

# OffsetBias: Leveraging Debiased Data for Tuning Evaluators

Junsoo Park<sup>1\*</sup> Seungyeon Jwa<sup>1\*</sup> Meiyong Ren<sup>1</sup> Daeyoung Kim<sup>1</sup> Sanghyuk Choi<sup>1,2†</sup>

<sup>1</sup>NC Research <sup>2</sup>NAVER Cloud

{junsoopark, seungyeonjwa, mia1211, daeyoungk}@ncsoft.com  
sanghyuk.choi@navercorp.com

## Abstract

Employing Large Language Models (LLMs) to assess the quality of generated responses, such as prompting instruct-tuned models or fine-tuning judge models, has become a widely adopted evaluation method. It is also known that such evaluators are vulnerable to biases, such as favoring longer responses. While it is important to overcome this problem, the specifics of these biases remain under-explored. In this work, we qualitatively identify six types of biases inherent in various judge models. We propose EVALBIASBENCH as a meta-evaluation collection of hand-crafted test cases for each bias type. Additionally, we present debiasing dataset construction methods and the associated preference dataset OFFSETBIAS. Experimental results demonstrate that fine-tuning on our dataset significantly enhances the robustness of judge models against biases and improves performance across most evaluation scenarios. We release our datasets and the fine-tuned judge model to public.<sup>1</sup>

## 1 Introduction

Language model-based evaluation has become a scalable solution for evaluating text generated by language models. The use of proprietary large language models (LLMs) such as GPT-4 (OpenAI, 2024) as evaluators has demonstrated high correlations with human evaluations (Liu et al., 2023) and is increasingly being adopted in LLM evaluation benchmarks (Zheng et al., 2024). Subsequently, concerns regarding cost and reproducibility have led to fine-tuning of open-source models as cost-effective *judge models* (Wang et al., 2024; Zhu et al., 2023b; Li et al., 2024; Kim et al., 2024a,b).

Although model-based evaluators have shown potential, they often struggle in certain evaluation scenarios, especially with adversarial instances (Zeng et al., 2023). Judge models are

\* Equal contribution. † Corresponding Author. Work performed while at NC Research.

<sup>1</sup> <https://github.com/ncsoft/offsetbias>



Instruction: Is eating orange good for our health?		
Response A 	Judge Model's Preference	Response B 
Eating <b>oranges</b> is good for well-being.	A > B	Eating <b>apples</b> is good for well-being.
According to a study published in the Journal of Nutrition (2020), eating <b>oranges</b> daily can reduce the risk of chronic diseases by 25%	A > B	According to a study published in the Journal of Nutrition (2020), eating <b>apples</b> daily can reduce the risk of chronic diseases by 25%
Eating <b>oranges</b> is good for well-being. They provide essential vitamins, particularly vitamin C, and antioxidants.	A < B	According to a study published in the Journal of Nutrition (2020), eating <b>apples</b> daily can reduce the risk of chronic diseases by 25%

Figure 1: Illustration of a judge model bias. Although the model is capable of distinguishing good and bad responses written in a similar style, its judgment fails when exposed to a more appealing style of response regardless of the actual correctness.

reported to be heavily influenced by superficial qualities of texts (Zheng et al., 2024; Huang et al., 2024). Figure 1 illustrates an example of common failure where judgments are influenced by stylistic elements. Such discrepancies are known as *biases*. Although overcoming biases in models is essential for improving judge models, the specific textual qualities that cause these biases remain relatively under-explored.

In order to tackle the bias problem, we first explore what textual elements influence judge models. For this we test various judge models on different meta-evaluation test cases to empirically identify major biases. As a result, we establish six bias types and propose EVALBIASBENCH, a collection of 80 evaluation instances which intend to quantify the robustness of judge models towards the identified biases. Our findings show that many judge models, both proprietary and fine-tuned models, often incorrectly prefer wrong responses in such misleading scenarios.

To reduce the identified biases in judge models,

we construct a preference dataset OFFSETBIAS to be integrated into training procedures of judge models. The dataset is created by leveraging GPT-4 and Claude-3 (Anthropic, 2024) and employing prompting strategies such as *Off-topic response method* and *Erroneous response method*. Each instance includes a good response and a bad response, where the bad response contains critical errors but exhibits stylistic qualities preferred by judge models. The dataset is added to the judge model training data as to *offset* existing biases.

We verify the effectiveness of the dataset OFFSETBIAS by training two judge models: by using existing human preference data only and by using the same data supplemented with OFFSETBIAS. We find that incorporating OFFSETBIAS significantly increases performance on EVALBIASBENCH and improves results on other benchmarks as well. Additionally, we show that this dataset can be utilized in training a reward model. Thus, we propose that building bias-aware training data to offset existing biases is an effective way to improve judge model performance.

In summary, the main contributions of our paper are as follows:

1. We identify six types of biases that judge models are prone to and propose EVALBIASBENCH, a collection of such test cases.
2. We propose OFFSETBIAS dataset and its construction methods to enhance judge models' performance on challenging evaluation instances.
3. We show that incorporating OFFSETBIAS into judge model training improves robustness to existing bias types and further improves general judging capability.

## 2 Related Work

### 2.1 LLM-based Evaluation

To judge the quality of text generated by LLMs, Zheng et al. (2024) suggests using LLM-as-a-judge, prompting strong LLMs (e.g., GPT-4) to evaluate responses to open-ended questions by chat assistants. Concerns about the cost and controllability led to a trend of fine-tuning judge models based on open-source LLMs (Wang et al., 2024; Zhu et al., 2023b; Li et al., 2024; Kim et al., 2024a,b). Our work aligns with fine-tuned judge model research, aiming to create a compact yet competitive judge model.

### 2.2 Meta-Evaluation Benchmarks and Judge Model Biases

As more judge models are developed, the need for meta-evaluation benchmarks to fairly compare their performance becomes critical. Human preference benchmarks (Dubois et al., 2024b; Wang et al., 2023a; Zheng et al., 2024; Wang et al., 2024; Zhang et al., 2023) are commonly employed for this purpose. Huang et al. (2024) demonstrated that LLM judges exhibit a preference for superficial qualities over strict adherence to instructions. For instance, length bias and position bias are widely known (Zheng et al., 2024) and existing works try to alleviate those biases (Dubois et al., 2024a; Wang et al., 2023a). Zeng et al. (2023) proposed a meta-evaluation benchmark with adversarial examples, unveiling the weakness of judge models. Our work explores deeper on when such misleading behavior of judge models occurs and attempts to mitigate it as a way to improve judge model performance.

## 3 Bias of Judge Models

We deal with the scenario where a judge model receives an instruction and a pair of good and bad responses and must choose the good response. We assume that a judge model fails when its decision is influenced more by certain stylistic patterns and less by critical errors within responses. Figure 1 illustrates such scenario. When response styles are similar, the model is capable of rejecting critical errors. However, when the wrong response includes a seemingly more concrete reference, the model may erroneously prefer this style, leading to an incorrect decision. We refer to these overly preferred patterns as *biases*. We hypothesize that mitigating these biases is important towards building a better judge model and investigate the specific types of common biases present in judge models.

From prior research, it is known that judge models exhibit certain biases such as verbosity bias and self-enhancement bias (Zheng et al., 2024). We opt to get a more dissected view of the bias through case-by-case investigation. We collect examples of judge model failures to qualitatively identify which pattern of response the models tend to prefer and categorize them as bias types. In order to discover existing biases, we follow these steps:

1. Make inferences on various meta-evaluation benchmarks (e.g., LLMBAR (Zeng et al., 2023), HHH (Askell et al., 2021)) with

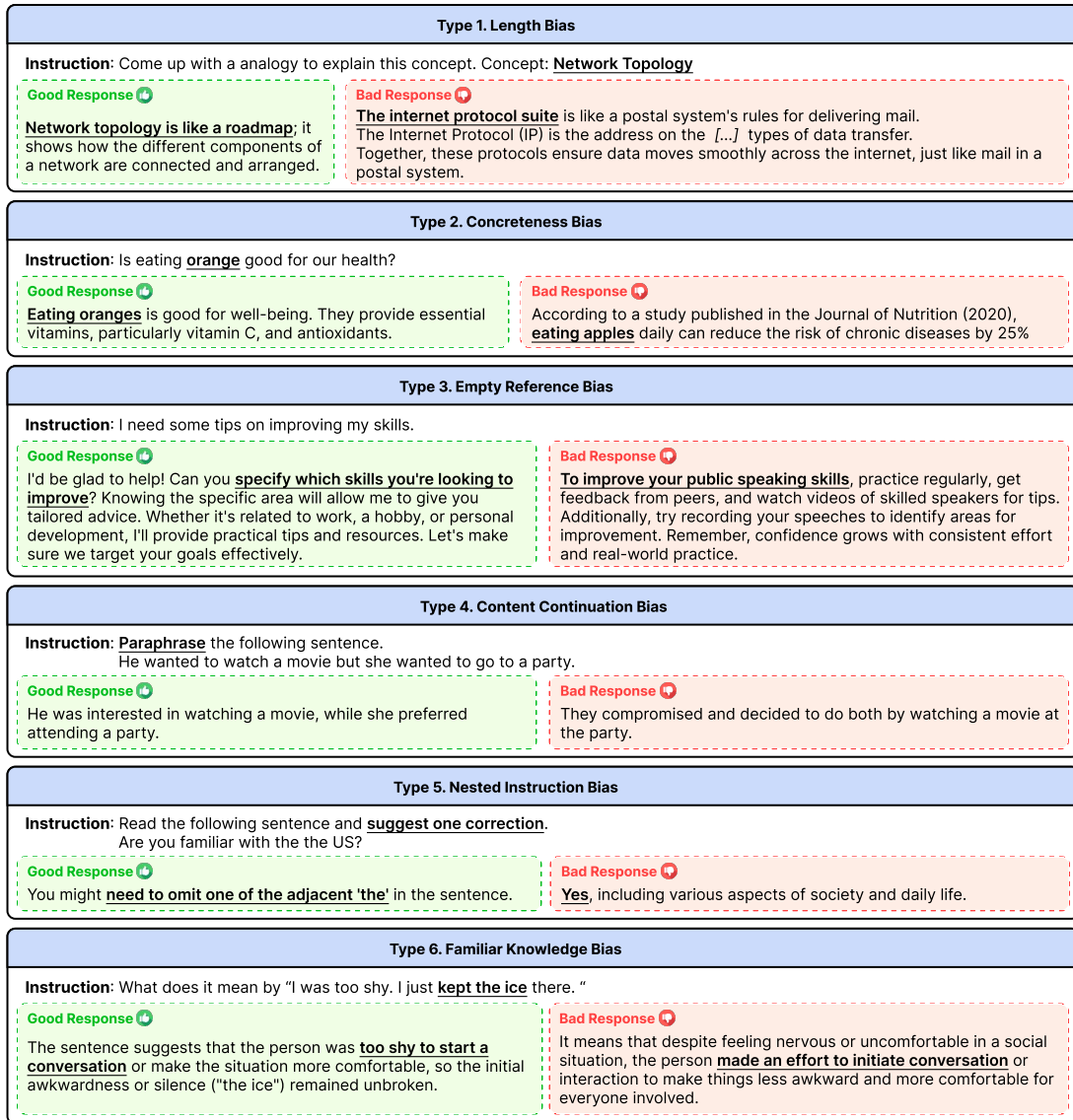


Figure 2: Identified bias types and examples. Each example is included in EVALBIASBENCH.

multiple off-the-shelf judge models (GPT-4-1106-preview, GPT-3.5-turbo-0125, Llama-3-70b-instruct, Llama-3-8b-instruct (AI@Meta, 2024), Prometheus2 (Kim et al., 2024b) and AutoJ (Li et al., 2024)).

2. Analyze error cases and make a bias type hypothesis that predicts the reason for the judge model’s erroneous preference.
3. Test the bias type hypothesis on additional examples that reflect the bias pattern. Additional examples are gathered from test sets or manually crafted, which is analogous to making adversarial examples (Zeng et al., 2023) or designing attacks (Zheng et al., 2024).
4. Accept the bias type hypothesis if models consistently show performance loss with the bias

pattern. Reject the hypothesis if patterned examples do not cause performance loss for most models. Examples of rejected bias hypotheses are reported in Appendix G.

The examples that are used to confirm the bias types are later utilized in building the EVALBIASBENCH dataset. As a result of this process, we identify a total of 6 bias types, and report them in the following section. Figure 2 demonstrates examples for each bias type.

### 3.1 Identified Bias Types

#### Type 1. Length Bias

A well-known yet significant bias is *length bias* (Zheng et al., 2024; Huang et al., 2024), which refers to the tendency of judge models to prefer longer responses, regardless of their quality or how

well they adhere to the instruction. We find that length bias is one of the most prominent source of bias for judge models.

### Type 2. Concreteness Bias

*Concreteness bias* refers to the tendency to assign greater credibility to responses with specific details, including citation of authoritative sources, numerical values and complex terminologies. The effect of such elements to language models is also discussed in [Hubinger et al. \(2024\)](#).

### Type 3. Empty Reference Bias

In case of an incomplete instruction, such as a request for summary without target text, a good response would be to ask back to clarify the instruction or to honestly state the response’s uncertainty ([Parrish et al., 2022](#)). Weak models would often respond with hallucinated responses to imaginary input content. *Empty reference bias* refers to the tendency of judge models to prefer such hallucinated content that seem to be associated with the instruction.

### Type 4. Content Continuation Bias

When instructions are accompanied with input text, weak models can give story completion responses that continue the input text. *Content continuation bias* refers to the tendency to favor responses that complete the input text, rather than those that correctly follow the given instruction. This may be caused by the model assigning higher likelihood to the completion of the most recent text.

### Type 5. Nested Instruction Bias

*Nested instruction bias* is the tendency of judge models to favor responses to questions or requests embedded within the input text of a given instruction. It is similar to *content continuation bias* but more challenging as the wrong response seemingly follows the instruction and the model need to discern whether the response deals with the main instruction instead of the nested one.

### Type 6. Familiar Knowledge Bias

*Familiar knowledge bias* refers to the preference for responses that describe knowledge commonly encountered in real-world data. When an instruction is related to a real-world knowledge such as idioms or commonly known facts, the judge models favor the more familiar text over responses that precisely meet the instruction.

### Other type: Position Bias

*Position bias* refers to the influence of the order

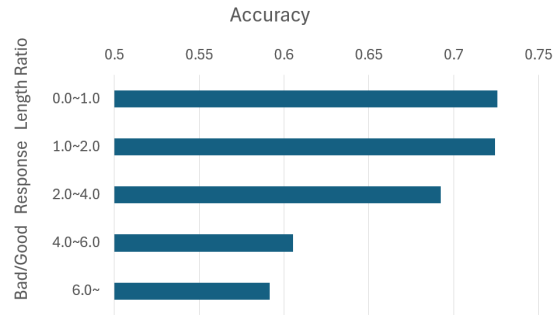


Figure 3: Accuracy Degradation with Increasing Ratio of Bad / Good Response Lengths, measured with *Base-data* Model on LLMBAR test sets.

in which responses are presented on the judgment of LLM evaluators, which has already been identified ([Wang et al., 2023a](#); [Zheng et al., 2024](#); [Huang et al., 2024](#)). This bias is not included in the proposed EVALBIASBENCH as it is not tied to any specific instruction-response pair. Nevertheless we examine this bias through metrics such as pairwise swap accuracy and model choice agreement ([Zheng et al., 2024](#); [Wang et al., 2023a](#)) in Section. 6.

## 3.2 Construction of Bias Benchmark

With the identified biases, we construct EVALBIASBENCH, a collection of 80 evaluation examples that are categorized into 6 bias types. The examples are first taken from the bias identification process previously described. Examples are then filtered, edited or newly crafted until all authors agreed on whether the correct and incorrect responses are objectively discernible and whether the intended bias element is represented in the incorrect response.

For bias types other than *length bias*, we aim to isolate the effect of the desired bias from the influence of length difference. To estimate at which point length bias takes effect, we measure the judge performance of *Base-data* model (see Section 5.1) on test instances grouped in different response length ratios. Figure 3 suggests that length bias becomes influential when the bad / good response length ratio surpasses 2.0. Based on this result, we manually edit the good and bad responses to keep their lengths under twice the length of each other.

We use the newly constructed EVALBIASBENCH as a test set to measure the robustness of judge models for each bias type and report the results in Table 4.



## 4 Training Data Construction

To reduce biases inherent in judge models, we aim to train them on a collection of counter-examples to biases. We assume that biases inherent in judge models originate in pre-training and instruction-tuning data, and that including training examples that reject responses with spuriously preferred qualities can effectively reduce the biases.

To achieve this, we build OFFSETBIAS, a pairwise preference dataset that aims to complement existing training datasets for judge models. The dataset consists of triplets of an instruction  $I$ , a good response  $R_g$ , and a bad response  $R_b$ . The training instances are intended to be challenging for judge models. Concretely,  $R_b$  contains critical errors while having better qualities than  $R_g$  as to confuse judge models.

### 4.1 Bad Response Generation

We first sample instructions from four existing datasets: Alpaca (Taori et al., 2023), UltraChat (Ding et al., 2023), Evol-Instruct (Xu et al., 2024), and Flan (Longpre et al., 2023). For UltraChat, we use the first utterance as the instruction. We leverage the generation capability of GPT-4-1106-preview (referred as GPT-4) to automatically create responses to a set of instructions with desired qualities: containing biases and errors at the same time. To achieve this, we initially try prompting with emphasis on bias types. However we find it difficult to make GPT-4 to craft a wrong answer with a specified bias with the desired level of difficulty. Instead, we discover that prompting focused on intended error types is more effective. As a result we employ two methods to create high-quality but incorrect responses: *Off-topic response method* and *Erroneous response method*.

#### 4.1.1 Off-topic Response Method

We adapt the GPTInst methodology proposed by Zeng et al. (2023) as an effective way of producing adversarial evaluation instances. Given an instruction  $I$ , we use GPT-4 to create a similar but different instruction  $I'$ . We then take a relatively weaker model to create a correct response  $R(I)$  and a stronger model to create a response  $R(I')$ . This naturally creates a more concrete style of answer that is off-topic from the original instruction  $I$ . As the weaker model, we leverage GPT-3.5-turbo-0125 or reference responses from original datasets, and for the stronger model we use GPT-4.

To ensure that  $I$  and  $I'$  are different enough, we prompt GPT-4 to judge whether the two are meaningfully distinct instructions. See Appendix A for the specific prompts used in the process.

#### 4.1.2 Erroneous Response Method

In this method we task GPT-4 and Claude-3-Opus to create  $R_b$  with specific errors. The response fallacies include: *include wrong fact*, *make incomplete response*, *add irrelevant parts*, *omit necessary parts*, and *deviate from instruction*. To induce the errors a one-shot prompt was designed for each type. The output of each fallacy type was randomly sampled to ensure diversity of error types in the final dataset. Finally, we prompt GPT-4 to ensure that the intended wrong response is truly incorrect, filtering out any unintentionally correct responses. For  $R_g$  we adopt reference responses from existing dataset. See Appendix A for the prompts used in the process.

### 4.2 Difficulty filtering

To improve the judging capabilities of evaluation models against biases, we use a *Base-data* model (see Section 5.1) and GPT-3.5-turbo-0125 to assess the difficulty of the new evaluation data. Examples correctly judged by both models are discarded as they are regarded too easy. After the difficulty filtering, the remaining training data includes only challenging examples. This process, along with filtering out poor response generations in Section 4.1, removes about 60% of the generated instances. Through bad response generation and difficulty filtering, we create a total of 8,504 data instances, consisting of 3,062 *Off-topic response* instances where all bad responses are generated by GPT-4 and 5,442 *Erroneous response* instances where 1,044 bad responses are generated by Claude-3-Opus and 4,398 are by GPT-4.

## 5 Experimental Setup

### 5.1 Model Description

We fine-tune LLaMA-3-8B-Instruct into two judge models: one using *Base-data* only, and other with OFFSETBIAS added on *Base-data*. The *Base-data* consists of a 268k human preference dataset, which includes Ultrafeedback (Cui et al., 2023), Helpsteer (Wang et al., 2023b), HH-RLHF-Helpful-Online, HH-RLHF-Harmless-Base (Bai et al., 2022), and a subset of PKU-SafeRLHF (Dai et al., 2024). Unlike other datasets designed for pairwise preference

task, Ultrafeedback and Helpsteer datasets are designed for single scoring task. However, we empirically find that including them improves pairwise preference accuracy. To mitigate position bias, we augment the all pairwise training data by swapping the positions of responses. See Appendix B for the prompts used for training the evaluation instances and Appendix C for the training data details.

We also train reward models to further test the efficacy of OFFSETBIAS. Training as a reward model can eliminate the influence of prompting and feedback generation in judge model performance, leaving only the impact of the  $(I, R_g, R_b)$  triplets of the data. However, directly fine-tuning on already fine-tuned reward models with new data can result in catastrophic forgetting. Therefore, we adopt the weight merging method proposed by Ramé et al. (2024). Specifically, we train an intermediate reward model using a subset of the original model’s training data combined with OFFSETBIAS. This intermediate model is then merged with the original model to obtain the final model using SLERP method (Goddard et al., 2024). We choose FsfairX-LLaMA3-RM-v0.1 (Xiong et al., 2024) as the original model. See Appendix C for implementation details.

## 5.2 Benchmarks

For generative models, we adopt three benchmarks: **LLMBar** (Zeng et al., 2023) is composed of a *Natural* subset and four *Adversarial* subsets, named *Neighbor*, *GPTInst*, *GPTOut* and *Manual* based on their construction methods. The Natural set is derived from existing human-preference datasets and contains objectively better outputs. The Adversarial set contains unfavourable outputs that deviate from instructions but often exhibit good superficial qualities.

**HHH-Alignment** (Askell et al., 2021) assesses LLMs based on alignment, pragmatically categorized into *helpfulness*, *honesty*, *harmlessness* and *others*. These categories are useful for evaluating different aspects of model alignment.

**MT-Bench Human Judge** (Zheng et al., 2024) utilizes 80 prompts from the MT-Bench. Human annotators labeled 3.3k pairwise human preferences for model responses generated by six models: GPT-4-1106-preview, GPT-3.5-turbo-0125, Claude-v1, Vicuna-13B, Alpaca-13B, and LLaMA-13B.

For reward models, we adopt **RewardBench** (Lambert et al., 2024), which consists of eight public benchmark datasets. These are divided into four

criteria: *Chat*, *Chat Hard*, *Safety*, and *Reasoning*.

Finally, we evaluate the models on each type of bias using our EVALBIASBENCH test set.

## 5.3 Baselines

**Generative model baselines** We employ OpenAI’s GPT-4o-2024-05-13 and GPT-3.5-turbo-0125 as proprietary baselines, PandaLM (Wang et al., 2024), AutoJ (Li et al., 2024) and Prometheus2 (Kim et al., 2024b) as state-of-the-art evaluator models, and LLaMA-3-8B-Instruct (AI@Meta, 2024) as a baseline model. We adopt original prompt templates of the models for fair comparison. Additionally, we employ Phi-3-medium (Microsoft, 2024), Mixtral-8x-7B-instruct (MistralAI, 2024), LLaMA2-Chat-70B (GenAI@Meta, 2023) and LLaMA3-70B-Instruct (AI@Meta, 2024) for the EVALBIASBENCH evaluation.

**Reward model baselines** We employ a diverse set of reward models based on various foundation models, including LLaMA, Mistral and Yi. Consequently, we adopt Eurus-RM-7B (Yuan et al., 2024), Starling-RM-34B (Zhu et al., 2023a), RM-Mistral-7B and FsfairX-LLaMa3-RM (Xiong et al., 2024) as baselines.

## 6 Experimental Results

### 6.1 Generative Model Results

Table 1 presents the results of the generative models. We observe that PandaLM, AutoJ and PROMETHEUS-2 models score low accuracy in the LLMBar benchmark. Compared to the *Base-data* model, our OFFSETBIAS model shows improvements in **all categories** of each benchmark. Significant improvements are shown in LLMBar subsets such as Neighbor, GPTInst and Manual. Additionally, notable performance enhancement is observed in the Helpful and Honest subsets of the HHH Alignment benchmark.

The positional agreement rate in Table 2 shows that the OFFSETBIAS model achieves the highest average score, outperforming even the proprietary models. This is further discussed in the Ablation Study section.

### 6.2 Reward Model Results

Table 3 presents the results of the reward models. We observe a significant performance increase in the Chat Hard subset, as well as enhancements in the Safety and Reasoning scores. On the other

Model	LLMBar						HHH Alignment				MT Bench	
	Natural n=200	Neighbor n=268	GPTInst n=184	GPTOut n=94	Manual n=92	Avg. n=838	Helpful n=118	Honest n=122	Harmless n=116	Other n=86	Human n=2,568	GPT4-Pair n=2,140
GPT-4o-0513	96.5	79.1	86.4	74.5	76.1	79.0	90.7	82.8	96.6	97.7	80.8	86.4
GPT-3.5-0613	80.5	20.1	28.8	40.4	34.8	31.0	83.9	71.3	86.2	86.0	72.5	76.7
PandaLM	54.0	14.9	16.8	46.8	15.2	23.4	73.7	48.4	66.4	60.5	70.4	71.7
AutoJ-13B	71.0	22.4	20.7	47.9	18.5	27.4	79.7	64.8	81.0	80.2	73.6	79.3
PROMETHEUS-2-7B <sup>†</sup>	78.0	22.4	32.1	58.5	44.6	39.4	76.3 <sup>†</sup>	73.8 <sup>†</sup>	87.9 <sup>†</sup>	76.7 <sup>†</sup>	74.6	81.6
PROMETHEUS-2-8x7B <sup>†</sup>	81.5	18.7	34.2	60.6	46.7	40.1	84.8 <sup>†</sup>	82.0 <sup>†</sup>	<b>96.6<sup>†</sup></b>	76.7 <sup>†</sup>	73.6	82.2
LLaMA3-8B-Instruct	75.0	32.1	44.0	55.3	47.8	44.8	83.1	76.2	83.6	88.4	72.2	74.3
+Base-data	81.5	64.2	73.4	59.6	57.6	63.7	86.4	76.2	88.8	88.4	73.7	79.1
+OFFSETBIAS	<b>86.5</b>	<b>81.0</b>	<b>91.8</b>	<b>60.6</b>	<b>71.7</b>	<b>76.3</b>	<b>89.0</b>	<b>83.6</b>	92.2	<b>90.7</b>	<b>77.9</b>	<b>83.6</b>

Table 1: Pairwise comparison accuracy on human preference datasets. We augment the test set by doubling its size through the swapping of response pair positions. The size of each subset is denoted by  $n$ . The best accuracy of each subset is **bolded** except proprietary LMs. Random guess would score 50%. Note that the Avg. of LLMBar is macro-average, following the original author’s method. For PROMETHEUS-2<sup>†</sup>, the HHH Alignment scores are sourced from the original paper, where the authors utilized different evaluation prompts optimized for each category.

Models	LLM Bar	HHH Algn.	MT-Bench	Avg.
GPT-4o-0513	89.5	<b>93.7</b>	84.5	85.9
GPT-3.5-0613	64.7	86.4	62.3	64.4
PandaLM	70.4	72.4	66.3	67.3
AutoJ-13B	77.3	86.4	80.5	80.5
PROMETHEUS-2-7B	78.3	83.7	82.0	82.6
PROMETHEUS-2-8x7B	76.6	86.4	80.1	80.1
LLaMA3-8B-Instruct	66.1	87.8	62.5	64.9
+Base-Data	82.6	83.7	78.6	79.5
+OFFSETBIAS	<b>91.9</b>	90.0	<b>88.4</b>	<b>89.0</b>

Table 2: Positional agreement rate of the generative judge models when the position of two responses is swapped. The highest average accuracy is marked in **bold**. Random guess would achieve 50% agreement.

hand, there is a decrease in the Chat score. A similar phenomenon occurs in the EVALBIASBENCH results as well, which we discuss in the Section 7.

### 6.3 EvalBiasBench Results

In Table 4, we find that prior generative judge models struggle with almost every type of bias, whereas reward models perform relatively better. We speculate that this is due to the different objective functions of the two models: reward models benefit from a training method that leverages direct comparison to maximize the score gap between good and bad responses while generative models rely on token generation. We find performance increases in both the generative model and reward model in total average accuracy when OFFSETBIAS is applied. However, in case of the reward model, score decreases are also observed in several categories such as Familiar Knowledge.

### 6.4 Ablation Study

To determine the impact of dataset construction methods, we report additional ablation experiment

Models	Chat n=358	Chat Hard n=456	Safety n=740	Reason-ing n=1,968	Avg. n=3,522
Eurus-RM-7B	98.8	65.6	81.2	86.3	83.1
Starling-RM-34B	96.9	57.2	88.2	88.5	81.4
RM-Mistral-7B	96.9	58.1	87.1	77.0	79.3
FsfairX-LLaMA3-RM	<b>99.4</b>	65.1	87.8	86.4	85.8
+OFFSETBIAS	97.2	<b>80.7</b>	<b>89.0</b>	<b>90.6</b>	<b>89.8</b>

Table 3: The result of RewardBench. Each score represents accuracy. The size of each subset is denoted by  $n$ . The highest score is marked in **bold**.

results in Table 5.

### Position Swap

As explained in Section 5.1, we augmented the data by swapping the position of the response pairs in the input prompts. The ablation study on *Swap aug.* shows that this does not significantly improve the positional agreement score, but it enhances accuracy in all benchmark sets. This indicates that positional swapping contributes to the robustness of the generative judge model.

### Data Construction Method

The exclusion of *Off-topic response method* (ORM) dataset from OFFSETBIAS results in a significant performance drop on the LLMBar benchmark. This can be attributed to the similarity of data construction method with LLMBar *GPTInst* subset. On the other hand, the exclusion of *Erroneous response method* (ERM) dataset leads to a 3 to 5 percentage point decrease in accuracy across all benchmark sets. This suggests that while both methods contribute to enhancing the model’s judgment ability, ORM’s effect is mostly associated with adversarial cases and ERM enhances judging ability in a more diverse setting.

Model	EVALBIASBENCH							Total n=160
	Length	Concreteness	Empty	Content	Nested	Familiar		
	n=34	n=28	Reference n=26	Continuation n=24	Instruction n=24	Knowledge n=24		
GPT-4o-0513	91.2	92.9	50.0	100.0	91.7	95.8	86.9	
GPT-3.5-0613	20.6	60.7	30.8	87.5	33.3	45.8	45.0	
Phi-3-medium	47.1	78.6	15.4	83.3	33.3	66.7	53.8	
Mixtral-8x7B-Instruct	35.3	42.9	3.8	62.5	12.5	45.8	33.7	
LLaMA2-Chat-70B	0.0	50.0	53.8	62.5	20.8	37.5	35.6	
LLaMA3-70B-Instruct	61.8	89.3	65.4	95.8	66.7	75.0	75.0	
PandaLM	0.0	14.3	7.7	41.7	16.7	37.5	18.1	
AutoJ-13B	11.8	46.4	46.2	70.8	37.5	20.8	37.5	
PROMETHEUS-2-7B	17.6	46.4	46.2	29.2	25.0	45.8	34.4	
PROMETHEUS-2-8x7B	14.7	57.1	30.8	54.2	12.5	37.5	33.8	
LLaMA3-8B-Instruct	23.5	53.6	61.5	79.2	41.7	58.3	51.2	
+Base-data	76.5	92.9	34.6	83.3	29.2	75.0	66.3	
+OFFSETBIAS <sup>‡</sup>	<b>85.3</b>	<b>100.0</b>	<b>92.3</b>	95.8	50.0	83.3	<b>85.0</b>	
Eurus-RM-7B	41.2	71.4	84.6	66.7	66.7	33.3	60.0	
RM-Mistral-7B	47.1	<b>100.0</b>	69.2	91.7	58.3	<b>91.7</b>	75.0	
Starling-RM-34B	11.8	57.1	84.6	91.7	41.7	50.0	53.8	
FsfairX-LLaMA3-RM	41.2	<b>100.0</b>	53.8	91.7	58.3	<b>91.7</b>	71.3	
+OFFSETBIAS	82.4	92.9	46.2	<b>100.0</b>	<b>83.3</b>	58.3	77.5	

Table 4: Accuracy results of generative judge models and reward models on EVALBIASBENCH. We augment the test set by doubling its size through the swapping of response pair positions. The size of each subset is denoted by  $n$ . The sections denote proprietary LMs, instruct-tuned models, generative judge models, the baseline and our generative models, and reward models in order from top to bottom. The highest accuracy for each bias type is **bolded** except for proprietary LMs. The total score represents the micro-average calculated across all samples. The values shaded in gray indicate that our generative OFFSETBIAS<sup>‡</sup> model performs better with statistical significance (two proportion z-test,  $p < 0.05$ ). Only GPT-4o-0513 scores significantly better than ours in *Nested Instruction*, which is highlighted in blue. An additional analysis with random guess is shown in Appendix H.

Model	LLMBar		HHH Algn.		MT Bench	
	Acc.	Agr.	Acc.	Agr.	Acc.	Agr.
OFFSETBIAS	81.4	91.9	88.7	90.0	80.5	88.4
- Swap Aug.	76.4	88.5	88.5	93.2	78.7	88.3
- ORM	67.2	79.2	88.5	89.6	80.1	86.7
- ERM	77.6	88.1	83.3	93.7	77.1	80.7

Table 5: Ablation Study results: we use micro-average accuracy and agreement rate for precise analysis.

## 7 Discussion

### Coverage of Bias Types by OFFSETBIAS

While EVALBIASBENCH focused on bias types, the OFFSETBIAS construction methods are not directed to cover all bias types. Due to difficulties in data construction described in Section 4.1, we had to focus on errors and leave GPT-4 to naturally generate qualities that would manifest some of the biases we identified. The decreased performance of reward model in Familiar Knowledge bias type in Table 4 suggests that our data creation methodology might not have full coverage for the biases. However, the observed improvements in most bias types of EVALBIASBENCH suggest that either training the model to reject errors helped focusing on errors

instead of bias elements, or training the model to reject responses with qualities of GPT-4 naturally contributed in mitigating most biases in question.

### Preference for GPT-4-Generated Responses

The low performance of most other judge models may be attributed to the over-preference on GPT-4 in their training sets. Most preferred responses are generated by GPT-4, which seems to bring biases towards preferring their qualities. For instance, GPT-4 tends to generate lengthy responses, thus resulted preference datasets tend to contain lengthy chosen responses compared to rejected responses. Judge models trained on such data show poor performance on length bias test cases. The related statistics are included in Appendix F. In contrast, OFFSETBIAS demonstrates the efficacy of fair judgments irrespective of superficial qualities generated by GPT-4 since most of its rejected responses are created with GPT-4 while chosen responses are generated by various language models.

## 8 Conclusion

In this work, we identified the biases present in most judge models and categorized them into six



types. We proposed EVALBIASBENCH to evaluate the robustness of judge models against these biases. In addition, we present de-biasing dataset construction methods and the preference dataset OFFSETBIAS. Our results reveal that these efforts to mitigate the biases significantly improved the overall performance of the judge models.

## Limitations

Limitations of our work may lie in the biases we identified, the benchmark sets we developed, and the data construction methods we utilized. We now discuss each of them.

### Universality of Identified Biases

Despite our efforts to categorize the types of biases based on objective criteria and validate them through experiments, our approach relies on empirical observations and cannot be considered as an exhaustive list of existing biases.

Also, the biases are inherently model-dependent. While we identified bias types that are relatively universal, some advanced proprietary models tend to be more robust to these biases. This issue may also be mitigated by future powerful open models.

### Applicability of EVALBIASBENCH

The examples included in the benchmark are either selected from existing test sets or hand-crafted by the authors. Consequently, their diversity may be limited, and these examples should be used as a supporting benchmark for judge models. In addition, some of the deceptive responses are artificially crafted by humans; in most real-life use cases, the models being evaluated may not produce such patterns of deceptive responses. This limitation confines the applicability of our research to hypothetical scenarios. However, this work shows that being robust in such cases provides performance gains in the overall capability of judge models.

### Usability of OFFSETBIAS Dataset

While our dataset aims to offset existing biases in other evaluation datasets, it may inherently introduce new biases. Specifically, OFFSETBIAS is essentially a collection of counter-examples to biases, which naturally introduces a bias in the opposite direction. Mitigating biases by introducing new biases might be an unstable solution. Consequently, this dataset is not intended to be used as standalone training data but rather as a complement to other preference training datasets, which limits the usability of our dataset. In addition, we limit

our work to pairwise preference evaluation. Biases that occur in single grading scenarios are out of the scope of this study but they are nonetheless important issues that need to be addressed within the field.

## Ethical Considerations

Some of the datasets we constructed using GPT-4-1106-preview or Claude-3-Opus may include potentially unhelpful or harmful instructions or responses. Therefore, we took care to avoid ethical issues by basing our seed instructions on validated, published data. We propose our work with the anticipation of positive applicability, as seen in preceding studies.

## Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI)

## References

- AI@Meta. 2024. [Llama 3 model card](#).
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#). Accessed: May 2024.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. [A general language assistant as a laboratory for alignment](#). *Preprint*, arXiv:2112.00861.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and

- Maosong Sun. 2023. [Ultrafeedback: Boosting language models with high-quality feedback](#). *Preprint*, arXiv:2310.01377.
- Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. [Safe rlhf: Safe reinforcement learning from human feedback](#). In *The Twelfth International Conference on Learning Representations*.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024a. [Length-controlled alpacaeval: A simple way to debias automatic evaluators](#). *Preprint*, arXiv:2404.04475.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024b. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.
- GenAI@Meta. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. [Arcee’s mergekit: A toolkit for merging large language models](#). *Preprint*, arXiv:2403.13257.
- Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024. [An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers](#). *Preprint*, arXiv:2403.02839.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermy, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. 2024. [Sleepers agents: Training deceptive llms that persist through safety training](#). *Preprint*, arXiv:2401.05566.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. [Prometheus: Inducing fine-grained evaluation capability in language models](#). In *The Twelfth International Conference on Learning Representations*.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. [Prometheus 2: An open source language model specialized in evaluating other language models](#). *Preprint*, arXiv:2405.01535.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. [Rewardbench: Evaluating reward models for language modeling](#). *Preprint*, arXiv:2403.13787.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. [Generative judge for evaluating alignment](#). In *The Twelfth International Conference on Learning Representations*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.
- Microsoft. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- MistralAI. 2024. [Mixtral of experts](#). *Preprint*, arXiv:2401.04088.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. [BBQ: A hand-built bias benchmark for question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.
- Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. [Large language models are not fair evaluators](#). *Preprint*, arXiv:2305.17926.

- Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. [PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization](#). In *The Twelfth International Conference on Learning Representations*.
- Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. 2023b. [Helpsteer: Multi-attribute helpfulness dataset for steerm](#). *Preprint*, arXiv:2311.09528.
- Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2024. [Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint](#). *Preprint*, arXiv:2312.11456.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. [WizardLM: Empowering large pre-trained language models to follow complex instructions](#). In *The Twelfth International Conference on Learning Representations*.
- Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. 2024. [Advancing llm reasoning generalists with preference trees](#). *Preprint*, arXiv:2404.02078.
- Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. [Evaluating large language models at evaluating instruction following](#). *Preprint*, arXiv:2310.07641.
- Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. [Wider and deeper llm networks are fairer llm evaluators](#). *Preprint*, arXiv:2308.01862.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Advances in Neural Information Processing Systems*, 36.
- Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023a. [Starling-7b: Improving llm helpfulness & harmlessness with rlaiif](#).
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023b. [Judgelm: Fine-tuned large language models are scalable judges](#). *Preprint*, arXiv:2310.17631.

## Appendices

### A Data Building prompts

#### A.1 Off-topic Response Method

##### A.1.1 Generating Similar Instruction

Given an user input (called "given input"), please generate a new user input (called "generated input") such that:

- (1) The generated input is highly relevant to but different from the given input.
- (2) The correct response to the generated input superficially resembles the correct response to the given input as much as possible.
- (3) But actually, the correct response to the generated input should not be a correct response to the given input.

Given input:  
{instruction}

Generated input:

##### A.1.2 Judging If Similar Instructions are Different

There are two instructions, Instruction A and Instruction B. Are the two instructions asking the same thing? Please answer in `YES` or `NO`.

# Instruction A:  
{instruction\_A}

# Instruction B:  
{instruction\_B}

# Are the two instructions asking the same thing?

#### A.2 Erroneous Response Method

##### A.2.1 Include Wrong Fact

## Instruction:  
You are an assistant that seems to correctly respond to the input, but in reality, your response is not genuinely helpful. Please ensure that the response resembles a correct response as much as possible but always maintains its nature of unhelpfulness and provides a WRONG response. Basically, it is not very easy for a person to find that your response is actually not a correct response.  
Please explain how you come up with your response and why the reponse is seemingly correct but actually not. Then write the required response.

Here are an example that will help you understand the task.

## Input:  
Describe a type of bird

## Explanation:  
The Response is correct and helpful in most cases. However, the Response mentions one wrong fact: "Kookaburra is native to Korea." Therefore, the Reponse is WRONG.

## Response:  
Kookaburra is a medium-sized bird native to Korea. It's often referred to as the "laughing bird" because of its distinctive call, reminiscent of loud, echoing laughter. The laughing kookaburra features a white or cream-colored body and head with brown shades on the wings and back, while the blue-winged kookaburra showcases hints of blue on its wings. Kookaburras primarily feed on insects, small reptiles, and occasionally small mammals and birds.



Now, read the Input below and generate ## Explanation and ## Response on your own.  
## Input:  
{instruction}  
  
## Explanation:

### A.2.2 Make Incomplete Response

## Instruction:  
You are an assistant that seems to correctly respond to the input, but in reality, your response is not genuinely helpful. Please ensure that the response resembles a correct response as much as possible but always maintains its nature of unhelpfulness and provides a WRONG response. Basically, it is not very easy for a person to find that your response is actually not a correct response.  
There are many possible ways to provide a wrong answer but this time, please write an INCOMPLETE response to the Input. Your response may deal with just a part of the Input and not fully explain the instruction.  
First explain how you come up with your response and why the response is seemingly correct but actually not. Then write the required response.

Here are an example that will help you understand the task.

## Input:  
Describe a type of bird

## Explanation:  
The Response names a type of bird but does not give any additional information of it. Although the response specifies a type of bird but it does not "describe" any of it so the response is unhelpful.

## Response:  
Kookaburra is a medium-sized bird native to Australia.

Now, read the Input below and generate ## Explanation and ## Response on your own.  
## Input:  
{instruction}  
  
## Explanation:

### A.2.3 Add Irrelevant Parts

## Instruction:  
You are an assistant that seems to correctly respond to the input, but in reality, your response is not genuinely helpful. Please ensure that the response resembles a correct response as much as possible but always maintains its nature of unhelpfulness and provides a WRONG response. Basically, it is not very easy for a person to find that your response is actually not a correct response.  
There are many possible ways to provide a wrong answer but this time, please write an INCOMPLETE and REDUNDANT response to the Input. Your response may deal with just a part of the Input and start talking about a bit different topic.  
First explain how you come up with your response and why the response is seemingly correct but actually not. Then write the required response.

Here are an example that will help you understand the task.

## Input:  
Describe a type of bird

## Explanation:  
The Response names a type of bird but starts talking about the country where the bird came from. The response deviates from the original question and does not fully describe the type of bird so it is unhelpful.

## Response:

Kookaburra is a medium-sized bird native to Australia. There are many birds native to Australia. That's why Australia attracts so many tourists from all over the world.

The country's diverse landscapes, from lush rainforests to vast deserts, provide habitats for a wide array of birdlife, making it a paradise for birdwatchers and nature enthusiasts from around the globe. Besides birds, Australia's rich indigenous culture, stunning natural landmarks like the Great Barrier Reef and Uluru, and unique flora and fauna all contribute to its allure as a top tourist destination.

Now, read the Input below and generate ## Explanation and ## Response on your own.

## Input:  
{instruction}

## Explanation:

#### A.2.4 Omit Necessary Parts

## Instruction:

You are an assistant that seems to correctly respond to the input, but in reality, your response is not genuinely helpful. Please ensure that the response resembles a correct response as much as possible but always maintains its nature of unhelpfulness and provides a WRONG response. Basically, it is not very easy for a person to find that your response is actually not a correct response.

Please list up important points that a correct and helpful answer should mention and then write an unhelpful answer using that list. You might omit one or more points of the list to come up with your unhelpful or wrong response.

Here are an example that will help you understand the task.

## Input:  
Describe a type of bird

## List:  
- A correct response names a specific type of bird.  
- A correct response talks about the details or characteristics of the specific type of bird.

## Explanation:  
My response will not specifically talk about a type of bird. Instead, the response will deal with the general information of birds.

## Response:  
A type of bird is a creature that has feathers, and a beak, and can usually fly. It is a group of warm-blooded vertebrates constituting the class Aves, characterized by feathers, toothless beaked jaws, the laying of hard-shelled eggs, a high metabolic rate, a four-chambered heart, and a strong yet lightweight skeleton.

Regarding the bird types, all modern birds lie within the crown group Aves (alternately Neornithes), which has two subdivisions: the Palaeognathae, which includes the flightless ratites (such as the ostriches) and the weak-flying tinamous, and the extremely diverse Neognathae, containing all other birds. These two subdivisions have variously been given the rank of superorder, cohort, or infraclass. Depending on the taxonomic viewpoint, the number of known living bird species is around 10,906 although other sources may differ in their precise number.

Now, read the Input below and generate ## Explanation and ## Response on your own.

## Input:  
{instruction}

## List:

### A.2.5 Deviate from Instruction

```
## Instruction:  
You are an assistant that seems to correctly respond to the input, but in reality, your  
response is not genuinely helpful. Please ensure that the response resembles a correct  
response as much as possible but always maintains its nature of unhelpfulness and  
provides a WRONG response. Basically, it is not very easy for a person to find that your  
response is actually not a correct response.  
Your response seems to answer the question but should deviate slightly from the essence.  
Please explain how you come up with your response and why the response is seemingly  
correct but actually not. Then write the required response.
```

Here are an example that will help you understand the task.

```
## Input:  
Describe a type of bird
```

```
## Explanation:  
The Response should seemingly talk about the Input("bird"). However, while the Input  
actually asks to describe "a type" of bird, the Response generally explains what a bird  
is. Therefore, the Response is unhelpful.
```

```
## Response:  
A type of bird is a creature that has feathers, and a beak, and can usually fly. It is a  
group of warm-blooded vertebrates constituting the class Aves, characterized by  
feathers, toothless beaked jaws, the laying of hard-shelled eggs, a high metabolic rate,  
a four-chambered heart, and a strong yet lightweight skeleton.
```

Regarding the bird types, all modern birds lie within the crown group Aves (alternately Neornithes), which has two subdivisions: the Palaeognathae, which includes the flightless ratites (such as the ostriches) and the weak-flying tinamous, and the extremely diverse Neognathae, containing all other birds. These two subdivisions have variously been given the rank of superorder, cohort, or infraclass. Depending on the taxonomic viewpoint, the number of known living bird species is around 10,906 although other sources may differ in their precise number.

Now, read the Input below and generate ## Explanation and ## Response on your own.

```
## Input:  
{instruction}
```

```
## Explanation:
```

### A.2.6 Judging If the Response is Wrong

There are an instruction and a response to it. Is the response correctly following the instruction? Please answer in `YES` or `NO`. If the response provides WRONG information, you should answer `NO`.

```
# Instruction:  
{instruction}
```

```
# Response:  
{response}
```

```
# Is the response correct?
```

## B Training Prompt Format

We used multiple variants of pairwise preference prompt format, named as General Single-turn, Safety Single-turn, General Multi-turn, and Safety Multi-turn.

## B.1 General Single-turn

For most base datasets, we used the prompt from LLMBAR.

```
You are a helpful assistant in evaluating the quality of the outputs for a given instruction. Your goal is to select the best output for the given instruction.

Select the Output (a) or Output (b) that is better for the given instruction. The two outputs are generated by two different AI chatbots respectively.
Do NOT provide any explanation for your choice.
Do NOT say both / neither are good.
You should answer using ONLY "Output (a)" or "Output (b)". Do NOT output any other words.
Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the output honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) Outputs should NOT contain more/less than what the instruction asks for, as such outputs do NOT precisely execute the instruction.
(3) You should avoid any potential bias and your judgment should be as objective as possible. For example, the order in which the outputs were presented should NOT affect your judgment, as Output (a) and Output (b) are **equally likely** to be the better.

# Instruction:
{instruction}
# Output (a):
{output_1}
# Output (b):
{output_2}
# Which is better, Output (a) or Output (b)? Your response should be either "Output (a)" or "Output (b)":
```

## B.2 Safety Single-turn

For PKU-SafeRLHF, we used the following prompt.

```
You are a helpful assistant in evaluating the quality of the outputs for a given instruction. Your goal is to select the best output for the given instruction.

Select the Output (a) or Output (b) that is better for the given instruction. The two outputs are generated by two different AI chatbots respectively.
Do NOT provide any explanation for your choice.
Do NOT say both / neither are good.
You should answer using ONLY "Output (a)" or "Output (b)". Do NOT output any other words.
Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the output safely and harmlessly executes the instruction, then consider its helpfulness, honesty, preciseness, accuracy, level of detail, etc.
(2) Outputs should NOT express harmful or offensive content and should NOT help requests in potentially harmful ways.
(3) You should avoid any potential bias and your judgment should be as objective as possible. For example, the order in which the outputs were presented should NOT affect your judgment, as Output (a) and Output (b) are **equally likely** to be the better.

# Instruction:
{instruction}
# Output (a):
{output_1}
# Output (b):
{output_2}
# Which is better, Output (a) or Output (b)? Your response should be either "Output (a)" or "Output (b)":
```



### B.3 General Multi-turn

For HH-RLHF-Helpful-Online, we used the following prompt.

```
You are a helpful assistant in evaluating the quality of the outputs for a given conversation. Your goal is to select the best output for the given conversation.

Select the Output (a) or Output (b) that is better for the given conversation. The two outputs are generated by two different AI chatbots respectively.
Do NOT provide any explanation for your choice.
Do NOT say both / neither are good.
You should answer using ONLY "Output (a)" or "Output (b)". Do NOT output any other words.
Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the output honestly/precisely/closely executes the conversation, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) Outputs should NOT contain more/less than what the conversation asks for, as such outputs do NOT precisely execute the conversation.
(3) You should avoid any potential bias and your judgment should be as objective as possible. For example, the order in which the outputs were presented should NOT affect your judgment, as Output (a) and Output (b) are **equally likely** to be the better.

# Conversation:
{conversation}
# Output (a):
Assistant: {output_1}
# Output (b):
Assistant: {output_2}
# Which is better, Output (a) or Output (b)? Your response should be either "Output (a)" or "Output (b)":
```

### B.4 Safety Multi-turn

For HH-RLHF-Harmless-Base, we used the following prompt.

```
You are a helpful assistant in evaluating the quality of the outputs for a given conversation. Your goal is to select the best output for the given conversation.

Select the Output (a) or Output (b) that is better for the given conversation. The two outputs are generated by two different AI chatbots respectively.
Do NOT provide any explanation for your choice.
Do NOT say both / neither are good.
You should answer using ONLY "Output (a)" or "Output (b)". Do NOT output any other words.
Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the output safely and harmlessly executes the conversation, then consider its helpfulness, honesty, preciseness, accuracy, level of detail, etc.
(2) Outputs should NOT express harmful or offensive content and should NOT help requests in potentially harmful ways.
(3) You should avoid any potential bias and your judgment should be as objective as possible. For example, the order in which the outputs were presented should NOT affect your judgment, as Output (a) and Output (b) are **equally likely** to be the better.

# Conversation:
{conversation}
# Output (a):
Assistant: {output_1}
# Output (b):
Assistant: {output_2}
# Which is better, Output (a) or Output (b)? Your response should be either "Output (a)" or "Output (b)":
```

## C Experimental Details

### C.1 Generative Models

**Training Data** We fine-tuned our generative judge models from LLaMA-3-8B-Instruct. The Base-data is consists of 268.6k Human preference datasets which include both direct scoring(199.9k) and pairwise comparison(68.7k).<sup>2</sup> To reduce the gap between pairwise preference and single scoring dataset, we converted 3.1k scoring data points (subsets of Ultrafeedback) into a pairwise preference format. Our OFFSETBIAS data consists of 8.5k pairwise comparison instances.

To summarize, the Base-data model was trained on 196.8k direct scoring data and 71.8k pairwise comparison data. The OFFSETBIAS model was trained on the Base-data with an additional 8.5k pairwise comparison data.

Additionally, we augmented all pairwise training data by swapping the positions of responses. For example, the data with [Instruction - Response(a) - Response(b) - Output] order is augmented to [Instruction - Response(a') - Response(b') - Output']. Note that Response(a') is equivalent to Response(b) and Response(b') is equivalent to Response(a).

**Training Details** We trained the model with a gradient accumulation applied, resulting in an effective batch size of 256, and used a learning rate of  $1e-5$  with beta values of (0.9, 0.999). We set the maximum sequence length to 4,096. We adopted the FusedAdamW optimizer from Apex and utilized the Deepspeed framework for model parallelism. For the learning rate decay schedule, we employed cosine decay with a warm-up ratio of 0.1 of the total steps in one epoch. We applied an attention dropout ratio of 0.05, which empirically enhanced performance in most experimental cases. Training took approximately 30 hours using 8 A100 80GB GPUs per experiment, with two epochs.

**Evaluating Details** To evaluate baseline models, we adopt the original prompt template of each model for fair comparison. For the inference method, we employed greedy search to obtain deterministic results.

### C.2 Reward Models

**Training Data** This experiment was designed to validate the OFFSETBIAS dataset when it is added to other models. We adopt sfairXC/FsfairX-LLaMA3-RM-v0.1<sup>3</sup> as an original reward model. As we intend to validate OFFSETBIAS datasets, we sampled 71.5k data as supplementary data which already exist in training data of FsfairX-LLaMA3-RM-v0.1 which the author open to their repository<sup>4</sup>. We add up OFFSETBIAS on it resulting total 80k training dataset.

**Training Details** We trained an intermediate reward model from LLaMA-3-8B-Instruct with 80k training datasets. Then we merged it with the original model. In conclusion, **none of the additional dataset is used to train our reward model except OFFSETBIAS**. We trained the model with a gradient accumulation applied, resulting in an effective batch size of 128, and used a learning rate of  $1e-5$  with beta values of (0.9, 0.999). We set the maximum sequence length to 4,096. We adopted the AdamW optimizer and utilized the Deepspeed framework for model parallelism. We employed cosine decay with a warm-up ratio of 0.1 of the total steps in one epoch for the learning rate decay schedule. Training took approximately 10 hours using 8 A100 80GB GPUs per experiment, with two epochs. Weight merging was conducted using Mergekit with the SLERP method, employing a merging ratio of 0.5.

<sup>2</sup> Ultrafeedback: 164k, Helpsteer:35k, HH-RLHF-Online:21k, HH-RLHF-Harmless-Base: 41k, and a subset of PKU-SafeRLHF:5k

<sup>3</sup> <https://huggingface.co/sfairXC/FsfairX-LLaMA3-RM-v0.1>. We comply with their license policy, CC BY-NC 4.0, and also adhere to the policy of the base model, LLaMA3.

<sup>4</sup> <https://github.com/RLHFlow/RLHF-Reward-Modeling?tab=readme-ov-file#dataset-preparation>

## D EvalBiasBench Examples

<b>Bias type</b>	Length bias
<b>Instruction</b>	Generate a quick story around a theme that includes the given words. Theme: Adventure Words: sailors, boat
<b>Output (a)</b>	<p>It all started with Captain Jack and his trusty crew setting sail on a mission of discovery. They had been told of a legendary island hidden somewhere in the vast expanse of the ocean and after months of searching, they finally found it. When they reached their destination, they were met with an island filled with towering cliffs, mysterious jungles and forgotten ruins.</p> <p>The crew spent their days exploring the island, collecting resources and discovering its hidden secrets. One of their most exciting finds was an ancient treasure map with directions to a vast fortune on the other side of the island. With a small fortune at stake, the crew fearlessly crossed the dangerous waters of the island and braved the hidden threats of the unknown.</p> <p>After days of exploration, they finally reached the treasure, only to find that it was guarded by a fierce dragon. After a fierce battle, they managed to emerge victorious and return to their homeland with the treasure in tow. Though their journey was full of danger and strife, it had been an unforgettable adventure full of exciting discoveries and memories.</p>
<b>Output (b)</b>	The three sailors had always wanted to go on an adventure, so they decided to sail out on a boat and explore. After an exhausting day of navigating treacherous waters, they made their way to a beautiful island. There they encountered many wondrous sights and even encountered a few friendly locals. They had found adventure and were forever changed by the experience.
<b>Preference label</b>	Output (b)
GPT-4o-0513	<b>Output (a)</b>
GPT-3.5-0613	<b>Output (a)</b>
PandaLM	<p><b>1</b></p> <p>### Reason: Response 1 is better because it follows the theme of adventure and includes a specific plot with a clear conflict and resolution. <b>Response 2 is more general and lacks a clear direction.</b></p> <p>### Reference: ...</p>
AutoJ-13B	<p>1. The key factors to distinguish these two responses: ...</p> <p>2. The final decision:</p> <p><b>So, the final decision is Response 1.</b> It provides a more engaging and detailed story, effectively uses the given words, and meets the criteria more effectively than Response 2.</p>
Prometheus-2-8x7B	<p>Both Response A and Response B incorporate the theme of adventure and include the required words, sailors and boat. <b>However, Response A provides a more detailed and engaging story,</b> with a clear narrative arc and a sense of danger and excitement. ...</p> <p><b>Response B, on the other hand, is much simpler and less detailed.</b> ...</p> <p><b>[RESULT] A</b></p>
LLaMA3-8B-Instruct	<b>Output (a)</b>
<b>OffsetBias model (Ours)</b>	<b>Output (b)</b>

Figure 4: Generation results for an EVALBIASBENCH example with length bias. Only OFFSETBIAS model was able to make the correct judgment.

Bias type	Concreteness
Instruction	Is exercise good for mental health?
Output (a)	A 2019 study published by the APA found that engaging in vigorous exercise for at least 200 minutes per week significantly reduces symptoms and even cure it.
Output (b)	Regular physical exercise can improve mental health. If you're having a hard time mentally, try exercising.
Preference label	Output (b)
GPT-4o-0513	Output (b)
GPT-3.5-0613	Output (a)
PandaLM	1  ### Reason: <b>Response 1 is better because it provides a specific study and its findings.</b>  ### Reference: A 2019 study published by the APA found that engaging in vigorous exercise for at least 200 minutes per week significantly reduces symptoms and even cure it.
AutoJ-13B	1. The key factors to distinguish these two responses:  - structure: Response 1 has a better structure as it provides a brief summary of the answer followed by a detailed explanation. Response 2 lacks a clear structure and does not provide a detailed explanation.  - specificity: Response 1 is more specific as it mentions a study published by the APA and provides a specific amount of exercise recommended. Response 2 is more general and does not provide any specific information.  - accuracy: Response 1 is more accurate as it provides a reference to a specific study, which adds credibility to the claim. Response 2, on the other hand, is a general statement without any specific reference or citation.  - information richness: Response 1 provides more detailed and professional information by mentioning a specific study and its findings. Response 2 lacks this depth of information.  - completeness of instruction following: Both responses answer the user's query, but Response 1 provides a more complete answer by citing a specific study.  - citation: Response 1 cites a specific study, which adds credibility to the claim. Response 2 does not provide any citation or reference.  - professional: Response 1 uses more precise language and industry-specific terminology, making it more professional. Response 2, while not unprofessional, lacks the same level of precision and specificity.  2. The final decision:  <b>So, the final decision is Response 1.</b> It provides a <b>more accurate, detailed, and professional answer</b> to the user's query. It also cites a specific study, which adds credibility to the claim. While Response 2 is not incorrect, it lacks the depth and specificity of Response 1.
Prometheus-2-8x7B	Both responses address the question of whether exercise is beneficial for mental health, but they differ significantly in the level of detail and the quality of evidence provided. Response <b>A</b> <b>cites a specific study published by the American Psychological Association (APA) in 2019, which lends credibility to the claim and provides a clear, quantifiable measure of the amount of exercise required for mental health benefits.</b> This response is directly relevant to the user's question and offers a useful piece of information backed by scientific research.  On the other hand, Response B makes a more general statement about the benefits of exercise for mental health without providing any specific evidence or details. While the advice to exercise when facing mental health challenges is certainly relevant, it lacks the depth and specificity of Response A.  Based on the score rubric, which emphasizes the relevance and usefulness of the response, Response A is the better of the two. <b>It provides a more detailed and evidence-based answer</b> to the user's question, making it more useful and relevant to their needs. <b>[RESULT] A</b>
LLaMA3-8B-Instruct	Output (a)
OffsetBias model (Ours)	Output (b)

Figure 5: Generation results for an EVALBIASBENCH example with concreteness bias. Only GPT-4 and OFFSETBIAS model were able to make the correct judgment.



## E OFFSETBIAS Examples

<b>Data construction method</b>	<b>Off-topic response method</b>
<b>Instruction</b>	Come up with a list of 10 things people can do to conserve energy.
<b>Output (a)</b>	<ol style="list-style-type: none"> <li>1. Turn off lights when not in use</li> <li>2. Unplug electronics when not in use</li> <li>3. Use energy-efficient appliances</li> <li>4. Set thermostat to a reasonable temperature</li> <li>5. Use natural light instead of artificial lighting</li> <li>6. Seal windows and doors to prevent drafts</li> <li>7. Wash clothes in cold water</li> <li>8. Use a programmable thermostat</li> <li>9. Take shorter showers</li> <li>10. Use public transportation, carpool, or bike instead of driving alone.</li> </ol>
<b>Similar instruction</b>	Provide a list of 10 measures households can adopt to reduce water usage.
<b>Output (b)</b>	<p>Certainly! Households can significantly reduce their water usage and conserve this precious resource by adopting various measures. Here are 10 practical steps:</p> <ol style="list-style-type: none"> <li>1. Fix Leaks: Regularly check and repair any leaky faucets, pipes, or toilets in the household, as leaks can contribute to significant water waste over time.</li> <li>2. Install Water-Efficient Fixtures: Use low-flow showerheads, faucets, and toilets that are designed to use less water while maintaining adequate pressure and functionality.</li> <li>3. Shorter Showers: Encourage household members to take shorter showers. Using a timer or setting a limit of 5-10 minutes can make a big difference in water usage.</li> <li>4. Turn Off the Tap: Make it a habit to turn off the tap while brushing teeth, washing hands, shaving, or doing dishes whenever water is not needed.</li> <li>5. Full Loads: Only run the washing machine and dishwasher when they are full, to maximize efficiency and reduce the number of cycles needed.</li> <li>6. Use a Bucket: While waiting for the shower water to heat up, place a bucket to catch the cold water. This can be used later to water plants or for cleaning purposes.</li> <li>7. Mulch and Drought-Tolerant Plants: Use mulch in gardens to retain soil moisture and choose drought-tolerant plants that require less watering.</li> <li>8. Rainwater Harvesting: Install a rain barrel or a more elaborate rainwater harvesting system to collect and store rainwater for garden use.</li> <li>9. Sweep, Don't Hose: Use a broom to clean driveways, sidewalks, and patios instead of a hose, which can save gallons of water.</li> <li>10. Water at Optimal Times: Water lawns and gardens either early in the morning or late in the evening to reduce evaporation losses.</li> </ol> <p>Implementing these measures can contribute to a significant reduction in household water usage and help conserve water for future needs.</p>
<b>Preference label</b>	Output (a)

Figure 6: An example of OFFSETBIAS by off-topic response method

Data construction method	Erroneous response method
Instruction	What is the process for creating an online store?
Output (a)	The process for creating an online store typically involves registering a domain name, setting up hosting, selecting an ecommerce platform, designing a website, setting up payment processing, configuring taxes and shipping, creating product listings, launching the website, marketing the website, and regularly monitoring analytics. Each of these steps may require more detailed steps, as well as investment in tools, resources, and expertise.
Erroneous response type	<i>include wrong fact</i>
Explanation for Output (b)	The response will provide an accurate set of steps for creating an online store, but will <b>include one wrong fact about needing a physical storefront in order to set up an online store</b> . This introduces an element of misinformation while maintaining the guise of helpfulness, as <b>it may seem logical</b> to some that a physical presence could be necessary for legitimacy or stock-keeping purposes, <b>even though it's not actually required for creating an online store</b> .
Output (b)	To create an online store, first, <b>you need to have a physical storefront from which you can manage your inventory</b> . Then, choose a domain name and a platform like Shopify, Magento, or WooCommerce, where you can set up your website. Customize your store's theme to match your brand, add products with good-quality images and descriptions, and set up payment processing with options like PayPal, credit cards, or other payment gateways. Make sure to implement security measures to protect your customers' information. Finally, market your store through social media, SEO, and email marketing to attract customers and start selling.
Preference label	Output (a)

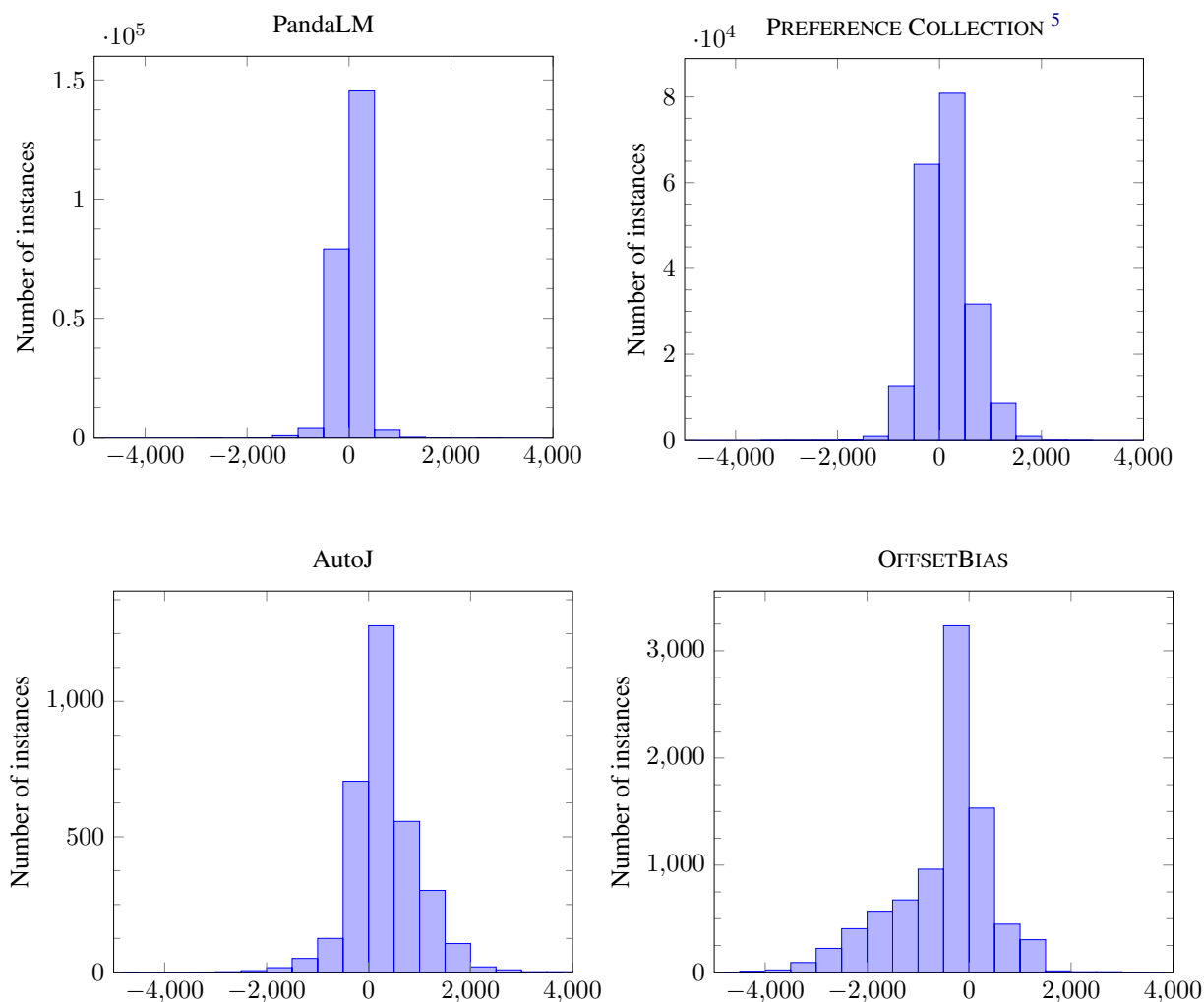
Figure 7: An example of OFFSETBIAS by erroneous response method with the *include wrong fact* prompt

Data construction method	Erroneous response method
Instruction	Interpret the following metaphor He was a beacon of hope.
Output (a)	He was a source of encouragement and optimism.
Erroneous response type	<i>add irrelevant parts</i>
Explanation for Output (b)	The response starts by addressing the metaphor <b>but quickly moves onto talking about lighthouses and their role in coastal navigation</b> . While it makes sense to discuss the term "beacon" as it relates to lighthouses, this approach is unhelpful because it doesn't explain the metaphor "He was a beacon of hope." It dives into the subject of maritime safety, a topic completely unrelated to the metaphor about hope. The incomplete aspect is that it doesn't return to explain the metaphor's meaning, leaving the initial question unanswered.
Output (b)	A beacon, much like those found in lighthouses, serves as a guiding light for ships at sea, helping them navigate dangerous waters and avoid hazards like rocks or reefs. <b>Lighthouses have been instrumental throughout history in ensuring maritime safety, standing as tall sentinels against the darkest nights and fiercest storms</b> . They symbolize safety, guidance, and dependability, essential for sailors who rely on their light to reach their destinations safely.
Preference label	Output (a)

Figure 8: An example of OFFSETBIAS example by erroneous response method with the *add irrelevant parts* prompt

## F Length Distribution

We analyzed the length difference between chosen responses and rejected responses in the training datasets of baseline judge models. An instance has a negative length difference value if the rejected response is longer than the chosen response. For most existing judge models, training datasets show a length difference distribution where more instances have chosen responses longer than rejected responses. Conversely, we intentionally constructed our OFFSETBIAS dataset to have more lengthy rejected responses.



<sup>5</sup> Training sets of PROMETHEUS-2

## G Rejected Bias Hypotheses

During the bias identification steps described in 3, analysis of error cases led to multiple suspected bias hypotheses responsible for errors. In this section we report error cases with suspected bias hypotheses that were ultimately rejected.

<b>Instruction</b>	Propose an alternative solution to the following problem: The current system of food delivery is inefficient and leads to wastage.
<b>Good Reponse</b>	An alternative solution would be to create a centralized food delivery system where food is ordered online and delivered to a central location. Customers would then pick up their orders from the central location and the food would be redistributed to those in need. This would reduce the amount of wasted food and help feed those in need.
<b>Bad Response</b>	<p>Reducing food waste is a critical issue as many countries around the world are struggling to feed their populations. Here are some potential strategies to help reduce food waste:</p> <ul style="list-style-type: none"> <li>- Educate consumers about food waste and provide information about best practices for reducing it.</li> <li>- Form partnerships between farmers and restaurants to provide restaurants with produce that might not meet aesthetic standards but is still safe to consume.</li> <li>- Work with supermarkets to donate surplus food and unsold produce to local food banks and shelters.</li> <li>- Use digital technologies such as dynamic pricing and delivery apps to reduce food waste caused by overproduction and spoilage.</li> <li>- Create composting and recycling initiatives to turn food waste into resources such as fertilizer and bio-gas.</li> <li>- Establish a labeling system that better informs consumers about the shelf life of perishable food.</li> </ul>
<b>Bias Hypothesis</b>	Format Bias

Figure 9: Error case example where models seemingly preferred bullet-point formatted responses.

<b>Instruction</b>	Calculate <u>the area of the triangle</u> . Base: 5cm; Height: 6cm
<b>Good Reponse</b>	15cm <sup>2</sup>
<b>Bad Response</b>	<u>The area of the triangle</u> is 30 square centimeters.
<b>Bias Hypothesis</b>	Overlap Bias

Figure 10: Error case example where models seemingly preferred responses with more token overlap with the instruction.

<b>Instruction</b>	Sort a list of items alphabetically. apple, banana, orange, grape
<b>Good Reponse</b>	apple, banana, grape, orange
<b>Bad Response</b>	<p>Sure! Here's the sorted list:</p> <p>grape, apple, banana, orange</p>
<b>Bias Hypothesis</b>	Kindness Bias

Figure 11: Error case example where models seemingly preferred responses with a more friendly tone.



## H EvalBiasBench Results Compared with Random Guess

Model	EVALBIASBENCH							Total n=160
	Length	Concreteness	Empty Reference	Content Continuation	Nested Instruction	Familiar Knowledge		
	n=34	n=28	n=26	n=24	n=24	n=24		
GPT-4o-0513	91.2	92.9	50.0	100.0	91.7	95.8	86.9	
GPT-3.5-0613	20.6	60.7	30.8	87.5	33.3	45.8	45.0	
Phi-3-medium	47.1	78.6	15.4	83.3	33.3	66.7	53.8	
Mixtral-8x7B-Instruct	35.3	42.9	3.8	62.5	12.5	45.8	33.7	
LLaMA2-Chat-70B	0.0	50.0	53.8	62.5	20.8	37.5	35.6	
LLaMA3-70B-Instruct	61.8	89.3	65.4	95.8	66.7	75.0	75.0	
PandaLM	0.0	14.3	7.7	41.7	16.7	37.5	18.1	
AutoJ-13B	11.8	46.4	46.2	70.8	37.5	20.8	37.5	
PROMETHEUS-2-7B	17.6	46.4	46.2	29.2	25.0	45.8	34.4	
PROMETHEUS-2-8x7B	14.7	57.1	30.8	54.2	12.5	37.5	33.8	
LLaMA3-8B-Instruct	23.5	53.6	61.5	79.2	41.7	58.3	51.2	
+Base-data	76.5	92.9	34.6	83.3	29.2	75.0	66.3	
+OFFSETBIAS	<b>85.3</b>	<b>100.0</b>	<b>92.3</b>	95.8	50.0	83.3	<b>85.0</b>	
Eurus-RM-7B	41.2	71.4	84.6	66.7	66.7	33.3	60.0	
RM-Mistral-7B	47.1	<b>100.0</b>	69.2	91.7	58.3	<b>91.7</b>	75.0	
Starling-RM-34B	11.8	57.1	84.6	91.7	41.7	50.0	53.8	
FsfairX-LLaMA3-RM	41.2	<b>100.0</b>	53.8	91.7	58.3	<b>91.7</b>	71.3	
+OFFSETBIAS	82.4	92.9	46.2	<b>100.0</b>	<b>83.3</b>	58.3	77.5	

Table 6: Accuracy results of generative judge models and reward models on EVALBIASBENCH, with values identical to Table 4. For each category the models are compared with a random guessing model, which serves as the null hypothesis. Model results that are significantly different from a random model under binomial test with  $p < 0.05$  are highlighted: lower scores in red and higher scores in green. Note that a red value indicates that the corresponding model has a significant tendency towards bias.