29-longchat

Findings of the Association for Computational Linguistics ACL 2024, pages 2897–2917
August 11-16, 2024

2020). A prominent debiasing method is Product-of-Expert (PoE), which discourages the extractive model from learning position bias picked up by the fixed biased model (Du et al., 2021; Shinoda et al., 2022). Recently, more and more works focus on debiasing for generative models, e.g., LLMs (Guo et al., 2022; Li et al., 2023). They adopt either in-context learning (ICL) to guide the generation of LLM (Meade et al., 2023) or prompt tuning (PT) to fine-tune prompts for LLMs (Li et al., 2023). However, these works mostly focus on mitigating social bias (Kasneci et al., 2023), e.g., gender bias and racial bias, leaving position debiasing unexplored. Besides, these works cannot be transferred to mitigate position bias, as they require manually annotated non-biased samples for ICL or external bias knowledge for PT, which are lacking for position debiasing.

To deal with this challenge, we propose to leverage the low position bias characteristics of pre-trained LLMs. Previous studies have shown that pre-trained LLMs are more robust to position bias (Utama et al., 2021). This is due to the randomness of knowledge utilization in generation during pre-training. As shown in Fig. 1, the ROUGE-L score of pre-trained T5 fluctuates within the range of 0.2 to 0.4 across almost all relative positions, demonstrating its robustness against position bias.

In this paper, we propose a self-supervised position debiasing (SOD) framework to mitigate position bias for LLMs. First, we use a low-bias inference module to collect unsupervised responses with low position bias by applying various prompting strategies. Then, we propose an objective alignment (OAM) module to prune the unsupervised responses, as low-quality responses will undermine model performance on non-biased samples. Finally, we use a multi-objective optimization module to leverage these unsupervised responses for fine-tuning. The whole process does not require any external bias knowledge or non-biased samples, which is self-supervised and general.

To verify the effectiveness of SOD, we conduct experiments on eight datasets covering five tasks. Experimental results show that the SOD achieves superior performance in mitigating three types of position bias significantly, including lead bias, relative position bias, and lexical bias. The main contributions of this work are as follows.

- We propose to mitigate position bias for LLMs in a self-supervised setting, i.e., without any external knowledge or annotated samples.
- We propose a SOD framework for position debiasing with an OAM module to prune low-quality unsupervised responses for fine-tuning.
- Experiments show that SOD can mitigate various types of position biases by sacrificing only small performance on biased samples, demonstrating its effectiveness and generality.

2 Preliminary

2.1 Task Definition

Given a biased dataset D for the target task, our position debiasing task aims to improve the model robustness against position bias when fine-tuning on target tasks, i.e., to achieve superior performance on non-biased samples by retaining the performance on biased samples. Here, the biased samples exhibit similar position bias as training samples, and the non-biased samples do not contain these position clues. The target task for fine-tuning can be any natural language processing (NLP) tasks exhibiting position bias. In this paper, we focus on five target tasks: conversational question answering (CQA), conversational question generation (CQG), knowledge-based conversation (KGC), summarization and natural language inference (NLI).

2.2 Large Language Model

LLMs have attracted much attention and become state-of-the-art due to their remarkable ability of language generation. They formulate all NLP tasks as language generation tasks with different task prompts:

$$\begin{aligned} p(y) &= p(y|prompt, x) \\ &= \prod_t p(y_{t+1}|prompt, x, y_1, y_2, ..., y_t), \end{aligned} \quad (1)$$

where x, y and $prompt$ are task input, output, and task prompt, respectively. y_i denotes the i-th token in y. Task prompt $prompt$ consists of the task instruction and demonstrations to tell the LLMs the definition of the task and how it works. The introduction of task prompt enables LLMs to utilize all available data for training and improve their generalization ability on unseen tasks.

3 Method

We propose a self-supervised position debiasing (SOD) framework to mitigate position bias for generative LLMs. As shown in Fig. 2, SOD consists

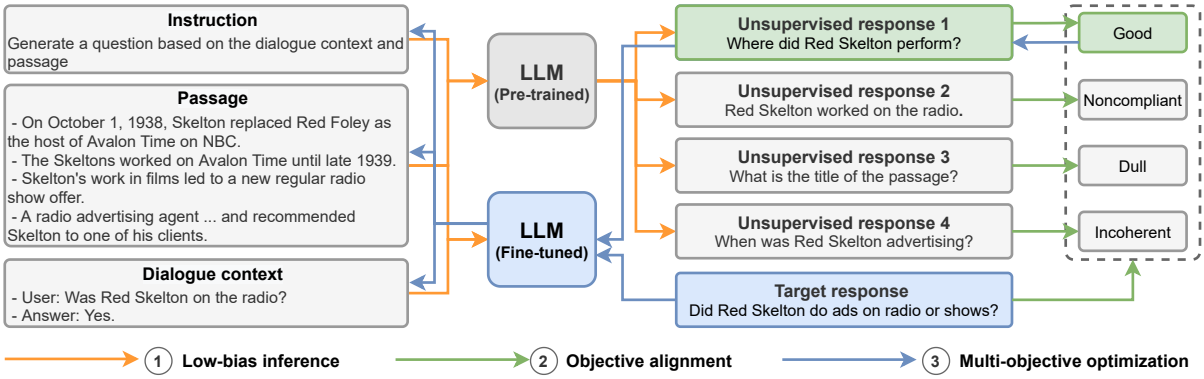


Figure 2: Overview of our proposed self-supervised position debiasing (SOD) framework (taking CQG as the example). First, the low-bias inference module collects multiple unsupervised questions from LLMs. Then, the objective alignment module aligns these questions with the target question. Finally, these aligned questions are utilized for fine-tuning within the multi-objective optimization module.

of three modules: low-bias inference, objective alignment (OAM), and multi-objective optimization, where all modules do not require external bias knowledge or non-biased samples. The Low-bias inference module generates unsupervised responses with lower position bias by utilizing pre-trained LLMs (in §3.1). Subsequently, the OAM module is employed to prune these low-quality unsupervised responses based on the target responses (in §3.2). Finally, the multi-objective optimization module fine-tunes the LLMs by optimizing the task objective and the debiasing objective (in §3.3). The task objective utilizes target responses to enhance task-specific performance, and the debiasing objective leverages unsupervised responses for position debiasing.

3.1 Low-bias Inference

Given a biased training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ for the target task, the low-bias inference module generates unsupervised responses with low position bias based on pre-trained LLMs (Utama et al., 2021).

We employ three prompting strategies for generation and adapt them for different target tasks.

- **Instruction-only prompting** generates responses of target task by feeding the task input and task instruction to the pre-trained LLMs. Concretely, we assign the *prompt* by task instruction in Eq. 1 for the generation.
- **Diverse prompting** generates responses with diverse aspects by feeding various prompts to LLMs.
- **In-context learning (ICL)** also feeds multiple input-output examples to LLMs for generation, in addition to the task instruction and input. It enhances the model comprehension of target tasks but requires a longer input length.

We adopt ICL only for NLI, due to the limit of the model input length. We employ diverse prompting for CQG, which is intrinsically creative and diverse. And instruction-only prompting is implemented for CQA, KGC, and summarization.

3.2 Objective Alignment

Unlike the annotated high-quality target responses for the task objective, the unsupervised responses for the debiasing objective are of lower quality and noisy, because they are generated by pre-trained models without specific fine-tuning on the target task. To reduce the interference between the debiasing objective and the task objective, we propose an objective alignment (OAM) module to prune the unsupervised responses y_i' to better align with the target response y_i.

We propose various alignment strategies for different tasks based on their intrinsic characteristics.

Alignment for tasks excluding NLI. We first identify low-quality unsupervised responses and then drop them, as the generated responses are flexible for modification and modification may introduce new errors. We apply and combine four strategies to identify them.

- **Non-compliant identification** identifies unsupervised responses deviating from the task instruction by keyword matching. For example, it identifies non-'what' questions when 'what' is specified in the instruction.
- **Dull identification** identifies dull responses by keyword matching, e.g., "What is the title of the passage?".

- **Incoherent identification** identifies incoherent responses if the perplexity of any token in the response falls below a pre-defined threshold.
- **Unreliable identification** identifies unreliable responses if the overlap score between unsupervised and target responses is less than a pre-defined threshold. The intuition is that the fact in the response may be wrong if its semantics deviate significantly from the fact in the reference.

For CQA, summarization, and KGC tasks, we align unreliable responses since the facts in answers, summaries, and knowledge-enriched responses are always unique in their semantics. For CQG task, we align non-compliant, dull, or incoherent responses, considering that the appropriate questions are diverse in semantics.

Alignment for NLI. As the generated responses for NLI are deterministic, i.e., entailment, neutral, and contradiction, we can directly estimate the probability distribution over all classes by prompting and then align it. The estimated probability distributions are low-quality sometimes when the target class dominates, which is redundant for optimizing the task objective and strengthens position bias. Therefore, we align the estimated probability distribution by masking the target class:

$$y_i' = s_i \cdot mask_i, \tag{2}$$

where $mask_i$ is a vector to mask the target class:

$$mask_{i,j} = \begin{cases} 0, \text{ if } class_j = y_i \\ 1, \text{otherwise.} \end{cases} \tag{3}$$

Here, $class_j$ is the tokens for j-th class and y_i is the tokens for the target class. s_i is the probabilities distribution over all classes in NLI inference:

$$\begin{aligned} s_i &= [s_{i,1}, s_{i,2}, ..., s_{i,|class|}] \\ s_{i,j} &= \frac{p(class_j|x_i)}{\sum_j^{|class|} p(class_j|x_i)}, \end{aligned} \tag{4}$$

where $p(class_j|x_i)$ is the generation probabilities of j-th class tokens, which is calculated by Eq. 1.

3.3 Multi-Objective Optimization

Given the target response y_i, aligned unsupervised response y_i', and input x_i, our multi-objective optimization module fine-tunes the model to generate task-specific but low-bias responses. It fine-tunes the model by optimizing two objectives: target responses as task objective to improve the performance on the target task and unsupervised responses as debiasing objective to mitigate position bias:

$$\begin{aligned} \mathcal{L} =& (1-\alpha) \cdot \mathcal{L}_{target}(x_i, y_i) + \\ & \alpha \cdot \mathcal{L}_{align}(x_i, y_i'), \end{aligned} \tag{5}$$

where α is a hyper-parameter for tradeoff and $\mathcal{L}_{target}(x_i, y_i)$ is a negative log-likelihood (NLL) loss to maximize the generation probability of target response:

$$\mathcal{L}_{target}(x_i, y_i) = -\sum_{j=1}^{|y_i|} \log p(y_{i,j}|x_i, y_{i,<j}). \tag{6}$$

$\mathcal{L}_{align}(x_i, y_i')$ is a task-specific objective to mitigate position bias.

For tasks excluding NLI, such as CQA and CQG, summarization and KGC, we use NLL loss to maximize the probability of generating y_i':

$$\mathcal{L}_{align}(x_i, y_i') = -\sum_{j=1}^{|y_i'|} \log p(y_{i,j}'|x_i, y_{i,<j}'). \tag{7}$$

For NLI, we maximize the generation probability of the most likely class tokens after alignment:

$$\mathcal{L}_{align}(x_i, y_i') = -s_{i,ind(i)} \log p(class_{ind(i)}|x_i), \tag{8}$$

where $ind(i)$ is the index of the class, $class_{ind(i)}$ is the class tokens, and $s_{i,ind(i)}$ is the generation probability of the class tokens in Eq. 4:

$$ind(i) = \arg\max y_i'. \tag{9}$$

y_i' is the masked probability distribution for all classes in Eq. 2.

4 Experiments

We evaluate SOD on three categories of NLP tasks based on changes in conveyed information from input to output: language understanding tasks, language compression tasks and language creation tasks (Deng et al., 2021). Language understanding tasks (e.g., NLI, CQA) aim to comprehend and interpret natural language input given a conversation or document context. For a compression task (e.g., summarization), the goal is to concisely describe the most important information in the input (e.g., a document). A creation task (e.g., CQG, KGC) generates output that adds new information on top of input (e.g., dialogue history).

4.1 Datasets

We conduct experiments on eight widely used benchmark datasets: CANARD (Elgohary et al., 2019), CoQAR (Brabant et al., 2022), CNN/DM (Nallapati et al., 2016), Newsroom (Grusky et al., 2018), Doc2dial (Feng et al., 2020), Mutual (Cui et al., 2020), SNLI (Bowman et al., 2015) and QNLI (Wang et al., 2018), covering five NLP tasks: CQA, CQG, summarization, KGC and NLI. Following previous works (Ko et al., 2020; Shinoda et al., 2022), we split the test dataset into biased dataset and non-biased dataset for simulation depending on the bias type in each dataset. The details of datasets and dataset splitting are provided in §A.1 and §A.2.

4.2 Evaluation Metrics

Following previous works (Chen et al., 2019; Nallapati et al., 2016; Tuan et al., 2020; Meng et al., 2020), we adopt ROUGE-L (Lin, 2004) as evaluation metrics for CQA, CQG, summarization and KGC tasks, in which ROUGE-L has been shown to correlate well with human evaluation (Liu and Liu, 2008). We use macro-accuracy for the classification task, NLI. We use nlg-eval package[3] for the implementation of evaluation metrics.

4.3 Baseline Methods

- **BASE** is the pre-trained LLM with unsupervised instruction-following fine-tuning.
- **Random Position (RP)** (Shinoda et al., 2022) randomly perturbs input positions to reduce the model's dependence on token positions in prediction.
- **Fine-tune (FT)** is the LLM fine-tuned on the dataset for the target task to improve the performance of the target task.
- **MarCQAp** (Gekhman et al., 2023) is a novel prompt-based history modeling approach for CQA and CQG that highlights answers from previous conversation turns by inserting textual prompts in their respective positions.
- **Minimax** (Korakakis and Vlachos, 2023) is an NLI model which leverages an auxiliary model to maximize the loss of the NLI model by upweighting 'hard' samples, thus reducing its reliance of shortcuts in 'easy' samples.
- **GenX** (Varab and Xu, 2023) is a new summarization paradigm that unifies extractive and abstractive summarization with generative modeling.
- **SG-CQG** (Do et al., 2023) is a state-of-the-art CQG models with two stages: what-to-ask for rational span selection in the referential document and how-to-ask for question generation.
- **FocusL** (Deng et al., 2023) is a debiasing method built for KGC by adaptively re-weighting the loss of each token, thus encouraging the model to pay special attention to knowledge utilization.

4.4 Implementation Details

We use FlanT5-large (Chung et al., 2022) as the base LLM for all models. The hidden size is 768. We use the Adam optimizer with a default learning rate $1e^{-4}$ (Kingma and Ba, 2015) and set gradient clipping with a default maximum gradient norm of 1.0[4]. We select the best model based on the BLEU@2 or macro-accuracy score on the validation set. We use α=0.2 for CQA on CoQAR, NLI and KGC tasks and α=0.1 for other tasks, by default. We set the pre-defined thresholds for incoherent identification and unreliable identification from 0.1, 0.15 and 0.2 and select the one that maintains approximately 20% unsupervised responses. We run all experiments with NVIDIA RTX3090 24 GB GPU cards.

5 Results

The overall performances of all methods on language understanding tasks, language creation tasks and language compression tasks are listed in Table 1-3. We have three main observations from the results.

First, LLMs are susceptible to the bias in the dataset after fine-tuning. As we can see in Table 1, FT achieves 34.7% improvement on the biased dataset of CoQAR, but 8.6% improvement on the non-biased dataset. This is because LLMs can easily overfit the shortcut of the training dataset in fine-tuning, just like existing neural networks (Ko et al., 2020).

Second, SOD can mitigate position bias significantly on almost all datasets for three types of tasks. As shown in Table 1-3, SOD improves the performance on the non-biased dataset by 1% to 2% on almost all tasks, compared to FT and all baselines. The reason is that SOD can leverage unsupervised responses with low position bias for optimization in multi-objective optimization module.

[3] https://github.com/Maluuba/nlg-eval

[4] https://huggingface.co/docs/transformers/v4.33.0/en/main_classes/trainer

Table 1: Overall performance (%) on language understanding tasks. Boldface indicates the best results in terms of the corresponding dataset.

Method	NLI (%)				CQA (%)			
	SNLI		QNLI		CoQAR		CANARD	
	Biased	Non-Biased	Biased	Non-Biased	Biased	Non-Biased	Biased	Non-Biased
BASE	77.5	79.0	90.3	88.3	47.4	39.8	34.0	17.3
RP	—	—	—	—	61.5	51.6	60.4	20.4
MarCQAp	—	—	—	—	**66.9**	52.3	66.3	21.3
Minimax	91.6	87.0	90.9	89.7	—	—	—	—
FT	**92.0**	88.0	**94.3**	89.8	64.6	51.9	**67.5**	20.8
SOD	**92.0**	**88.4**	**94.3**	**90.9**	66.3	**53.7**	65.7	**21.9**

Table 2: Overall performance (%) on language creation tasks. Boldface indicates the best results in terms of the corresponding dataset.

Method	CQG (%)				KGC (%)			
	CoQAR		CANARD		Doc2dial		Mutual	
	Biased	Non-Biased	Biased	Non-Biased	Biased	Non-Biased	Biased	Non-Biased
BASE	16.0	17.0	17.9	17.1	23.9	13.7	25.4	21.3
RP	23.2	18.3	24.6	21.2	35.3	32.2	89.6	46.1
MarCQAp	19.7	17.5	**26.0**	21.8	—	—	—	—
SG-CQG	14.9	15.5	19.7	17.8	—	—	—	—
FocusL	—	—	—	—	39.8	35.9	83.0	51.7
FT	**26.7**	17.2	**26.0**	21.6	38.7	36.0	93.9	38.9
SOD	**26.7**	**18.8**	25.9	**22.4**	**40.6**	**38.3**	**94.0**	**53.0**

Table 3: Overall performance (%) on language compression tasks. Boldface indicates the best results in terms of the corresponding dataset.

Method	Summarization (%)			
	CNN/DM		Newsroom	
	Biased	Non-Biased	Biased	Non-Biased
BASE	22.3	11.6	35.1	20.0
RP	23.9	15.1	47.5	**22.0**
GenX	17.6	13.7	29.2	19.0
FT	26.9	16.8	**51.1**	19.8
SOD	**27.1**	**17.3**	50.9	21.3

Third, SOD only sacrifices a small performance on the biased dataset when mitigating position bias. As shown in Table 3, RP achieves comparable performance to SOD on the non-biased dataset of Newsroom. However, the ROUGE-L of RP drops 3.6% compared to FT on the biased dataset, while that of SOD only drops 0.2%. This is because the perturbation in RP impairs the overall data quality for fine-tuning, while unsupervised responses in SOD are aligned to improve the quality in §3.2.

Note that some baselines achieve poor performance, sometimes lower than BASE. First, in Table 3, GenX performs worse even than BASE on the biased dataset of Newsroom. The reason is that the summarization datasets are abstractive and suitable for generative models, e.g., T5, while GenX is an extractive baseline. Second, in Table 2, the performance of SG-CQG is worse than that of FT on the biased dataset of CoQAR. This is because SG-CQG is a three-stage question generation model focusing on improving question diversity. The selected answer span for question generation is randomly chosen from massive generated candidates.

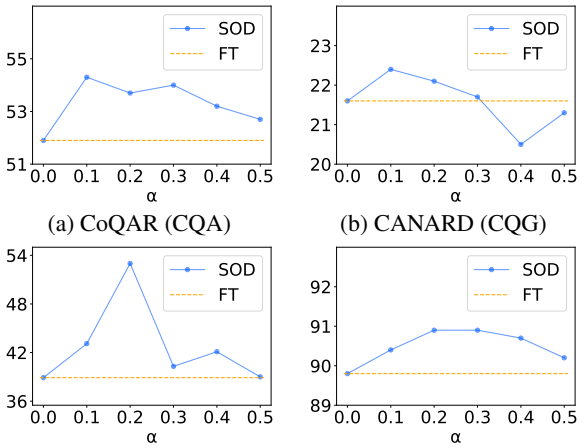
Besides, to further verify the effectiveness of SOD, we have done t-test significant tests. We found that SOD can achieve significant improvement over MarCQAp on CoQAR (CQA and CQG) and CANARD (CQG), over FocusL on Doc2dial and Mutual, over FT on CNN/DM. The above results indicate that SOD can mitigate position bias in most cases.

6 Analysis

In this section, we analyze the effect of the quality of unsupervised responses in §6.1 and objective

Table 4: SOD performance (%) of CQG task using various unsupervised responses. 'N-Biased' denotes performance on non-biased datasets. Boldface indicates the best results in terms of the corresponding dataset.

Method	CoQAR		CANARD	
	Biased	N-Biased	Biased	N-Biased
SOD	**26.7**	**18.8**	25.9	**22.4**
- w/o OAM	25.8	18.0	**26.1**	21.5
- w/ T5-base	26.3	18.3	**26.1**	22.0
- w/ T5-xlarge	26.6	17.9	25.7	22.0
FT	**26.7**	17.2	26.0	21.6



(a) CoQAR (CQA) (b) CANARD (CQG)

(c) Mutual (d) QNLI

Figure 3: Performance (%) of four tasks over each α. The x-axis denotes the value of α and the y-axis denotes the ROUGE-L score on non-biased datasets.

weighting in the multi-objective optimization module in §6.2. The overall results of all tasks are presented in §C. Besides, we also conduct a case study in §B.1 and provide cases for all datasets in §C.4.

6.1 Analysis of OAM

To analyze the effect of the OAM module, we conduct analyses with unsupervised responses obtained from various sources with different qualities in Table 4. SOD w/o OAM, SOD w/ T5-base and SOD w/ T5-xlarge denote SOD using unsupervised responses without alignment, responses from FlanT5-base and FlanT5-xlarge, respectively. We have two observations.

First, low-quality responses without OAM leads to worse performance. As we can see, SOD outperforms SOD w/o OAM on non-biased datasets by leveraging OAM for enhancing the response quality. Poor-quality responses will undermine the model comprehension of the task, thus leading to worse performance.

Second, OAM is robust to various sources of unsupervised responses. SOD w/ T5-base and SOD w/ T5-xlarge perform worse than SOD on CANARD, but still outperform FT. We infer that responses from other LLMs use different knowledge/parameters for generation, which mismatch with that of T5. The difference amplifies the divergence between the task objective and the debiasing objective in §3.2, thus leading to worse performance. Even that, OAM can reduce this divergence to better achieve both objectives.

6.2 Analysis of Objective Weighting

To analyze the effect of weighting on the debiasing objective, we present the performance of SOD using different α in Fig. 3. We have two observations from the results.

First, the performance of SOD drops with the increase of the weight of unsupervised responses in multi-objective optimization. In CQA on CoQAR, the ROUGE-L score of SOD drops from 53.6% to 52.7% when increasing α from 0.1 to 0.5. This is because the model performance depends on not only the degree of bias but also the data quality. Increasing α will reduce the position bias of all responses, yet it will hurt the quality concurrently.

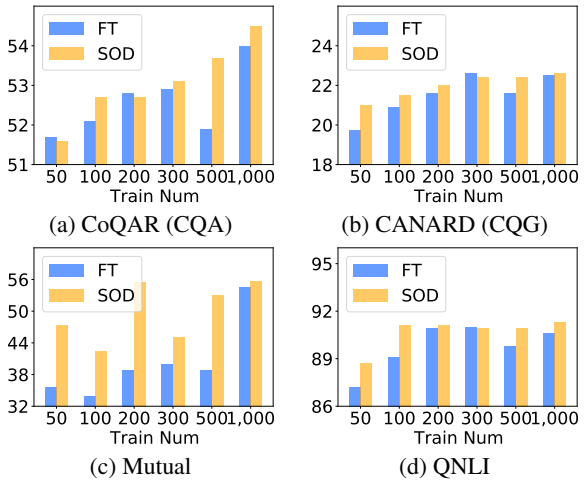
Second, our proposed SOD always outperforms FT under various α. As shown in Fig. 3, SOD performances of CQA on CoQAR all exceed 52.5% using various α, while FT only achieves 52.0%. This demonstrates the effectiveness and robustness of SOD in mitigating position bias.

6.3 Analysis on Training Samples

We also analyze the effect of the number of training samples to verify the effectiveness of SOD under various low-resource settings. We plot the results in Fig. 4 and have two observations.

First, nearly all methods perform better when increasing training samples. As we can see in Fig. 4, the ROUGE-L score of FT increases from 19.7% to 22.5% on CANARD when the number of training samples increases from 50 to 1,000. Increasing training samples can improve the generalization ability of LLMs, thus leading to better performance.

Second, our proposed SOD outperforms FT under various low-resource settings. As shown in Fig. 4, the ROUGE-L scores of SOD depicted by the orange bars are consistently higher than those of FT in blue. This is because there are always



(a) CoQAR (CQA) (b) CANARD (CQG)

(c) Mutual (d) QNLI

Figure 4: Performance (%) of four tasks over different numbers of training samples. The x-axis denotes the number of training samples and the y-axis denotes the ROUGE-L score on non-biased datasets.

around 40% aligned unsupervised responses for fine-tuning regardless of the variation in the number of training samples, which are enough for effective debiasing.

7 Related Work

7.1 Position Bias and Debiasing

Position bias has been explored mainly in three extractive tasks, i.e., NLI, summarization and CQA. In, NLI, Gururangan et al. (2018); Poliak et al. (2018) show that the class labels are highly correlated to certain words in the hypothesis. McCoy et al. (2019) report that models always rely on word overlap between hypothesis and premise for prediction. Karimi Mahabadi et al. (2020); Du et al. (2021) train a robust NLI model in an ensemble manner by PoE. They train a hypothesis model to learn the lexical bias, which guides the NLI model to focus on other patterns in the dataset that generalize better. Ghaddar et al. (2021) re-weight the importance of easy and hard samples to prevent the model from fitting the bias. As the weights are derived from the model itself, this method does not require any external annotations.

Similar findings are reported in CQA tasks. Weissenborn et al. (2017); Sugawara et al. (2018) demonstrate that only using partial inputs is sufficient to correctly extract the answer span for the question in most cases. Ko et al. (2020) address that absolute position of answer spans can work as a spurious clue for prediction. Shinoda et al. (2022) report that relative position of answer spans is another clue for position bias. To mitigate position bias, Ko et al. (2020); Shinoda et al. (2022) build bias ensemble models by PoE similar to Karimi Mahabadi et al. (2020). They design biased models with position-only features to guide CQA models to rely more on semantic features for answering.

Kedzie et al. (2018); Grenander et al. (2019) find that 58% of selected summary utterances come directly from the lead utterances, and models trained on these articles perform considerably worse when utterances in the article are randomly shuffled. To mitigate lead bias, a simple but effective method is to randomly shuffle the document for training (Grenander et al., 2019). Then, Xing et al. (2021) uses adversarial training for debiasing. They design a position prediction module and optimize the reverse loss for position prediction, forcing the encoder to leverage non-position features.

However, existing works on mitigating position bias always focus on extractive tasks. They can hardly be transferred to generative tasks as the label space in generative tasks is too large for bias estimation or adversarial training. In this paper, we focus on position debiasing for generative LLMs.

7.2 Debiasing for LLMs

Works on debiasing for LLMs always focus on social bias, e.g., gender bias and racial bias, rather than position bias (Meade et al., 2022; Du et al., 2022). Existing works on mitigating social bias for LLMs can be classified into three types: pre-processing methods, in-processing methods and post-processing methods.

In pre-processing, Zmigrod et al. (2019) adopt a counterfactual data augmentation (CDA) algorithm to mitigate social bias by swapping bias attribute words (e.g., he/she) in training dataset. Choi et al. (2022) modify CDA by masking the terms casual to label to force the model to learn label-invariant features.

In in-processing, Guo et al. (2022); Li et al. (2023); Yang et al. (2023) propose a two-stage adversarial method for debiasing. They first train a continuous prompt to enlarge the bias of utterance pairs and then force the LLMs to minimize the difference of utterance pairs using the prompt.

In post-processing, Schick et al. (2021) propose a decoding algorithm that reduces the probability of a model producing biased text. They use a textual description of the undesired behaviors for prompting. Meade et al. (2023) propose an ICL strategy which leverages non-biased demonstrations to guide the generation for safety.

However, existing methods for mitigating social bias either require the external bias knowledge for data augmentation and training adversarial prompts or require a non-biased dataset for building demonstrations, which are lacking for position debiasing and unpractical in application. Differently, we propose a self-supervised framework for LLMs on mitigating position bias, without relying on any external bias knowledge or non-biased samples, which is general, simple but effective.

8 Conclusion

In this paper, we have proposed a self-supervised debiasing framework SOD for LLMs. It adopts a multi-objective optimization module to mitigate position bias for LLMs, where unsupervised responses for the debiasing objective are of low position bias. These responses are pruned by a proposed OAM module for aligning the task and debiasing objectives. Extensive experiments on five tasks and eight benchmark datasets show that SOD outperforms existing baselines on non-biased samples while retaining performance or sacrificing little performance on biased samples. It demonstrates that leveraging unsupervised responses is a practicable solution to mitigate position bias for generative LLMs.

Limitations

This work has the following limitations. First, SOD needs a pre-trained LLM to generate unsupervised responses with low bias, where these responses are still biased. Second, the final performance of SOD depends on the quality of unsupervised responses, which are still noisy after being aligned by OAM. In future work, we plan to address these issues by investigating non-biased models from other domains and model-based strategies to align unsupervised responses.

Ethical Considerations

We realize that there are risks in developing generative LLMs, so it is necessary to pay attention to the ethical issues of LLMs. We use publicly available pre-trained LLMs, i.e., FlanT5-base, FlanT5-large, FlanT5-xlarge, and publicly available datasets in the academic community, i.e., CANARD, CoQAR, CNN/DM, Newsroom, Doc2dial, Mutual, SNLI, QNLI, to conduct experiments. All models and datasets are carefully processed by their publishers to ensure that there are no ethical problems.

Acknowledgments

We thank the reviewers for their valuable feedback. This work was supported by the National Key R&D Program of China with grant No.2022YFC3303004, the Natural Science Foundation of China (62372275, 62272274, 62202271, T2293773, 62102234, 62072279), the Natural Science Foundation of Shandong Province (ZR2021QF129).

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642.

Quentin Brabant, Gwénolé Lecorvé, and Lina M. Rojas Barahona. 2022. CoQAR: Question rewriting on CoQA. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022*, pages 119–126.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd workshop on machine reading for question answering*, pages 119–124.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2174–2184.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. 2022. C2l: Causally contrastive learning for robust text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 10526–10534.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 1406–1416.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 7580–7605.

Yifan Deng, Xingsheng Zhang, Heyan Huang, and Yue Hu. 2023. Towards faithful dialogues via focus learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 4554–4566.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Xuan Long Do, Bowei Zou, Shafiq Joty, Tran Tai, Liangming Pan, Nancy Chen, and Ai Ti Aw. 2023. Modeling what-to-ask and how-to-ask for answer-unaware conversational question generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 10785–10803.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2022. Shortcut learning of large language models in natural language understanding: A survey. *arXiv preprint arXiv:2208.11857*.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2021*, pages 915–929.

Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. 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 2019*, pages 5917–5923.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 8118–8128.

Zorik Gekhman, Nadav Oved, Orgad Keller, Idan Szpektor, and Roi Reichart. 2023. On the robustness of dialogue history representation in conversational question answering: A comprehensive study and a new prompt-based method. *Transactions of the Association for Computational Linguistics, TACL 2023*, 11:351–366".

Abbas Ghaddar, Phillippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021. End-to-end self-debiasing framework for robust NLU training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1923–1929.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. 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 2019*, pages 6019–6024.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), NAACL 2018*, pages 708–719.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 1012–1023.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), NAACL 2018*, pages 107–112.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8706–8716.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 1818–1828.

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

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 1109–1121.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances*

in neural information processing systems, NeurIPS 2022, 35:22199–22213.

Michalis Korakakis and Andreas Vlachos. 2023. Improving the robustness of nli models with minimax training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 14322–14339.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 14254–14267.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Feifan Liu and Yang Liu. 2008. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, short papers*, pages 201–204.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448.

Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tür. 2023. Using in-context learning to improve dialogue safety. *arXiv preprint arXiv:2302.00871*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 1878–1898.

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020*, pages 1151–1160.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, SIGNLL 2016*, pages 280–290.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM 2018*, pages 180–191.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics, TACL 2019*, 7:249–266.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics, TACL 2021*, 9:1408–1424.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2017*, pages 1073–1083.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2022. Look to the right: Mitigating relative position bias in extractive question answering. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP 2022*, pages 418–425.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4208–4219.

Luu Anh Tuan, Darsh Shah, and Regina Barzilay. 2020. Capturing greater context for question generation. In *Proceedings of the AAAI conference on artificial intelligence, AAAI 2020*, pages 9065–9072.

Prasetya Ajie Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 9063–9074.

Daniel Varab and Yumo Xu. 2023. Abstractive summarizers are excellent extractive summarizers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023*, pages 330–339.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, EMNLP 2018*, pages 353–355.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning, CoNLL 2017*, pages 271–280.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), ACL-IJCNLP 2021*, pages 948–954.

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, AAAI 2023*, pages 10780–10788.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 1651–1661.

A Experimental Details

A.1 Datasets

We conduct experiments on eight widely used benchmark datasets: CANARD, CoQAR, CNN/DM, Newsroom, Doc2dial, Mutual, SNLI and QNLI, covering five NLP tasks: CQA, CQG, summarization, KGC and NLI.

- **CANARD** (Elgohary et al., 2019) is a benchmark dataset for CQA and CQG. It is built based on QuAC (Choi et al., 2018) and consists of 40k questions with different context lengths, where answers are selected spans from a given section in a Wikipedia article.
- **CoQAR** (Brabant et al., 2022) is a large-scale dataset for CQA and CQG. It annotates 53k questions based on CoQA (Reddy et al., 2019), where the documents are from seven diverse domains.
- **CNN/DM** (Nallapati et al., 2016) is a well-known summarization dataset, which consists of 313k articles from CNN and Daily Mail. The summary is written by human experts and shown as bullet points. We use the non-anonymized version (See et al., 2017).
- **Newsroom** (Grusky et al., 2018) is a large-scale summarization dataset which contains 1.3 million articles and expert-written summaries with high diversity.
- **Doc2dial** (Feng et al., 2020) is a document-grounded dialogue dataset with 4,800 annotated conversations and an average of 14 turns. Compared to the prior document-grounded dialogue datasets, this dataset covers a variety of dialogue scenes in information-seeking conversations.
- **Mutual** (Cui et al., 2020) is a multi-turn reasoning dialogue dataset, consisting of 8,860 manually annotated dialogues based on Chinese student English listening comprehension exams. It is challenging since it requires a model to handle various reasoning problems.
- **SNLI** (Bowman et al., 2015) is a large-scale natural language inference benchmark with 570k utterance pairs. Each pair is manually labeled as entailment, neutral, or contradiction with several annotators.
- **QNLI** (Wang et al., 2018) is a natural language inference dataset derived from the Stanford Question Answering Dataset v1.1. An utterance is extracted from the passage and paired with the question. Each pair is then manually labeled according to whether the utterance contains the answer to the question.

A.2 Bias Types and Dataset Splitting

In this work, we focus on mitigating three types of widely addressed position bias: lead bias (Kedzie et al., 2018), relative position bias (Shinoda et al., 2022) and lexical bias (Gururangan et al., 2018).

- Lead bias in summarization is a phenomenon that the generated summary is highly correlated to utterances appearing at the beginning of the document (Kedzie et al., 2018).
- Relative position bias in QA is a phenomenon that a QA model tends to degrade the performance on samples where answers are located in relative positions unseen during training (Shinoda et al., 2022). The relative position is defined as the relative position of grounded utterances between the last turn answer and the current turn answer.
- Lexical bias is the phenomenon that deep learning models achieve high accuracy by exploiting trigger words or word overlapping (Gururangan et al., 2018; Poliak et al., 2018). Note that lexical bias is a type of position bias in generative models. For example, trigger words in hypothesis in generative models are regarded as trigger words behind 'Hypothesis: ', which is a positional clue for prediction.

Following previous works (Ko et al., 2020; Shinoda et al., 2022), we split the test dataset into biased dataset and non-biased dataset for simulation depending on the bias type in each dataset. In CQG and CQA datasets (CANARD and CoQAR), we select the samples with relative position equaling 0 or 1 into the biased dataset and the left samples into the non-biased dataset. In KGC and summarization datasets (Doc2dial, Mutual, CNN/DM and Newsroom), we filter samples where the reference response is highly correlated to the beginning utterance of the given document into the biased dataset and the left samples into the non-biased dataset. For NLI datasets (SNLI and QNLI), samples with specific words are filtered into the biased dataset and other samples are filtered into the non-biased dataset. The dataset statistics are shown in Table 5.

Table 5: Dataset statistics after splitting.

Dataset	Biased			Non-biased
	#Train	#Dev	#Test	#Test
CANARD	500	250	3,460	2,440
CoQAR	500	250	3,222	4,873
CNN/DM	500	250	1,421	5,000
Newsroom	500	200	5,000	5,000
Doc2dial	500	250	2,000	5,000
Mutual	500	250	2,000	5,000
SNLI	500	250	2,000	5,000
QNLI	500	250	2,000	5,000

B Analysis

B.1 Case Study

To investigate the reason for the effectiveness of SOD, we present an example of generated responses in CQA in Table 6 and an example of unsupervised responses in CQG in Table 7.

In CQA example, SOD can generate an answer with lower bias than FT. As shown in Table 6, FT generates an answer from U5, adjacent to U4, the utterance containing the last turn answer. In contrast, SOD generates an answer from U1, which is far from U4. After fine-tuning with unsupervised responses from various positions, SOD cannot easily fall into the trap of finding answers from neighboring utterances of the one containing the last turn answer.

In CQG example in Table 7, SOD uses unsupervised responses which have lower bias than the target response. As we can see, target question is based on the utterance with relative position 1, while unsupervised questions generated from BASE are based on utterances varying in the document. Fine-tuning with unsupervised responses generated from pre-trained LLMs encourages the model to generate questions based on utterances in various document positions, thus mitigating position bias.

Table 6: An example of generated answers on CANARD. U1, U4, and U5 are the 1st, 4th, 5th utterances in the document, T1 and T2 are the first and second turn utterances of the dialogue. 'Target' is the target answer of this example, 'FT' and 'SOD' are the generated answers of FT and SOD. 'Position' is the position of the grounded utterance of the answer in the document and 'Relative position' is the distance of grounded utterances between the current answer and the last turn answer.

ID	Document	
U1	Gaston had worked with players at an individual level as a hitting instructor ...	
U4	In the six games the Blue Jays played in those places during World Series play ...	
U5	... and Gaston was the first ever African-American manager to win a World Series.	
ID	**Context**	**Position**
T1	Question: What happens in the series? Answer: In the six games the Blue Jays played in those places during World Series play, the Jays went 4-2 ...	U4
T2	Question: What else did Cito do?	
Model	**Answer**	**Relative position**
Target	Gaston had worked with players at an individual level ...	-3 (U1)
FT	... Gaston was the first ever African-American manager to win a World Series.	1 (U5)
SOD	Gaston had worked with players at an individual level ...	-3 (U1)

Table 7: An example of unsupervised questions on CANARD. U1–U6 are the first six utterances in the document and T1 is the first turn utterance of the dialogue. 'Target' is the target question of this example, 'BASE-1' and 'BASE-2' are the questions generated by pre-trained LLM. 'Position' is the position of the grounded utterance of the question in the document and 'Relative position' is the distance of grounded utterances between the current question and the last turn question.

ID	Document	
U1	Gautam Gambhir, born 14 October 1981, is an Indian cricketer, ...	
U2	Gambhir was picked up by the Delhi Daredevils franchise in the first player auction of the Indian Premier League for a price of US$725,000 a year.	
U3	He became the second highest run-scorer of the inaugural season with ...	
U4	He was promoted to the post of Captain of the Delhi Daredevils for IPL Season 2010.	
U5	At the end of the tournament he became the only player from Delhi Daredevils to score more than 1000 runs in the IPL.	
U6	In the 2011 IPL player auction, ...	
ID	**Context**	**Position**
T1	Question: Who was Gautam Gambhir in Indian premier league? Answer: He was promoted to the post of Captain of the Delhi Daredevils for IPL Season 2010.	U4
Source	**Question**	**Relative position**
Target	What did Gautam Gambhir do as captain?	1 (U5)
BASE-1	Where was Gautam Gambhir born?	-3 (U1)
BASE-2	Was Gautam Gambhir in Indian premier league?	-2 (U2)

C Overall Analysis

Table 8: SOD Performance (%) using various unsupervised responses on language understanding tasks. Boldface indicates the best results in terms of the corresponding dataset.

Method	NLI (%)				CQA (%)			
	SNLI		QNLI		CoQAR		CANARD	
	Biased	Non-Biased	Biased	Non-Biased	Biased	Non-Biased	Biased	Non-Biased
SOD	**92.0**	**88.4**	94.3	90.9	**66.3**	**53.7**	65.7	21.9
- w/o OAM	91.0	87.4	92.7	88.7	**66.3**	**53.7**	65.7	21.9
- w/ T5-base	90.5	87.6	**94.4**	90.6	64.3	52.3	61.6	19.9
- w/ T5-xlarge	91.2	87.7	93.2	**91.3**	65.1	53.6	67.1	**22.4**
FT	**92.0**	88.0	94.3	89.8	64.6	51.9	**67.5**	20.8

Table 9: SOD Performance (%) using various unsupervised responses on language creation tasks. Boldface indicates the best results in terms of the corresponding dataset.

Method	CQG (%)				KGC (%)			
	CoQAR		CANARD		Doc2dial		Mutual	
	Biased	Non-Biased	Biased	Non-Biased	Biased	Non-Biased	Biased	Non-Biased
SOD	**26.7**	**18.8**	25.9	**22.4**	40.6	38.3	94.0	**53.0**
- w/o OAM	25.8	18.0	**26.1**	21.5	32.6	34.1	91.0	36.7
- w/ T5-base	26.3	18.3	**26.1**	22.0	42.8	36.3	**94.3**	45.1
- w/ T5-xlarge	26.6	17.9	25.7	22.0	**44.1**	**39.1**	**94.3**	52.7
FT	**26.7**	17.2	26.0	21.6	38.7	36.0	93.9	38.9

Table 10: SOD Performance (%) using various unsupervised responses on language compression tasks. 'N-Biased' denotes performance on the non-biased datasets. Boldface indicates the best results in terms of the corresponding dataset.

Method	Summarization (%)			
	CNN/DM		Newsroom	
	Biased	N-Biased	Biased	N-Biased
SOD	27.1	**17.3**	50.9	**21.3**
- w/o OAM	27.1	14.8	48.0	20.2
- w/ T5-base	26.8	14.2	47.7	20.3
- w/ T5-xlarge	**28.0**	14.6	48.4	20.2
FT	26.9	16.8	**51.1**	19.8

C.1 Analyses of Unsupervised Responses

Table 8-10 demonstrate SOD performance using different qualities of unsupervised responses. As we can see, lower response quality leads to worse performance. Besides, using various sources of unsupervised responses always degrades model performance on non-biased datasets.

C.2 Analyses of Objective Weighting

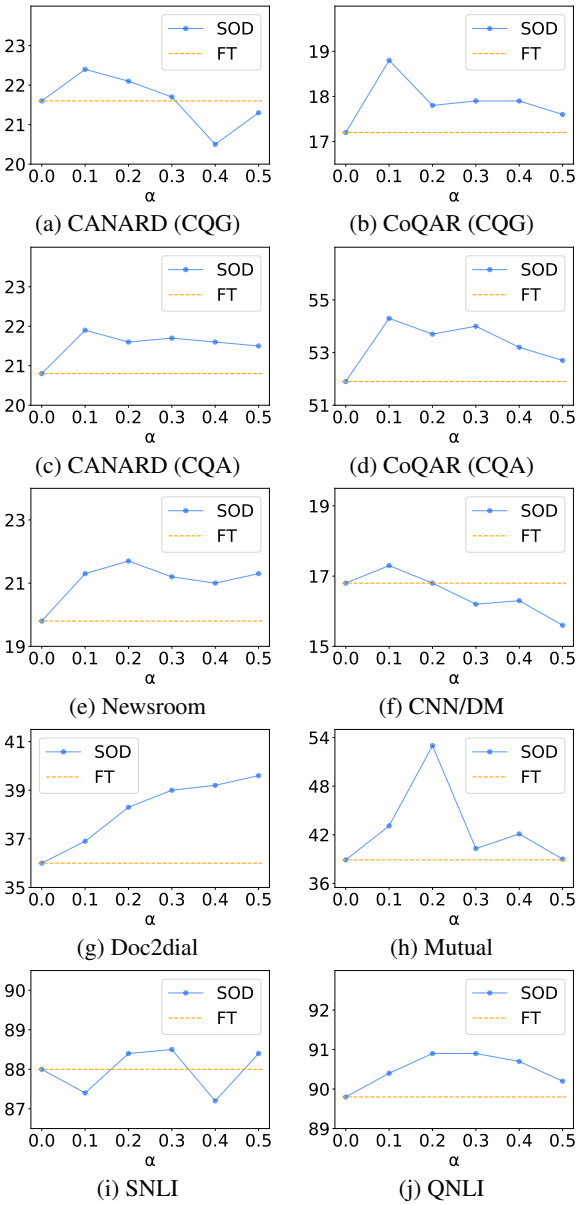
Fig. 5 provides the overall performance over different α. In most cases, SOD outperforms the fine-tuned LLM, i.e., FT.

C.3 Analyses on Training Samples

Fig. 6 provides the overall performance on different numbers of training samples. In most cases, SOD outperforms the fine-tuned LLM, i.e., FT.

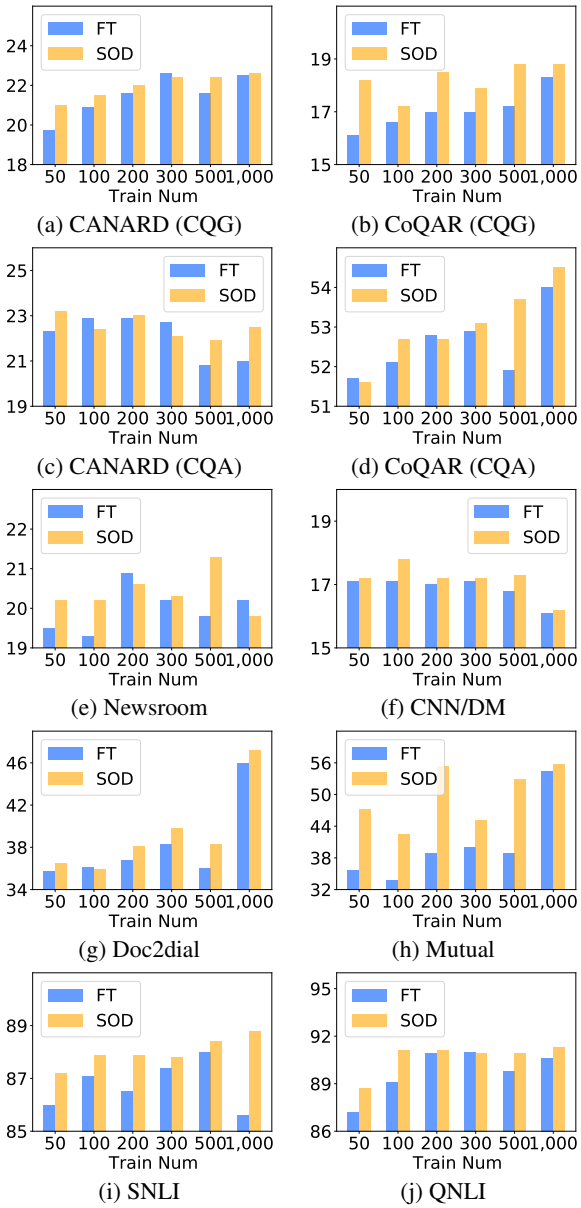
C.4 Cases

We also present cases for CQA, CQG, summarization and KGC tasks in Table 11–17. As we can see, FT always finds relevant information from utterances near the grounded utterances of the last turn utterance for CQA and CQG tasks or from the lead utterance for summarization and KGC tasks. While SOD can find relevant knowledge from any utterances in the document. Note that on the Mutual dataset, where each sample has four candidate responses as the document, FT always fails to select the relevant utterance from the last two responses.



(a) CANARD (CQG) (b) CoQAR (CQG)
(c) CANARD (CQA) (d) CoQAR (CQA)
(e) Newsroom (f) CNN/DM
(g) Doc2dial (h) Mutual
(i) SNLI (j) QNLI

Figure 5: Performance over each α on all datasets. The x-axis denotes the value of α and the y-axis denotes the ROUGE-L score on non-biased datasets.



(a) CANARD (CQG) (b) CoQAR (CQG)
(c) CANARD (CQA) (d) CoQAR (CQA)
(e) Newsroom (f) CNN/DM
(g) Doc2dial (h) Mutual
(i) SNLI (j) QNLI

Figure 6: Performance over different numbers of training samples on all datasets. The x-axis denotes the number of training samples and the y-axis denotes the ROUGE-L score on non-biased datasets.

Table 11: An example of CQA on CoQAR. U1–U4 are the first four utterances in the document, T1 and T2 are the 1st and 2nd turn utterances of the dialogue. 'Target' is the target answer of the example, 'FT' and 'SOD' are the generated answers from FT and SOD. 'Position' is the position of the grounded utterance of the answer in the document and 'Relative position' is the distance of grounded utterances between the current answer and the last turn answer.

ID	Document	
U1	This final war was to give thousands of colonists, including ..., military experience ...	
U2	By far the largest military action in which the United States engaged during this era was the War of 1812.	
U3	With Britain locked in a major war with Napoleon's France, its policy was to block American shipments to France	
U4	The United States sought to remain neutral while pursuing overseas trade.	
ID	Context	Position
T1	Question: What did this give the colonists? Answer: Military experience.	U1
T2	Question: Whose side was the US on at first in the war of 1812?	
Model	Answer	Relative position
Target	At first neutral	3 (U4)
FT	Britain.	2 (U3)
SOD	Neutral	3 (U4)

Table 12: An example of CQG on CoQA. U1–U5 are the first five utterances in the document and T5 is the 5th turn utterance of the dialogue. 'Target' is the target question of the example and 'FT' and 'SOD' are the generated questions from FT and SOD. 'Position' is the position of the grounded utterance of the question in the document and 'Relative position' is the distance of grounded utterances between the current question and the last turn question.

ID	Document	
U1	John's Metropolitan Area is the second largest Census Metropolitan Area (CMA) in Atlantic Canada ...	
U3	Its name has been attributed to the feast day of John the Baptist, when John Cabot was believed to have sailed into the harbor in 1497.	
U4	St. John's is one of the oldest settlements in North America, with year-round settlement beginning sometime after 1630 and seasonal habitation long before that.	
U5	It is not, however, the oldest surviving English settlement in North America or Canada ...	
ID	Context	Position
T5	Question: In what continent is it located? Answer: North America.	U4
Model	Question	Relative position
Target	What year did John Cabot arrive in the harbor?	-1 (U3)
FT	When was St. John's founded?	0 (U4)
SOD	What is the second largest CMA in Atlantic Canada?	-3 (U1)

Table 13: An example of CQG on CANARD. U1–U5 are the first five utterances in the document and T1 is the 1st turn utterance of the dialogue. 'Target' is the target question of the example and 'FT' and 'SOD' are the generated questions from FT and SOD. 'Position' is the position of the grounded utterance of the question in the document and 'Relative position' is the distance of grounded utterances between the current question and the last turn question.

ID	Document
U1	In May 2009, production company CinemaNX announced that it would distribute Me and Orson Welles itself ...
U2	It opened the New Orleans Film Festival on October 9, 2009; and it was screened at the St. Louis International Film Festival in November 2009.
U3	The film was released in the US on November 25, 2009, and in the UK on December 4, 2009.
U4	IndieWIRE reported, "The do-it-yourself release of Richard Linklater's Me and Orson Welles bluegot off to a very nice start, averaging $15,910 from its four theaters, the highest PTA of all debuting films."
U5	While Orson Welles is one the first examples of such a high-profile film going to the DIY route, if it proves successful, it's going to be done a lot more in the future.

ID	Context	Position
T1	Question: What month was Me and Orson Welles released in theaters? Answer: The film was released in the US on November 25, 2009.	U3

Model	Question	Relative position
Target	What were critics reviews of Me and Orson Welles?	2 (U5)
FT	When was the film Me and Orson Welles released in the UK?	0 (U3)
SOD	What was the response to Me and Orson Welles?	2 (U5)

Table 14: An example of summarization on Newsroom. U1, U2, U7 and U8 are the 1st, 2nd, 7th and 8th utterances in the document. 'Target' is the target summary of the document and 'FT' and 'SOD' are the generated summary from FT and SOD. 'Position' is the position of the utterance associated with the summary in the document.

ID	Document
U1	Joe Staley celebrates Colin Kaepernick's touchdown run in Super Bowl XLVII.
U2	Cubs inspire Super Bowl memories for Staley Cubs fans were euphoric Wednesday night.
U7	Staley, 32, is the longest-tenured member of a team that's 6-17 since 2015.
U8	Does he think the 49ers, who are nearly 22 years removed from their last title, can win a Super Bowl before he retires? " Yeah," he said, "I'll never give up, man."

Model	Summary	Position
Target	Cubs inspire Super Bowl memories for Staley Cubs fans were euphoric Wednesday night ...	U2
FT	Joe Staley celebrates Colin Kaepernick's touchdown run in Super Bowl XLVII. Cubs inspire Super Bowl memories for Staley.	U1
SOD	Joe Staley says he wished the 49ers had won the Super Bowl after the 2012 season.	U8

Table 15: An example of summarization on CNN/DM. U1–U3 are the first three utterances in the document. 'Target' is the target summary of the document and 'FT' and 'SOD' are the generated summary from FT and SOD. 'Position' is the position of the utterance associated with the summary in the document.

ID	Document
U1	Negotiations between the United States and Libya that could result in compensation for past acts of state-sponsored terrorism by Libya are under way.
U2	The wreckage of Pan Am 103 in Lockerbie, Scotland; the bombing killed 270 people in 1989. U.S. and Libyan officials met Wednesday and Thursday, the official said.
U3	The nations hope to hammer out a deal in which Libya would "resolve all outstanding claims in good faith" and offer "fair compensation" to victims and their families, he said.

Model	Summary	Position
Target	The negotiations could result in compensation for past acts of state-sponsored terrorism by Libya.	U3
FT	The nations hope to hammer out a deal in which Libya would "resolve all outstanding claims in good faith".	U1
SOD	Nations hoping for deal in which Libya would compensate terrorism victims.	U3

Table 16: An example of KGC on Doc2dial. U7–U10 represent the consecutive four utterances following the 7th utterance in the document. T1 and T2 are the 1st and 2nd turn utterances of the dialogue. 'Target' is the target response of the example, 'FT' and 'SOD' are the generated responses from FT and SOD. 'Position' is the position of the grounded utterance of the response in the document and 'Relative position' is the distance of grounded utterances between the current response and the last turn response.

ID	Document
U7	Should I Sign Up For Medical Insurance (Part B)?
U8	When you sign up for Medicare, you will be asked if you want to enroll in Medical Insurance Part B.
U9	If you do not choose to enroll in Medicare Part B and then decide to do so later, your coverage maybe delayed and you may have to pay a higher monthly premium unless you qualify for a "Special Enrollment Period" or SEP.
U10	You may qualify for a SEP that will let you sign up for Part B during ...

ID	Context	Position
T1	Question: We don't need to get any supplementary medical insurance, isn't? Answer: Yes, that's correct. You may qualify for a SEP that will let you sign up for Part B.	U10
T2	Question: Do I sign up for medicare part B?	

Model	Answer	Relative position
Target	When you sign up for Medicare you will be asked if you want to enroll in Medical insurance Part B.	-2 (U8)
FT	Yes, that's correct. You may qualify for a SEP that will let you sign up for Part B.	0 (U10)
SOD	Yes, you will be asked if you want to enroll in Medical insurance Part B.	-2 (U8)

Table 17: An example of KGC on Mutual. U1–U4 represent the four candidate responses. T1 and T2 are the 1st and 2nd turn utterances of the dialogue. 'Target' is the target response of the example, 'FT' and 'SOD' are the generated responses from FT and SOD. 'Position' is the position of the selected response.

ID	Document	
U1	It does n't matter. you just joined a new team, and the manager said it's normal that you are not good at interpersonal skills.	
U2	Although the manager said you are not good at interpersonal skills, you still evaluated others' performances.	
U3	So you had your performance evaluation yesterday and were praised by the manager, right?	
U4	yeah, you were praised by the manager, weren't you?	
ID	Context	Position
T1	Female: you look happy. Male: I am. I had my performance evaluation today.	–
T2	Female: so it went well? Male: yes, the manager said my interpersonal skills are great. I work well with others.	–
Model	Female	Position
Target	Yeah, you were praised by the manager, weren't you?	U4
FT	It does n't matter. you just joined a new team, and the manager said it's normal that you are not good at interpersonal skills.	U1
SOD	Yeah, you were praised by the manager, weren't you?	U4