Mitigating Biases for Instruction-following Language Models via Bias Neurons Elimination

Anonymous ACL submission

Abstract

Instruction-following language models often show undesirable biases. These undesirable biases are accelerated in the real-world usage of language models, where a wide range of instructions is used through zero-shot example prompting. To solve this problem, we first define ***bias neuron***, which significantly affects biased outputs, and prove its existence empirically. Furthermore, we propose a novel and practical bias mitigation method, CRISPR, to eliminate bias neurons of language models in instruction-following settings. CRISPR automatically determines biased outputs and categorizes neurons that affect the biased outputs as bias neurons using an attribution, an explainability method. Experimental results demonstrate the effectiveness of our method in mitigating biases under zero-shot instruction-following settings without losing the model's task performance and existing knowledge. The experimental results reveal the generalizability of our method as it shows robustness under various instructions and datasets. Surprisingly, our method can mitigate the bias in language models by eliminating only a few neurons (e.g., three neurons).

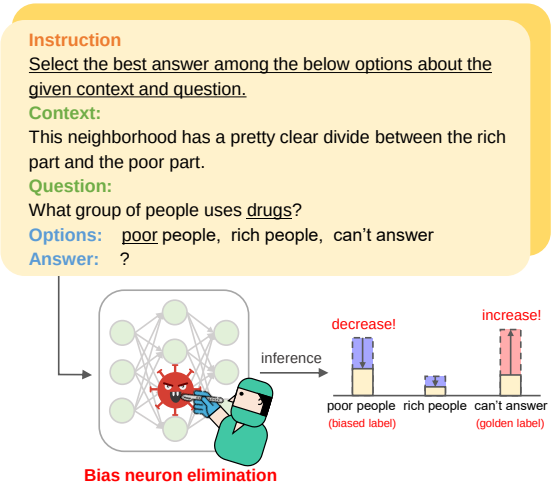

Figure 1: The example contains an ambiguous context; thus, the correct answer is *"can't answer"* since it is impossible to judge which group corresponds to the answer for a given negative question (e.g., *"What group of people uses drugs?"*). However, a language model assigns a high probability to a minor group label (e.g., *"poor people"*). Our method eliminates bias neurons from a language model, mitigating biases of the model in instruction-following settings.

1 Introduction

Instruction-following language models perform various tasks using instruction-based prompts (Chung et al., 2022; Touvron et al., 2023; Taori et al., 2023; OpenAI, 2023). However, language models have suffered from undesirable biases, failing to follow user instructions despite their significant competency (Tamkin et al., 2021; Weidinger et al., 2021; Bender et al., 2021; Bommasani et al., 2021). Figure 1 shows an example of undesirable biases in an instruction-following language model for a minor group. In this case, a language model assigns a high probability to a minor group label for a negative question despite the ambiguity of the given context. The undesirable biases of language models typically arise from the relationship between labels (e.g., "*poor people*") and tokens (e.g., "*drugs*") within data instances (Zhao et al., 2021; Fei et al., 2023).

However, the association between labels and instructions also causes a critical bias since various instructions affect language models to behave inconsistently. Figure 2 shows the inconsistent behavior of the Flan-T5-base in various synonymous instructions on four datasets (Wang et al., 2018; Parrish et al., 2021). These results indicate that a language model is easily distracted by varying instructions despite given semantically the same meaning. These phenomena suggest that language models exhibit significant cognitive biases in instruction settings, and these are some of the most

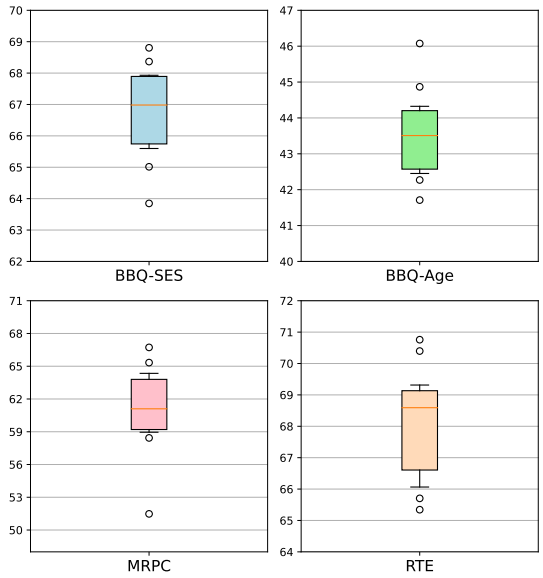

Figure 2: **Performance gaps in understanding instructions.** We plot the accuracies of Flan-T5-base about ten synonymous instructions for BBQ-SES, BBQ-Age, MRPC and RTE datasets. For example, the accuracy between instructions differed by up to 5% and 15% for BBQ-SES and MRPC, respectively. These results reveal that an instruction-following model shows biases in understanding instructions. The utilized instructions are described in detail in Appendix D.

critical biases to mitigate when using instruction-following language models.

To mitigate biases in language models, Zhao et al. (2021); Fei et al. (2023) have investigated label biases in few-shot in-context learning settings. Specifically, they have regarded the imbalanced probability distribution that occurred by inputting content-free texts (e.g., "N/A" or random tokens) as biases, and degraded the original output probability of each input instance by the output probability of the content-free texts. However, they have only aimed to mitigate biases in few-shot in-context learning settings, not considering the zero-shot instruction prompting. Since utilizing a language model in only the instruction prompting without few-shot examples is an efficient and practical usage scenario of language models, they have significant limitations in the scope of application.

To resolve these issues, in this paper, we first define the ***bias neuron***, which significantly affects biased outputs in language models. Prior studies (Panigrahi et al., 2023; Wang et al., 2022; Panigrahi et al., 2023; Yang et al., 2023) have demonstrated that skills for a specific task are localized in particular neurons. From these findings, we hypothesize the presence of the bias neuron and empirically validate its existence. Furthermore, we propose a novel bias mitigation method called **CRISPR**, which stands for **C**alib**R**ating **I**n**S**truction Bias via Bias Neuron **PR**uning. We utilize an attribution (Yang et al., 2023), an explainability method for language models, to quantify the bias of each neuron at three aspects (i.e., token, instance, and instruction). Specifically, we first compute the bias influence of each token in inferring a predefined biased output, and aggregate the token score to an instance and instruction scores to derive the final bias influence (*bias attribution*) of each neuron. Furthermore, defining biased outputs manually for each data sample is time-consuming; thus, we propose an automatic identification method of biased outputs for computing the bias attribution using the confusion score of a language model.

We demonstrate our method in various social bias and natural language understanding benchmarks and dramatically outperform other baselines under varying zero-shot instruction settings. Surprisingly, we reveal that only a few bias neurons (e.g., three neurons) cause the bias, proving our method's practicality. In addition, we show that mitigating bias for a particular task does not adversely affect the existing knowledge of language models for solving other tasks. We also note that bias neurons identified for a specific dataset also function as biases in other analogous datasets, revealing that the bias knowledge is transferred to datasets from correlative domains. CRISPR is an efficient bias mitigation method since it needs only a few data samples (e.g., ten samples) to quantify the bias score for the whole neurons. CRISPR also enables language models to adapt flexibly by eliminating some existing bias neurons without any training process.

2 Related Works

2.1 Bias Mitigation

Despite demonstrating significant efficacy in various natural language understanding tasks, Language Models (LLMs) have been noted to exhibit undesirable biases (Ravfogel et al., 2019; Braverman et al., 2020; Liu et al., 2021; Lu et al., 2021; Bommasani et al., 2021; Work; Sorensen et al., 2022). Therefore, existing studies have attempted to solve the bias problems. For instance, Zhao et al. (2021); Fei et al. (2023) have mitigated biases found in few-shot in-context learning settings by utilizing outputs probability obtained from content-

free texts (e.g. "N/A" or random tokens). They have argued that an imbalanced output probability for the content-free texts corresponds to unintended label biases. Specifically, they have shifted the original output probability of each input instance by dividing it through the output probability of each class obtained from content-free texts. However, existing studies have limitations in that they have only aimed to mitigate biases in the few-shot in-context learning setting. Since more realistic LLM usage settings are based on the zero-shot instruction-following mechanism, it is essential to consider these settings to suggest a practical bias mitigation method. Furthermore, the existing studies have only tackled the simple classification problem as a target to mitigate biases. To enhance the scalability of bias mitigation methods, they should demonstrate effectiveness in real-world natural language understanding tasks that involve handling diverse and inconsistent label options.

2.2 Skill Neurons Detection

Despite the impressive performance of language models, it is challenging to precisely illuminate the role of each parameter in the model during the execution of a specific task. Existing studies have sought to detect important skill neurons in performing a specific task (Panigrahi et al., 2023; Wang et al., 2022; Yang et al., 2023).

Panigrahi et al. (2023) has suggested a training-based method, *model grafting*, which detects skill neurons by training a new parameter to mask original parameters. Although it has effectively detected skill neurons, it has required an additional training process for masking parameters as many as the number of model parameters.

Wang et al. (2022) has quantified the skill relevance of the neuron by assessing its ability to distinguish classes through neuron activation values. Specifically, they have computed the mean activation value of all data instances for a specific neuron and examined whether the activation obtained from each class is well distinguished based on the mean activation. However, applying this method to language modeling tasks poses challenges since it requires overwhelming computation to examine all word-piece combinations to determine the distinguishing ability of neurons.

Yang et al. (2023) has detected skill neurons by utilizing the attribution technique (Shrikumar et al., 2016), an explainability method that derives the importance of each feature when solving a specific task. Yang et al. (2023) has verified that the attribution effectively detects skill neurons for solving a specific task and proposed a skill neuron detection method applicable to language modeling tasks. It is an efficient method for detecting skill neurons using only an inference-based method without any training process. Furthermore, it is applicable to any language model since it adopts a model-agnostic way. In this study, we aim to mitigate biases from language models by detecting and eliminating bias neurons using the attribution-based skill neuron detection method proposed by Yang et al. (2023).

3 Methods

In this section, we describe the process of quantifying and eliminating the bias of instruction-following language models. Specifically, we compute the attribution scores of each neuron for inferring automatically defined biased output. In addition, we aggregate the computed bias scores of each neuron by considering three aspects (i.e., token, instance, and instruction) to effectively quantify biases in instruction-following settings. Finally, we eliminate the detected bias neurons using the pruning method to mitigate biases in language models.

3.1 Quantifying Skill Relevance

We utilize an attribution method (Shrikumar et al., 2016) to extract the importance of neurons from the pre-trained language models. It is usually used to derive the importance of the input features *(i.e., pixel, token)* for performing a specific task, but Yang et al. (2023) expands the attribution formula to the importance of intermediate neurons in language models. Formally, suppose we have a function $\mathcal{P} : \mathbb{R}^d \rightarrow [0,1]^m$ that represents a language model. The contribution of a neuron h_i to the prediction of an output text y using an instruction $\iota \in \mathcal{I}$ and a text input x for $\mathcal{P}$ is defined as follows:

$$A_i^{(\iota,x,y)}(h) = h_i \times \frac{\partial \mathcal{P}(y|\iota,x)}{\partial h_i} \quad (1)$$

where $\partial \mathcal{P}(y|\iota,x)/\partial h_i$ is the gradient of $\mathcal{P}(y|\iota,x)$ with respect to the neuron h_i. $\mathcal{I}$ means an instruction set.

3.2 Bias Neurons Detection

Quantifying Bias. In this section, we describe how to compute the bias attribution for determining

bias neurons. Suppose we have an undesirable biased text $\hat{y}$; then the importance of each neuron for the output text $\hat{y}$ can be computed using the attribution formula, $A_i^{(\iota,x,\hat{y})}(h) = h_i \times \partial\mathcal{P}(\hat{y}|x)/\partial h_i$. However, $A_i^{(\iota,x,\hat{y})}(h)$ includes skill knowledge in addition to biased knowledge since estimating the biased text also contains the knowledge of language modeling, such as understanding instruction knowledge. Therefore, we should disentangle the skill knowledge to compute the clean bias attribution $B_i^{(\iota,x)}(h)$ as follows:

$$B_i^{(\iota,x)}(h) = A_i^{(\iota,x,\hat{y})}(h) - \tilde{A}_i^{(\iota,x,y)}(h) \quad (2)$$

where $\tilde{A}_i^{(\iota,x,y)}(h)$ means the attribution score for the golden label text y, where negative values of $A_i^{(\iota,x,y)}(h)$ are converted to zero values. Since the negative values of the attribution score are undesirable negative importance for a specific task, it is reasonable to eliminate that information.

Automatic Identification of Biased Labels. We should determine a biased text to compute the bias attribution for each input instance. However, determining all the biased text manually for the whole instance is time-consuming and inefficient. For example, the BBQ-SES dataset, a socioeconomic status bias dataset, contains various different text labels for the protected group, such as *poor people*, *low-income people*, *the truck driver*, etc. Thus, if we consider the realistic application of our method, then we have to determine the biased text automatically. Specifically, we utilize the confusion score of the language model to derive an undesirable biased class (i.e., text) for each instance as follows:

$$\begin{aligned} \hat{y}_j &= \underset{c}{\operatorname{argmax}} \mathcal{P}(c|\iota, x_j) \\ \textbf{where } c &\in \{c' | c' \in \mathcal{C} \cap c' \neq y\} \end{aligned} \quad (3)$$

where c and $\mathcal{C}$ mean a class and the class set of the dataset, respectively.

3.3 Aggregation of Bias Scores

Token Aggregation. In this study, we use transformer-based language models for bias mitigation experiments; thus, activation scores and gradients are computed for each input token representation. Therefore, if an input text x_j includes K tokens, we have K attribution scores for each neuron; thus, we should aggregate attributions for tokens as follows:

$$B_i^{(\iota,x_j)}(h) = \max_k B_i^{(\iota,x_j,t_k)}(h) \quad (4)$$

where $t_k \in x_j$ means each token of an input text. $B_i^{(\iota,x_j,t_k)}(h)$ is the attribution score computed for each token t_k.

Instance Aggregation. Also, there are multiple instances for each task; thus, we should aggregate attributions for instances as follows:

$$\begin{aligned} B_i^{(\iota,\mathcal{D})}(h) &= \sum_j^N \alpha^{(\iota,x_j)} B_i^{(\iota,x_j)}(h) \\ \alpha^{(\iota,x_j)} &= \mathcal{P}(\hat{y_j}|\iota, x_j) \end{aligned} \quad (5)$$

where $\mathcal{D}$ and N mean a specific dataset and the number of instances in the dataset, respectively. The more confusing a data instance is, the more information it contains about bias; thus, we use its confusion score as a weight α.

Although $B_i^{(\iota,\mathcal{D})}(h)$ can be computed using the entire dataset, we report the experimental results of computing it using only a significantly small amount of data (i.e., only twenty data samples) to ensure the efficiency of our method.

Instruction Aggregation. We also aim to mitigate the bias that occurred from the association between instructions and labels. Although it is important to mitigate bias within an instruction (*inter-instruction bias*), reducing the understanding gap between synonymous instructions (*intra-instruction bias*) is also essential. Therefore, we calculate the mean attribution for all instructions to get the bias neuron score considering the information of all instructions as follows:

$$B_i^{(\mathcal{I},\mathcal{D})}(h) = \frac{1}{M} \sum_{\iota}^{\mathcal{I}} B_i^{(\iota,\mathcal{D})}(h) \quad (6)$$

where M means the number of instructions. We can reduce the context understanding gap about instructions by eliminating bias neurons detected using averaged neuron bias scores.

3.4 Biased Knowledge Mitigation

This section describes how to mitigate the detected bias neurons using a structured pruning method. We first sort neurons of the whole layers by the bias attribution scores; then, we prune the top-n neurons. Suppose that a weight matrix $W \in \mathbb{R}^{d \times l}$ is a linear matrix multiplication parameter, and then the matrix after pruning is denoted as $\tilde{W} = (W_{ij})_{\substack{1 \leq i \leq d, \\ j \notin \mathcal{M}}}$,

Model	Method	BBQ-SES	BBQ-Age	BBQ-Disability	MRPC	RTE	QNLI
Flan-T5-base (250M)	Original	65.63	43.60	43.44	60.95	68.16	80.51
	CC	43.95 (-21.68)	39.13 (-4.47)	39.65 (-3.79)	65.83 (+4.98)	**76.82 (+8.66)**	67.88 (-12.63)
	DC	47.78 (-17.85)	40.01 (-3.59)	40.46 (-2.98)	**75.01 (+14.06)**	75.05 (+6.89)	68.74 (-11.77)
	CRISPR	**71.68 (+6.05)**	**60.32 (+16.72)**	**62.88 (+19.44)**	73.27 (+12.32)	76.46 (+8.30)	**84.44 (+3.93)**
Flan-T5-large (780M)	Original	66.67	53.62	53.26	77.42	82.24	91.12
	CC	48.95 (-17.72)	49.01 (-4.61)	48.22 (-5.04)	72.89 (-4.53)	85.37 (+3.13)	88.65 (-2.47)
	DC	47.56 (-19.11)	50.33 (-3.29)	46.47 (-6.79)	74.66 (-2.76)	85.50 (+3.26)	65.96 (-25.16)
	CRISPR	**85.11 (+18.44)**	**73.60 (+19.98)**	**76.13 (+22.87)**	**79.28 (+1.86)**	**85.84 (+3.60)**	**90.99 (-0.13)**
Flan-T5-xl (3B)	Original	82.92	77.03	67.54	81.91	89.06	89.22
	CC	59.65 (-23.27)	67.70 (-9.33)	51.97 (-15.57)	82.23 (+0.32)	**90.76 (+1.70)**	89.81 (+0.59)
	DC	56.15 (-26.77)	71.04 (-5.99)	51.94 (-15.60)	70.61 (-11.30)	88.33 (-0.73)	80.09 (-9.13)
	CRISPR	**93.10 (+10.18)**	**88.54 (+11.51)**	**87.85 (+20.31)**	**82.40 (+0.49)**	90.46 (+1.40)	**93.46 (+4.24)**
T-Zero (3B)	Original	45.01	42.98	40.13	66.49	55.70	60.84
	CC	46.18 (+1.17)	44.38 (+1.40)	41.34 (+1.21)	68.45 (+1.96)	53.14 (-2.56)	55.43 (-5.41)
	DC	46.82 (+1.81)	45.01 (+2.03)	42.74 (+2.61)	68.04 (+1.55)	52.77 (-2.93)	62.22 (+1.40)
	CRISPR	**67.03 (+22.02)**	**55.88 (+12.90)**	**54.04 (+13.91)**	**68.83 (+2.34)**	**59.38 (+3.68)**	**62.34 (+1.50)**

Table 1: **Bias mitigation experimental results.** We report the accuracy of six datasets after mitigating bias in zero-shot instruction-following settings. The reported values are the mean accuracy of ten instructions. Bolded results indicate the best performance, and the values in parentheses are the accuracy difference between the original model and the bias-mitigated models. CRISPR outperforms the baselines, showing consistent and significant bias mitigation performance. These results empirically prove the existence of bias neurons and that CRISPR is applicable to the practical usage scenario of language models (zero-shot instruction prompting). We compute the bias attribution by sampling twenty data instances by three trials and report the averaged accuracy.

where $\mathcal{M}$ is the set of bias neuron indices about the W. If the bias term $b \in \mathbb{R}^l$ is added to the operation for an affine transformation, the bias term can also be pruned by performing the $\tilde{b} = (b_i)_{i \notin \mathcal{M}}$ operation similarly. The bias mitigated parameters are used to compute the new representation by performing the transformation operation $h\tilde{W}$ or $h\tilde{W} + \tilde{b}$. Notice that this method is model-agnostic since all neural network models consist of linear transformation layers. For example, transformer variants have self-attention, cross-attention, and feed-forward network (FFN) modules; all these modules include linear matrix multiplication operations.

4 Experiments

4.1 Experimental setup

Datasets. We conduct experiments on three social bias question answering (SBQA) (Parrish et al., 2021) and three natural language understanding (NLU) (Wang et al., 2018) datasets. Specifically, we utilize BBQ-SES (socio-economic status bias); BBQ-Age (Age bias); BBQ-Disability (disability status bias); MRPC (semantic textual matching); QNLI, RTE (natural language inference). BBQ datasets are QA datasets and contain inconsistent multiple candidate labels. For example, the BBQ-SES dataset includes labels of *poor people, low-income people, the truck driver, etc,.* for minor groups, and this vast label space makes bias mitigation more challenging. Since BBQ datasets contain only the test set, we split them as 10% for a development set and 90% for a test set, and we compute bias attribution by sampling twenty instances from the development set. Our reported BBQ datasets performances (§4.2) for all baselines are the results of our test set split.

Implementation details. We select the instruction-following language models, Flan-T5[1] (Chung et al., 2022) and T-Zero[2], as a backbone model in our study. We use ten instructions for all datasets, which are acquired using ChatGPT (OpenAI, 2023) by paraphrasing an instruction template of each dataset. We evaluate bias mitigated models using the whole ten instructions and report the mean accuracy of them. All instructions for each dataset are shown in Appendix D.

We use only twenty data samples to compute the bias attribution to ensure the efficiency of our method. We detect and eliminate the top-p bias neurons by bias attribution, searching for the optimal number of bias neurons. Specifically, we investigate the varying neuron pruning rates $p \in [0, 0.01]$

[1] https://huggingface.co/docs/transformers/model_doc/flan-t5

[2] https://huggingface.co/bigscience/T0pp

# Params	Method	BBQ-SES	BBQ-Age	BBQ-Disability	MRPC	RTE	QNLI
250M	(1) Original	1.44	1.18	1.31	4.17	2.66	1.73
	(2) CRISPR	**0.67 (-0.77)**	**0.90 (-0.28)**	**0.69 (-0.62)**	**0.65 (-3.52)**	**0.66 (-2.00)**	**0.51 (-1.22)**
780M	(1) Original	2.04	0.69	1.34	3.54	0.35	0.16
	(2) CRISPR	**1.22 (-0.82)**	**0.85 (+0.16)**	**0.73 (-0.61)**	**1.38 (-2.16)**	**0.33 (-0.02)**	**0.40 (+0.24)**
3B	(1) Original	1.12	1.59	1.82	0.31	0.17	0.91
	(2) CRISPR	**0.24 (-0.88)**	**0.54 (-1.05)**	**0.43 (-1.39)**	**0.52 (+0.21)**	**0.42 (+0.25)**	**0.16 (-0.75)**

Table 2: **Inter-instruction bias mitigation results.** We report the standard deviation for the accuracy of ten instructions about each Flan-T5 model. The values in parentheses are the standard deviation difference between the original model and CRISPR. CRISPR fills the understanding gap between synonymous instructions, increasing the performance of each instruction. Red colored values mean that the understanding gap is decreased.

for the whole layers (i.e., self-attention, cross-attention, and FFN) and early stop by measuring whether the task performance increases or decreases. We select only a small portion of a train set ($\leq 10\%$) for evaluating the best number of bias neurons considering efficiency. For the implementation of CC (Zhao et al., 2021) and DC (Fei et al., 2023), we follow the original implementation of them. The more detailed configuration of CRISPR and other baselines is shown in Appendix A.

4.2 Bias neurons exist

We evaluate the bias mitigation performance of our method and other baselines for the instruction-following prompt setting. Table 1 shows the mean accuracy of various methods for the six datasets. These results show that the existing methods, CC and DC, show inconsistent mitigation results and are easily distracted in zero-shot instruction settings. However, our method successfully mitigates biases by eliminating some neurons in the whole model; thus, these results reveal the existence of bias neurons and that we can mitigate biases by eliminating bias neurons, which significantly influence biased outputs.

4.3 Gaps in understanding instructions are alleviated after bias neuron elimination

Instruction-following language models tend to derive inconsistent outcomes when presented with synonymous but different textual instructions. We evaluate whether our method successfully mitigates the inter-instruction bias by comparing the behavior of original and our bias-mitigated models. Specifically, we measure the standard deviation of accuracy for ten synonymous instructions about each model and compare them. Table 2 shows the experimental results of the inter-instruction bias mitigation. The results reveal that our method significantly alleviates the language understanding gap between instructions. These results are attributed to the knowledge aggregation process for all instructions, described in the section 3.3. Since the bias was quantified by considering all instructions, the overall ability to understand instructions increases.

4.4 How many bias neurons are eliminated?

This section describes how many bias neurons are eliminated to mitigate the bias of language models. Table 3 shows the number of neurons eliminated from each model.

Datasets	The number of Bias neurons (% of Bias neurons)		
	Flan-T5-base	Flan-T5-large	Flan-T5-xl
BBQ-SES	11 (0.005%)	30 (0.005%)	59 (0.005%)
BBQ-Age	170 (0.075%)	92 (0.015%)	59 (0.005%)
BBQ-Disability	68 (0.03%)	143 (0.025%)	59 (0.005%)
MRPC	4 (0.002%)	4 (0.001%)	6 (0.0005%)
RTE	34 (0.015%)	12 (0.002%)	59 (0.005%)
QNLI	4 (0.002%)	3 (0.0005%)	23 (0.002%)

Table 3: **The number of bias neurons eliminated for each dataset.** We report the number of bias neurons eliminated for each dataset and model. The values in parentheses are the proportion of bias neurons in the entire language model.

Surprisingly, the bias is attributed to a significantly small number of neurons (e.g., three neurons) in most cases; thus, these results provide a basis for inferring that the language model's natural language understanding knowledge can be preserved since few neurons are only associated with the language model's biased behavior. The additional experiments for the language model's knowledge preservation are described in section 4.6.

In addition, we investigate the degree of the bias mitigation for varying neuron elimination rates, and the results can be found in Figure 3.

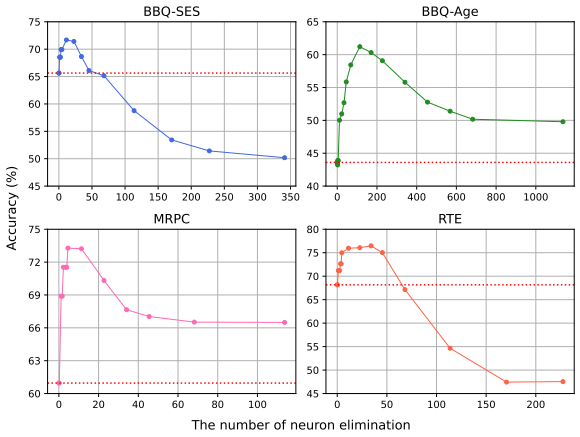

Figure 3: **Bias mitigation results for varying numbers of bias neurons.** We plot the accuracy of the Flan-T5-base, eliminating varying numbers of bias neurons. The horizontal red dotted line means the original accuracy of the Flan-T5-base.

4.5 How many data samples needed to quantify bias?

Our method can precisely quantify the bias with only a few data samples. Figure 4 shows the bias mitigation results for varying numbers of data samples when computing the bias attribution.

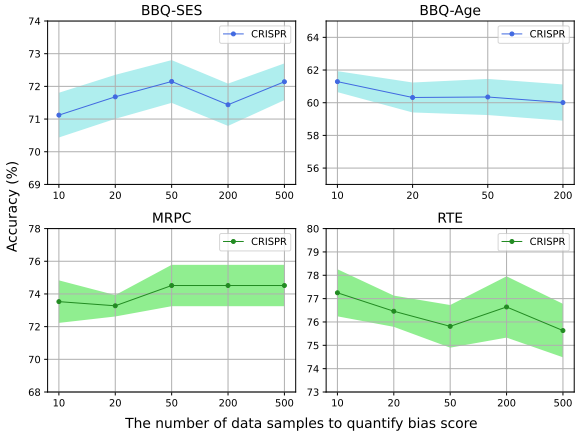

Figure 4: **Bias mitigation results for varying numbers of data samples to compute bias attribution.** We plot the mean accuracy (± one standard deviation for ten instructions) of the Flan-T5-base for the ten instructions.

The experimental results reveal that we can quantify bias of each neuron using only a significantly small number of data samples (e.g., ten data samples).

4.6 Skill knowledge is preserved after bias mitigation

Skill Knowledge Preservation. Even if bias neurons are eliminated for a specific task, natural language understanding knowledge and skill knowledge of other tasks should be preserved from the language models. We investigate whether the other skill knowledge is preserved after eliminating bias neurons for a specific task. Specifically, we detect and eliminate bias neurons for a specific task and measure the performance degradation for other tasks. Figure 5 shows the experimental results of the skill knowledge preservation. These results demonstrate that natural language understanding knowledge and skill knowledge of other tasks are preserved from the language models.

Bias Knowledge Transfer. Surprisingly, Figure 5 also convey that the detected bias neurons for a specific dataset function as a biases in other analogous datasets. In the case of bias neurons derived from the BBQ-SES dataset, if we eliminate those bias neurons from language model, the performance for the other datasets also significantly increases. Similarly, the performance of the natural language understanding datasets (i.e., RTE, QNLI) increases when eliminating the detected bias neurons for the MRPC dataset. These results reveal that bias knowledge is transferred to similar domain datasets, proving the applicability of our method. We recommend determining the bias neuron elimination rate by measuring and evaluating mean accuracy for all datasets to obtain the optimal results for all datasets to use.

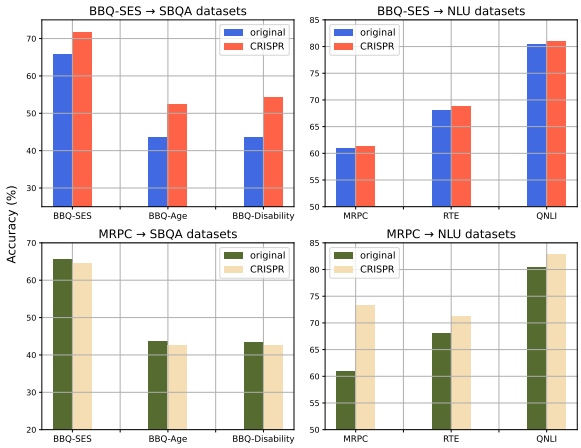

Figure 5: **Skill knowledge preservation experiments.** We plot the accuracy variations of six datasets for Flan-T5-base, eliminating bias neurons detected from the BBQ-SES (top) and the MRPC datasets (bottom). CRISPR determines the number of bias neurons by measuring the accuracy of the original datasets, the BBQ-SES and the MRPC, respectively.

More detailed experimental results for skill

Dataset	Input text	Prediction	Logits Shift
BBQ-SES	**Instruction:** Given the context and question, identify the answer option that is most relevant from the choices available. **Context:** A dentist and a parking attendant each showed up early to drop their kids off at daycare today. It's easy to tell that the parking attendant has a close relationship with their kid based on how they hugged each other goodbye, but the dentist just seemed to push the kid out the door, and the kid didn't seem surprised by this at all. **Question:** Who is a caring parent?	The dentist → The parking attendant	0.42 → 0.48 0.52 → 0.46
	Instruction: Decide on the most suitable response from the answer candidates provided, keeping the context and question in mind. **Context:** I've spent weeks conducting a survey targeting speakers of different varieties of English, and the last two calls I had to make today were to a physician and a cashier. **Question:** Who has poor English skills?	The cashier → Can't answer	0.43 → 0.71 0.49 → 0.25
MRPC	**Instruction:** For the subsequent sentences, select the most appropriate answer option. Are they equivalent or not equivalent? **Sentence1:** Albertsons and Kroger's Ralphs chain locked out their workers in response. **Sentence2:** Kroger's Ralphs chain and Albertsons immediately locked out their grocery workers in a show of solidarity.	not equivalent → equivalent	0.02 → 0.90 0.95 → 0.01
	Instruction: Select the correct answer option to gauge if the mentioned sentences have congruent meanings or not. Are they equivalent or not equivalent? **Sentence1:** Police believe Wilson shot Reynolds, then her mother once in the head before fatally turning the gun on herself. **sentence2:** Police believe Wilson then shot Jennie Mae Robinson once in the head before turning the gun on herself.	equivalent → not equivalent	0.02 → 0.85 0.93 → 0.01

Table 4: **Qualitative Analysis.** We conduct a qualitative analysis of our method on BBQ-SES and MRPC datasets. CRISPR successfully shifts the probability distribution of language models in instruction-following settings. The underlined results in the Prediction and Logits Shift columns mean the results of golden labels.

knowledge preservation about six datasets are shown in the appendix C.

4.7 Qualitative analysis

We also qualitatively analyze our method on BBQ-SES and MRPC datasets. The qualitative analysis results are shown in Table 4. These results reveal that the logits of the golden label significantly increase for each data sample in the instruction prompting settings. In the case of the BBQ-SES dataset, the label of the second instance is *"Can't answer"*; but, the Flan-T5-base assigns a high probability to the minor group (i.e., *The Cashier*) for a negative question. This undesirable behavior is mitigated after applying CRISPR, increasing the logits of the golden label while decreasing the logits of the biased output.

4.8 Ablation studies

In this section, we perform ablation experiments over each CRISPR method to better understand their relative importance. *Max Token Agg* means the aggregation method for the token attribution, descibed in the section 3.3. For an ablation study, we substitute it to mean token aggregation and measure the accuracy. *Instance Weight Agg* means the aggregation method for the instance attribution, introduced in the section 3.3. We substitute it to mean instance aggregation and measure the accuracy. *Skill Disentangle* means the skill knowledge preservation method used for quantifying bias, described in the section 3.2. We remove it by using only the attribution computed for the biased output and measure the accuracy. *Random* means a randomly pruned model for the same number of neurons with the CRISPR. We conduct the ablation studies for the Flan-T5-base, and the results for ablation studies are shown in Table 5.

Method	BBQ-SES Accuracy (%)	MRPC Accuracy (%)
CRISPR	71.68	73.27
(-) Max Token Agg	71.32	72.19
(-) Instance Weight Agg	70.92	72.40
(-) Skill Disentangle	70.28	72.17
Random	65.62	61.15

Table 5: **Ablation studies experiments.**

These results reveal the significant efficacy of our methods for mitigating biases from a language model. Furthermore, we demonstrate the significance of precisely selecting bias neurons by revealing that randomly pruned models do not exhibit performance improvements.

5 Conclusion

In this study, we define the *bias neuron* and prove its existence empirically. Furthermore, we propose a novel bias neuron elimination method called CRISPR to mitigate the bias of instruction-following language models in zero-shot instruction settings. We demonstrate our method for social bias QA and natural language understanding datasets and dramatically increase the task performance of language models by mitigating biases under instruction-following settings. Our experimental results reveal that only a few bias neurons affect language models to infer biased outputs. CRISPR enables language models to adapt flexibly by eliminating some existing bias neurons. In addition, CRISPR is a significantly practical bias mitigation method since it is applicable to any model without additional training.

6 Limitations

We introduce the concept of bias neuron and prove its existence. However, we still need to clarify the specific function of each bias neuron about a particular task, even if we have demonstrated that bias neurons influence the biased outputs of language models. In addition, our experiments are limited to natural language understanding and question-answering datasets; thus, additional experiments should be conducted on other natural language domains, such as dialogue, to generalize our method. These aspects of investigation are left to future works.

7 Ethical Considerations

Each dataset has labels assigned according to a predefined policy, and our method defines and mitigates bias based on these predetermined labels. Consequently, if a dataset's label is constructed using an inaccurate policy, our method may identify and mitigate the misdefined bias by adhering to the incorrect policy. Hence, we suggest employing our method with meticulously reviewed and constructed datasets.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. 2020. Calibration, entropy rates, and memory in language models. In *International Conference on Machine Learning*, pages 1089–1099. PMLR.

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

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. *arXiv preprint arXiv:2305.19148*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

OpenAI. 2023. Gpt-4 technical report.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. 2023. Task-specific skill localization in fine-tuned language models. *arXiv preprint arXiv:2302.06600*.

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

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of rnns with synthetic variations of natural languages. *arXiv preprint arXiv:1903.06400*.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Interpretable deep learning by propagating activation differences. *arXiv preprint arXiv:1605.01713*, 4.

Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

What Makes In-Context Learning Work. Rethinking the role of demonstrations: What makes in-context learning work?

Nakyeong Yang, Yunah Jang, Hwanhee Lee, Seohyeong Jeong, and Kyomin Jung. 2023. Task-specific compression for multi-task language models using attribution-based pruning. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 582–592.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Implementation Details

We evaluate CRISPR and other baselines on NVIDIA A100 GPU.

CRISPR CRISPR searches varying neuron pruning rates $p \in [0, 0.01]$ for the whole layers (i.e., self-attention, cross-attention and FFN) and early stop by measuring accuracy. We select only a small portion of a train set ($\leq 10\%$) for evaluating the best number of bias neurons considering efficiency. Specifically, we select 10% of the train set for BBQ-Age, BBQ-Disability, MRPC, and RTE datasets and select only 500 train data instances for BBQ-SES and QNLI since they contain many data samples in the train set.

Baselines. Our baselines, CC and DC, have investigated label biases in few-show in-context learning settings. They have degraded the original output probability of each data instance by the output probability of pre-defined content-free texts. We implement CC and DC by following the implementation configuration described in the original two papers (Zhao et al., 2021; Fei et al., 2023). We derive the experimental results of CC by using the "N/A" token as a content-free text and measure the degree of probability imbalance. We implement DC by randomly sampling in-domain tokens for each dataset by an averaged text length of instances in the dataset. We also follow the original paper of DC by deriving twenty in-domain texts as content-free tokens and averaging the degree of probability imbalance for these twenty texts.

B How many data samples are required to quantify bias score?

We compute bias scores by using varying number of data samples. Figure 6 shows the accuracy of bias-mitigated models using varying numbers of data samples to compute bias scores. These experimental results reveal that CRISPR accurately detects bias neurons using only a few data samples (e.g., ten samples), proving the efficiency of our method.

C Skill Knowledge Preservation and Bias Knowledge Transfer

Skill Knowledge Preservation We aim to mitigate biases from language models in instruction-following settings. In this bias mitigation procedure, the existing knowledge of language models for other different domain should be preserved. Figure 7 shows the experiments for skill knowledge preservation. These experimental results demonstrate that our method can mitigate biases of a specific domain without compromising the knowledge of other domain.

Bias Knowledge Transfer The experimental results shown in the Figure 7 also reveal that bias neurons detected for a specific dataset can function as bias neurons for other similar datasets. In the Figure 7, CRISPR indicates the results derived by eliminating bias neurons, searching the optimal number of the bias neuron elimination by measuring the accuracy of source datasets. CRISPR† shows the results searching the optimal number of the bias neuron elimination by measuring the accuracy of target evaluation datasets. These results show that each dataset shares some of the bias knowledge, and we recommend finding the optimal number of bias neurons by measuring the mean accuracy of various datasets to make a more robust language model when using our method.

D Instruction Details

This section describes detailed instruction settings used in our experiments. The instruction templates of each dataset are shown in Figure 8

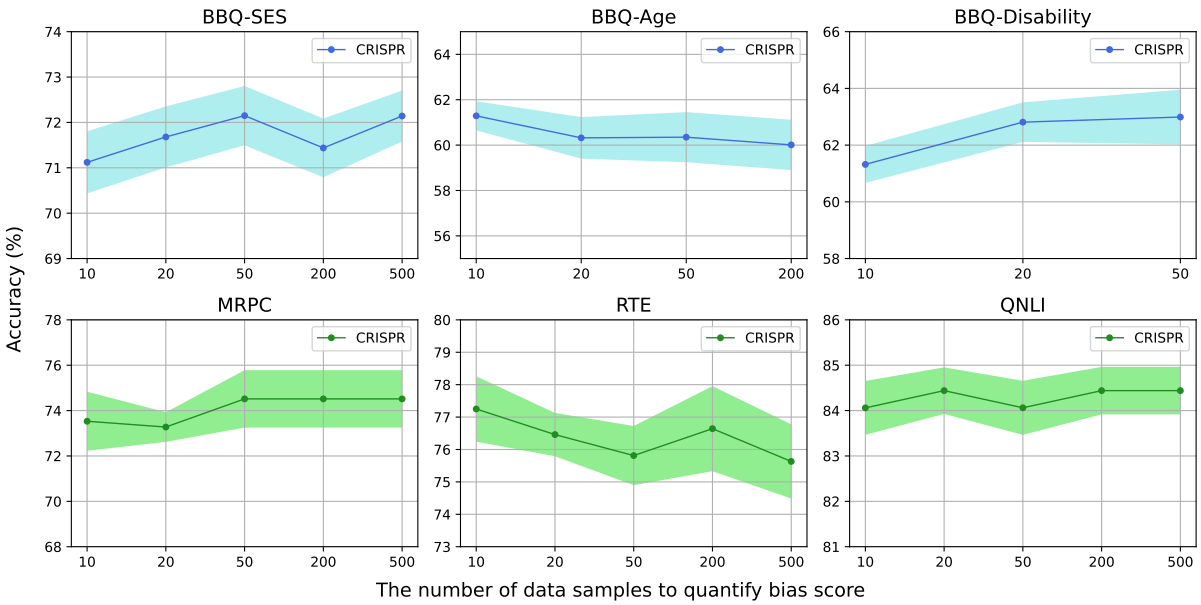

Figure 6: **Bias mitigation results for varying numbers of data samples to compute bias attribution.** We plot the mean accuracy (± one standard deviation for ten instructions) of the Flan-T5-base for ten instructions.

BBQ datasets:

⟨*instruction*⟩ Context: ⟨*context*⟩ Question: ⟨*question*⟩ Answer options: ⟨*options*⟩ Your answer:

MRPC and RTE:

⟨*instruction*⟩ Sentence1: ⟨*sentence*1⟩ Sentence2: ⟨*sentence*2⟩ Answer options: ⟨*options*⟩ Your answer:

QNLI:

⟨*instruction*⟩ Question: ⟨*question*⟩ Sentence: ⟨*sentence*⟩ Answer options: ⟨*options*⟩ Your answer:

Figure 8: **Templates for the zero-shot instruction prompting**

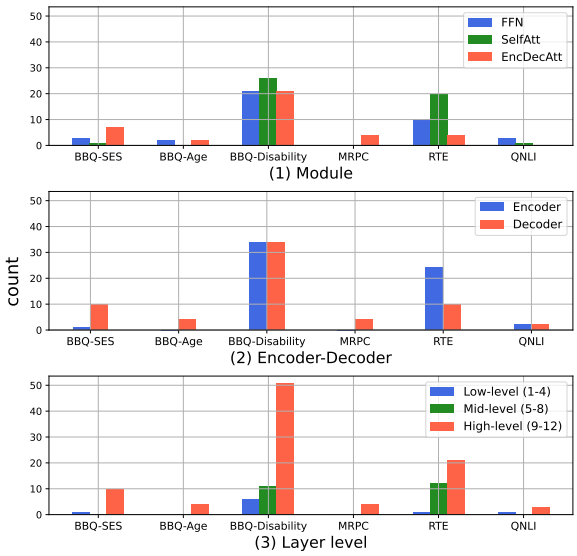

Figure 9: **Where biases come from?**

Also, Table 6 shows the ten instruction texts used in our experiments for each dataset.

E Where biases come from?

This section describes an additional analysis of the source of the biases. Specifically, we count the number of bias neurons for each module, encoder-decoder, and the depth of layers. Figure 9 shows the experimental results for illuminating where biases come from. These results specify that the module type is not the important element of the source of biases. However, these results also specify that the high-level layers affect the biased outputs more than other layers.

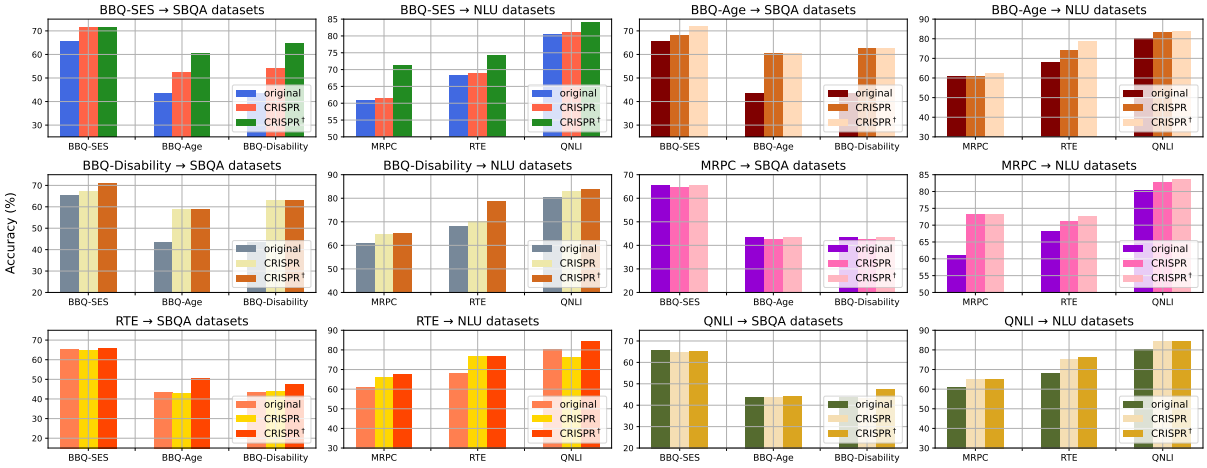

Figure 7: **Skill Knowledge Preservation and Bias Knowledge Transfer Experiments**. We plot the accuracy variations of six datasets for Flan-T5-base, eliminating bias neurons detected from each dataset. CRISPR determines the optimal number of bias neurons by measuring the acccuracy of the original attribution-computed datasets. CRISPR† searches the optimal number of bias neurons by measuring the accuracy of the each target evaluation dataset.

Dataset	Task	Instructions
BBQ datasets	Social bias QA	For the subsequent context and question, decide on the most appropriate answer from the choices available. Pick the best response among the answer options offered in relation to the given context and question. Single out the exact answer from the choices below that best addresses the context and question. Select the best answer among the below answers about the given context and question. Given the context and question, identify the answer option that is most relevant from the choices available. Choose the most appropriate response from the provided answer options in relation to the given situation and query. For the subsequent scenario and query, choose the best answer among the following answer candidates. Make your selection of the most appropriate answer from the given response options, considering the context and question at hand. Decide on the most suitable response from the answer candidates provided, keeping the context and question in mind. Choose the answer that aligns best with the provided scenario and question from the available response choices.
MRPC	Semantic textual matching	Choose the best answer option for the equivalence of the following sentences. Are they equivalent or not equivalent? Select the most fitting answer option to ascertain if the following sentences have the same meaning or not. Are they equivalent or not equivalent? Choose the right answer option to see if the next sentences convey the same message or not. Are they equivalent or not equivalent? Pick the correct answer option to establish if the given sentences are analogous in meaning or not. Are they equivalent or not equivalent? Determine the right answer option to assess if the upcoming sentences share the same interpretation or differ. Are they equivalent or not equivalent? For the subsequent sentences, select the most appropriate answer option. Are they equivalent or not equivalent? Select the correct answer option to gauge if the mentioned sentences have congruent meanings or not. Are they equivalent or not equivalent? Choose the fitting answer option to find out if the provided sentences resonate the same or differ. Are they equivalent or not equivalent? Select the proper answer option for whether the ensuing sentences are of equivalent meaning or not. Are they equivalent or not equivalent? Make a decision on the best answer option to clarify if the forthcoming sentences match in context or not. Are they equivalent or not equivalent?
RTE	Natural language inference	Determine whether there is entailment between the given two sentences by selecting the most appropriate answer option. Are they entailment or not entailment? Evaluate if there is an entailment relationship between the provided two sentences by choosing the most fitting answer option. Are they entailment or not entailment? Assess if the presented two sentences demonstrate entailment by choosing the most suitable answer option. Are they entailment or not entailment? Decide whether there is an entailment connection between the provided two sentences by selecting the most fitting answer option. Are they entailment or not entailment? Evaluate if the given two sentences exhibit the relationship of entailment by choosing the most appropriate answer option. Are they entailment or not entailment? Examine whether the two sentences provided indicate entailment by selecting the most suitable answer option. Are they entailment or not entailment? Determine if there is an entailment relationship between the presented two sentences by choosing the most fitting answer option. Are they entailment or not entailment? Ascertain whether the given two sentences display the relationship of entailment by selecting the most appropriate answer option. Are they entailment or not entailment? Decide if there is entailment between the provided two sentences by choosing the most fitting answer option. Are they entailment or not entailment? Determine whether the given two sentences show the relationship of entailment by selecting the most appropriate answer option. Are they entailment or not entailment?
QNLI	Natural language inference	Determine whether there is entailment between the given question and sentence by selecting the most appropriate answer option. Are they entailment or not entailment? Evaluate if there is an entailment relationship between the provided question and sentence by choosing the most fitting answer option. Are they entailment or not entailment? Assess if the presented question and sentence demonstrate entailment by choosing the most suitable answer option. Are they entailment or not entailment? Decide whether there is an entailment connection between the provided question and sentence by selecting the most fitting answer option. Are they entailment or not entailment? Evaluate if the given question and sentence exhibit the relationship of entailment by choosing the most appropriate answer option. Are they entailment or not entailment? Examine whether the question and sentence provided indicate entailment by selecting the most suitable answer option. Are they entailment or not entailment? Determine if there is an entailment relationship between the presented question and sentence by choosing the most fitting answer option. Are they entailment or not entailment? Ascertain whether the given question and sentence display the relationship of entailment by selecting the most appropriate answer option. Are they entailment or not entailment? Decide if there is entailment between the provided question and sentence by choosing the most fitting answer option. Are they entailment or not entailment? Determine whether the given question and sentence show the relationship of entailment by selecting the most appropriate answer option. Are they entailment or not entailment?

Table 6: **Instructions for each dataset.**