Benchmarking the Robustness of LLMs under Multi-Bias Scenarios

Anonymous ACL submission

Abstract

Despite significant progress, recent studies have indicated that current large language models (LLMs) may still utilize bias during inference, leading to the poor generalizability of LLMs. Some benchmarks are proposed to investigate the generalizability of LLMs, with each piece of data typically containing one type of controlled bias. However, a single piece of data may contain multiple types of biases in practical applications. To bridge this gap, we propose a multi-bias benchmark where each piece of data contains five types of biases. The evaluations conducted on this benchmark reveal that the performance of existing LLMs and debiasing methods is unsatisfying. Additionally, our analysis shows that relying on scaling up parameters and using slow thinking to improve the generalizability on multi-bias scenarios is not advisable considering the consumption of massive resources. These results highlight the challenge of eliminating multiple types of biases simultaneously.

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across diverse domains (Achiam et al., 2023). However, previous works have shown that LLMs would also learn **bias** during the training process (Schick et al., 2021; Navigli et al., 2023; Klimashevskaia et al., 2024), leading to *poor generalizability* of LLMs (Du et al., 2023; Cheng and Amiri, 2024; Yang et al., 2025). Thus, it is crucial to assess LLMs' generalizability with respect to biases and explore effective solutions for enhancement.

Recently, a series of benchmarks (Manerba et al., 2024; Bang et al., 2024) have been proposed to investigate the generalizability of LLMs with respect to biases. In most of these benchmarks, each piece of data typically contains only one type of controlled bias, with a small number of them containing two types. However, a single piece of data may simultaneously contain multiple types of biases in practical applications. As shown in Figure 1, this example contains five different types of biases that have been identified on the natural language inference (NLI) task (Gururangan et al., 2018; Dev et al., 2020; Sun et al., 2024b), and these types of biases have the same polarity 'entailment' (i.e., these types of biases will all cause an LLM to tend to predict the same answer 'entailment'). However, the gender of Oliver is not mentioned in the premise. Therefore, the relationship between the premise and the hypothesis is neutral. This type of data poses a greater challenge to LLMs because each type of bias can cause LLMs to tend to make certain predictions. If a single piece of data contains multiple types of biases, and these biases have the same polarity, it becomes extremely difficult for LLMs to utilize useful semantic information during inference. It also poses new challenges to the existing debiasing methods. This is because even if there is just one type of bias that has not been eliminated, it is likely to significantly affect the generalizability of LLMs. As a result, it is crucial to assess LLMs' generalizability by utilizing data that contains multiple types of biases.

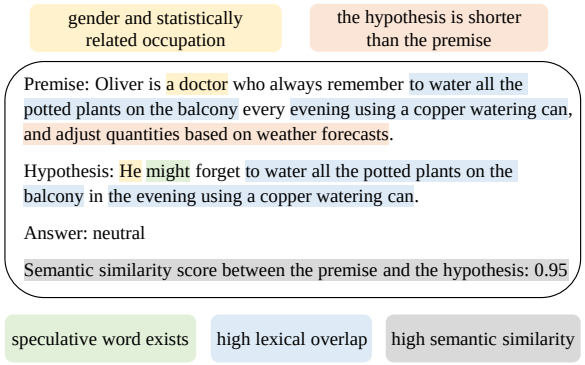

Figure 1: This figure presents an example that contains five different types of bias.

To address the limitations of existing evaluation,

we propose a multi-bias benchmark, in which each piece of data contains five types of biases, and these biases have the same polarity. To construct such a benchmark, it is necessary to take into account multiple types of biases simultaneously and ensure that their polarities are the same. Therefore, we choose the NLI task, on which previous researchers have conducted extensive studies regarding bias (Rajaee et al., 2022; Mckenna et al., 2023; Sun et al., 2024b), as the task format of this multi-bias benchmark. Subsequently, we explore and verify the polarity of the biases that have been discovered on the widely used LLM GPT-4o-2024-11-20 (Achiam et al., 2023) and select five biases with the same polarity. By manually controlling these biases within each piece of data, we are able to construct the multi-bias benchmark where each piece of data contains five types of biases. Then, to quantify the generalizability of LLMs, we thoroughly evaluate the mainstream LLMs and debiasing methods. Evaluation results demonstrate that even the most powerful LLMs and debiasing techniques are insufficient in handling multi-bias scenarios. Furthermore, our analysis indicates that relying on scaling up parameters and using slow thinking to improve the generalizability on multi-bias scenarios is not advisable considering the consumption of massive resources.

Additionally, to investigate scenarios where different biases do not share the same polarity, we also construct a dataset in which each data point exhibits biases with two distinct polarities. Experimental results demonstrate that these two biases offset each other to a certain extent. We will publicly release these data after the paper is accepted.

2 Bias Polarity Study

Formally, each type of bias contains at least one kind of bias feature. For example, the bias type 'lexical overlap' have two bias features 'high lexical overlap' and 'low lexical overlap'. To construct the multi-bias benchmark, we need to ensure that the polarities of these bias features are consistent. However, most of the previous researches (McCoy et al., 2019; Rajaee et al., 2022; Anantaprayoon et al., 2024) on bias polarity are based on small discriminative models. The conclusions drawn from such scenarios may not necessarily be applicable to generative large language models. Therefore, we first design experiments for the biases that have already been discovered by predecessors to explore and verify their polarities. Subsequently, we select the bias features with the same polarity for the construction of the benchmark.

GPT-4o	entailment	neutral	contradiction
hypothesis shorter	**43.7**	21.7	34.4
hypothesis longer	31.4	**40.1**	28.5
label distribution	33.3	33.3	33.3

Table 1: The distribution of predicted labels by GPT-4o and in the datasets in which each piece of data contains bias feature 'hypothesis longer' or 'hypothesis shorter'.

Take the bias type of sentence length for an example, Gururangan et al. (2018) discovered that the label neutral was associated with longer hypotheses, while the label entailment was associated with shorter hypotheses. Inspired by this finding, we take the relative length difference between the premise and hypothesis exceeding five words as the criterion for the sentence length bias and conduct explorations through experiments. For simplicity, we refer to the bias feature where the hypothesis is longer than the premise by more than five words as 'hypothesis longer', and the bias feature where the hypothesis is shorter than the premise by more than five words as 'hypothesis shorter'. Specifically, for the bias features 'hypothesis longer' and 'hypothesis shorter', we randomly select 3,000 samples with balanced labels from existing datasets such as MNLI (Williams et al., 2018) for each feature respectively. Then, we statistically analyze the label distribution predicted by GPT-4o-2024-11-20 and compare it with the label distribution of the dataset. Experimental results are shown in Table 1.

From the table, we can find that GPT-4o tends to predict entailment and neutral for the bias features 'hypothesis longer' and 'hypothesis shorter', respectively. So the polarity of the bias feature 'hypothesis longer' is neutral (and entailment for the bias feature 'hypothesis shorter').

For other types of biases such as lexical overlap, we also conducted similar experiments to investigate their polarities (More details about the experiments for other biases can be seen in Appendix C). Eventually, the types of biases that still exist in GPT-4o, as well as their polarities, are shown in Table 2. To ensure the consistency of polarities for selected bias features, we choose the bias features 'hypothesis shorter', 'high lexical overlap', 'high semantic similarity', 'speculative word exists', and 'male with male-biased occupations' to construct our multi-bias benchmark. Moreover, to investigate

Bias Type	Bias Feature	Bias Polarity
sentence length	hypothesis shorter: the hypothesis is shorter than the premise by more than five words	entailment
	hypothesis longer: the hypothesis is longer than the premise by more than five words	neutral
lexical overlap	high: the lexical overlap rate between the premise and hypothesis is higher than 0.8	entailment
	low: the lexical overlap rate between the premise and hypothesis is lower than 0.2	neutral
semantic similarity	high: the Bertscore between the premise and hypothesis is higher than 0.88	entailment
	low: the Bertscore between the premise and hypothesis is lower than 0.83	neutral
speculative word	speculative word exists: speculative word exists in the premise or the hypothesis	entailment
gender-occupation	male with male-biased occupations	entailment
	male with female-biased occupations	contradiction

Table 2: This table presents five types of biases with nine different bias features and their polarities. The bias of 'male with male-biased occupations' means that occupations statistically related to male appears in the premise, and male appears in the hypothesis (similar for male with female-biased occupations).

scenarios where different biases do not share the same polarity, we choose the bias features 'high semantic similarity' and 'low lexical overlap' to construct datasets for exploration.

3 Benchmark Construction

The challenge of constructing a multi-bias benchmark lies in how to control multiple types of biases simultaneously. To overcome this challenge, we design a data generation process that consists of three main steps, where each step introduces distinct biases to the data: (i) Template construction. (ii) vocabulary construction and template completion using the constructed vocabulary. (iii) verification of bias features on each generated data.

3.1 Template Construction

We construct four pairs of templates for the label neutral and the other two labels, respectively. Each template is explicitly incorporated with the bias feature of 'male with male-biased occupations' and 'speculative word exists'. Below is one pair of template, the first template is for the label neutral and the second is for the other two labels:

(1) Premise: N_1 is a P_1, V_1. Hypothesis: He S_1 V_2.

(2) Premise: N_1 is a P_1. He V_1. Hypothesis: He S_1 V_2.

where N_1 is a unisex name, P_1 is a male-biased occupation (occupations statistically related to male) used to control the bias feature of 'male with male-biased occupations', S_1 is a speculative word such as 'might', V_1 and V_2 are two verb phrases used to control the bias types of lexical overlap and sentence length. Note that we also control that the gender information is not introduced in V_1, so that the label of the samples constructed using the first template will always be neutral. Full list of the templates can be seen in Appendix A.

3.2 Vocabulary Construction and Template Completion

We drew our occupation vocabulary that includes 87 male-biased occupations from Anantaprayoon et al. (2024), and name vocabulary that contains 30 unisex names from the website[1]. For speculative vocabulary, we manually collect 6 speculative words (shown in Appendix B).

To control the answer of the generated data, we use GPT-4o-2024-11-20 to construct the vocabulary of verb phrase pairs (instead of verb phrases) and verify their validity. For each pair of them, the two verb phrases are used to fill in the premise and the hypothesis, respectively. To control the bias features of 'high lexical overlap' and 'hypothesis shorter', we utilize an automatic program to ensure that each verb phrase pair has a remarkably high degree of lexical overlap, and the verb phrase corresponding to the premise is more than three words longer than that corresponding to the hypothesis (since the other part of the premise is at least two words longer than that of the hypothesis in each template). Meanwhile, we manually verify whether each verb phrase pair always results in the premise entailing (or contradicting) the hypothesis after inserted into the templates. Three of us were involved in this verification process, and it was ensured that all three of us deemed that each verb phrase pair satisfied the requirements. Finally, we got a vocabulary containing 200 pairs of verb phrases (100

[1] https://www.behindthename.com/names/gender/unisex

Premise	Hypothesis	label
Noah is a plumber. He builds furniture to decorate the home and save costs, using recycled wood.	He might build furniture to decorate the home and save costs.	Entailment
Emma is a officer. He plans an extensive and detailed project outline for the community event.	He could plan an extensive project outline for the community event.	Entailment
Emerson, a judge by trade, considers the various intricate details of the new art exhibit in the gallery.	He could consider the intricate details of the art exhibit.	Neutral
Morgan is a scholar, visiting the park regularly to enjoy nature and unwind, usually along a shaded path.	He might visit the park regularly to enjoy nature and unwind.	Neutral
Liam works as a trader. He declared that delivering 30 food boxes by December 10th would organize holiday meals, however, the warehouse received only 11 boxes by the deadline.	He would organize holiday meals after delivering 30 food boxes by December 10th.	Contradict
Liam is a technician. He confirm that hosting 15 free coding workshops before December 15th would qualify for tech grants. however, there are only 6 workshops held by the deadline.	He may qualify for tech grants after hosting 15 free coding workshops before December 15th.	Contradict

Table 3: This table presents six examples in our constructed multi-bias benchmark.

pairs corresponds to the label 'entailment' and 100 pairs corresponds to the label 'contradiction').

After collecting the vocabulary, we can complete the template by randomly extracting data from the vocabulary. Note that when constructing data with the 'neutral' label, we also use verb phrase pairs corresponding to the label 'entailment'. Since the 'neutral' label is determined by templates, this does not affect the correctness of the data.

3.3 Verification of Bias Features

To ensure that the constructed data meets the requirements of the five bias features. Therefore, we employ an automatic program to select the data that satisfies the requirements of these bias features. Take the bias feature of 'high semantic similarity' as an example, we calculate the Bertscore between the premise and hypothesis and only retain data where the Bertscore exceeds 0.88 during this verification process. By doing these, we construct a multi-bias benchmark that contains 12000 samples with balanced labels. Some Dataset samples are provided in Table 3.

4 Experiments

In this section, we explore three core research questions with the proposed multi-bias benchmark. **RQ1**: Will the LLMs be more susceptible to the multi-bias scenarios? **RQ2**: Are existing state-of-the-art (SOTA) debiasing methods still effective in multi-bias scenarios? **RQ3**: Can the challenges of multi-bias scenarios be solved by scaling the model parameters? **RQ4**: Does slow-thinking enhance the robustness of LLMs in multi-bias scenarios?

4.1 Evaluation Setup

Models In this work, we utilize GPT-4o-2024-11-20 (Achiam et al., 2023). Llama3.1-70B-chat (Dubey et al., 2024), Gemma3-27B-it (Team et al., 2025), and Qwen3-8B (Yang et al., 2024) for the main experiments. We evaluate the LLMs in two ways: zero-shot, few-shots prompting (without slow thinking). For few-shot experiments, we randomly select three examples with balanced labels from the MNLI train set. All experiments are done for three times and we report the average of the accuracy.

Datasets In addition to the constructed benchmark (called 5-bias below) that contains five types of biases within each data, we also select the crowd-sourced dataset MNLI (Williams et al., 2018), and the meticulously curated HANS (McCoy et al., 2019) dataset in which each instance contains one type of controlled bias (lexical overlap, subsequence, or constituent) for experimental analysis. Since the gold labels for the MNLI test set are not publicly available, we follow previous work and

	GPT-4o			Llama3.1-70B-chat			Gemma3-27B-it			Qwen3-8B		
Zero-shot	MNLI	HANS	5-bias	MNLI	HANS	5-bias	MNLI	HANS	5-bias	MNLI	HANS	5-bias
Vanilla	84.5	83.8	51.7	80.5	75.4	40.4	82.6	75.4	43.8	77.9	80.8	33.5
DC	84.6	**84.1**	52.4	80.6	77.5	43.5	82.4	77.8	46.2	78.0	81.5	34.6
BC	**84.7**	84.0	**53.7**	80.8	**78.6**	44.6	82.6	**78.6**	47.9	**78.2**	**81.6**	35.2
Unibias	—	—	—	80.5	77.3	42.3	82.5	77.7	45.6	**78.2**	81.2	34.1
CAL	—	—	—	**81.2**	76.6	**46.7**	**82.9**	78.4	**50.1**	77.4	80.4	**35.6**
Few-shot	MNLI	HANS	5-bias	MNLI	HANS	5-bias	MNLI	HANS	5-bias	MNLI	HANS	5-bias
Vanilla	85.4	84.8	47.9	83.9	89.0	54.0	84.4	80.5	41.5	**80.6**	76.8	40.5
DC	**85.6**	85.0	50.2	83.8	89.5	54.3	84.3	80.9	44.7	80.4	77.6	41.5
BC	85.5	**85.2**	**52.8**	84.1	**89.7**	**54.6**	84.4	80.7	**46.4**	**80.6**	**77.8**	42.1
Unibias	—	—	—	**84.2**	89.4	54.1	**84.6**	80.6	42.7	80.5	77.1	41.2
CAL	—	—	—	83.6	89.6	54.5	84.2	**81.4**	45.8	80.2	77.5	**44.7**

Table 4: Evaluation results of different LLMs and different debiasing methods on MNLI, HANS, and 5-bias datasets. Best in bold, second with an underline. – denotes that this method is inapplicable to GPT-4o due to the requirement for internal representations of the model.

report the results on the development-matched.

Debiasing Methods In this work, we conduct a comprehensive analysis of the existing debiasing methods for LLMs, including DC (Fei et al., 2023), BC (Zhou et al., 2024a), Unibias (Zhou et al., 2024b), and CAL (Sun et al., 2024b) methods. Additionally, we also report the results of vanilla in-context learning.

More details about the specific prompts used for evaluation can be seen in Appendix D.

4.2 Main Results

Experimental results are shown in Table 4, from which we find that:

(1) Comparing the performance of four large language models (LLMs) on the HANS and 5-bias datasets in which each instance contains one and five types of controlled bias features, it is evident that the performance on 5-bias dataset is much lower (more than 20%) than that on HANS. This indicates that LLMs are more susceptible to multi-bias scenarios where each piece of data contains multiple types of biases.

(2) Compared to the vanilla zero-shot and few-shot prompting methods, current debiasing methods achieves better performance on HANS and 5-bias datasets in general (with BC and CAL methods performs the best). This demonstrates the effectiveness of current debiasing methods. However, the performance of these methods on 5-bias dataset is still unsatisfying, which indicates that current debiasing methods is not effective enough in multi-bias scenarios. This also highlights the importance of devising debiasing methods that can deal with multi-bias scenarios.

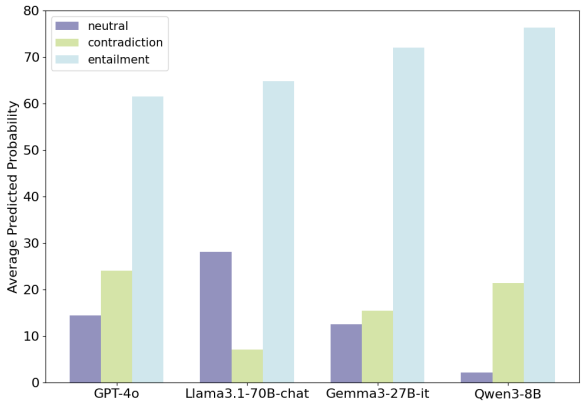

Figure 2: Average predicted probobilities of four LLMs for each label on the 5-bias dataset.

(3) When comparing vanilla zero-shot and few-shot methods, it can be find that few-shot method does not always achieve better performance than zero-shot method on the 5-bias dataset (especially for GPT-4o). This indicates that the random selection of few-shot samples may lead to the introduction of additional bias (which is consistent with the findings reported in Zhou et al. (2024a); Li et al. (2024)), thereby resulting in the unsatisfied performance. As a result, few-shot prompting can not improve the generalizability of LLMs on multi-bias scenarios easily.

(4) Compared to smaller LLMs, the performance of larger LLMs on the 5-bias dataset is higher in

	GPT-4o			Llama3.1-70B-chat			Gemma3-27B-it			Qwen3-8B		
Zero-shot	L-bias	S-bias	2-bias	L-bias	S-bias	2-bias	L-bias	S-bias	2-bias	L-bias	S-bias	2-bias
Vanilla	82.4	85.5	86.0	63.1	62.3	69.4	71.5	73.6	74.0	61.2	64.9	65.1
DC	82.3	85.3	85.9	67.5	67.2	72.4	71.7	73.8	75.7	62.3	66.2	66.4
BC	82.6	85.5	86.1	67.9	68.0	73.3	71.8	73.8	75.9	62.5	66.5	66.6
Unibias	—	—	—	65.9	65.4	70.7	71.7	73.4	75.3	61.9	65.5	65.6
CAL	—	—	—	65.5	68.6	71.8	71.6	74.1	74.7	66.4	73.7	70.5

Table 5: The table presents the accuracy of different LLMs and different debiasing methods on L-bias, S-bias, and 2-bias datasets. L-bias and S-bias represents datasets which only exhibit the bias of 'low lexical overlap' or 'high semantic similarity' while excluding those with the other four types of biases. − denotes that this method is inapplicable to GPT-4o due to the requirement for internal representations of the model.

	Qwen3-8B		Qwen3-14B		Qwen3-32B		GPT-5		Claude-3.5		Gemini-2.5	
Mode	2-bias	5-bias	2-bias	5-bias	2-bias	5-bias	2-bias	5-bias	2-bias	5-bias	2-bias	5-bias
No Thinking	65.1	33.5	**83.3**	**52.0**	83.3	45.6	**86.4**	**58.0**	**79.3**	47.0	**72.6**	**44.5**
Thinking	**65.6**	**40.8**	83.2	51.9	**83.7**	**46.5**	70.2	51.0	77.9	**50.8**	70.1	40.7

Table 6: Evaluation results of different thinking modes on 2-bias and 5-bias datasets for six different LLMs.

general, which indicates that larger LLMs is more robust in multi-bias scenarios. However, scaling up model size requires massive resources and current LLMs' performance is far from expectations even for GPT-4o, indicating that relying on scaling up parameters to improve the generalizability on multi-bias scenarios is not advisable.

4.3 Analysis of the Average Predicted Probability

For a more in-depth analysis, we statistically analyzed the average predicted probabilities of four LLMs for each label. Specifically, we conduct a statistical analysis on the zero-shot evaluation results for the 5-bias dataset. As shown in Figure 2, the average predicted probabilities of labels 'neutral' and 'contradiction' are significantly lower than those of the label 'entailment', which suggests that the bias features contained in these data causes LLMs to exhibit a tendency to predict 'entailment'. This aligns with the original intention of constructing the 5-bias dataset. As the five injected bias features all cause LLMs to tend to predict 'entailment', these LLMs will predict 'entailment' most of the time, thereby exhibiting lower average predicted probabilities on other label categories. Furthermore, comparing the average predicted probabilities of the label 'entailment' among four LLMs, it can be find that the probability is lower for larger LLMs, which also indicates that the increase in the number of model parameters exerts a certain effect on reducing the bias of LLMs. However, the decrease of the average predicted probability is relatively small. Considering the significant increase in required computational resources, it is not a good direction to improve the robustness of LLMs in multi-bias scenarios only by scaling up parameters.

4.4 Analysis of LLMs' Behavior with Different Polarity of Bias Features

To investigate scenarios where different bias features do not share the same polarity, we choose the bias features of 'high semantic similarity' and 'low lexical overlap', whose polarities are entailment and neutral respectively, to construct a 2-bias dataset for exploration. Specifically, we filtered out data samples from the SNLI dataset (Bowman et al., 2015) that exhibit these two bias features while excluding those with the other three types of biases. Ultimately, a label-balanced dataset with 12000 samples was constructed. We did not adopt this method to develop the multi-bias benchmark, primarily because existing datasets rarely contain samples that simultaneously exhibit more than two types of bias. Additionally, we also filtered out samples from the SNLI dataset which only exhibit the bias feature of 'high semantic similarity' or 'low lexical overlap' while excluding those with the other four types of biases (the two datasets are short for S-bias and L-bias respectively in Ta-

Figure 3: Error rates of six LLMs (no thinking) for each label on the 5-bias dataset.

Figure 4: Error rates of six LLMs (slow thinking) for each label on the 5-bias dataset.

ble 5). Finally, we get two label-balanced 1-bias datasets with 12000 samples. We conduct experiments on these two 1-bias datasets and the 2-bias dataset using GPT-4o-2024-11-20. Llama3.1-70B-chat, Gemma3-27B-it, and Qwen3-8B. The experimental results are shown in Table 5.

From the table, we can find that GPT-4o is relatively robust to the bias features of 'high semantic similarity' and 'low lexical overlap', while the other three LLMs are susceptible to these two bias features. Comparing the vanilla zero-shot method between two 1-bias datasets and the 2-bias dataset, it can be find that the performance on the 2-bias dataset is higher than that on two 1-bias datasets for four LLMs. This demonstrates that bias features with different polarities (entailment and neutral) will offset each other to a certain extent, so that the bias features' impact on LLMs is relatively lower for the 2-bias dataset. Additionally, we also observe that previous debiasing methods is also effective for data with different polarity of bias features. However, the performance improvement of these debiasing methods on 2-bias dataset is relatively low, which indicates that current debiasing methods is also not effective enough in scenarios with different polarity of bias features.

4.5 Comparison Among Different Thinking Modes

To investigate whether slow thinking enhances the robustness of LLMs against biases, we utilize GPT-5-2025-08-07 (OpenAI, 2025), Qwen3-32B, Qwen3-14B, Qwen3-8B (Yang et al., 2024), Claude-3-7-sonnet-20250219 (Anthropic, 2025), and Gemini-2.5-flash-preview-09-2025 (Comanici et al., 2025) for experiments. For these models, we conduct experiments on the 2-bias and 5-bias datasets with and without slow thinking. For GPT-5, we set the reasoning effort to high when conducting experiments of slow thinking. For experiments of slow-thinking, we set the max tokens to 8192. Experimental results are shown in Table 6. From the table, we can find that slow thinking cannot always improve the performance on multi-bias scenarios for the same LLM. Considering the significant increase in inference time, it is inappropriate to enhance the robustness of LLMs in multi-bias scenarios by increasing thinking time.

For a more in-depth analysis, we statistically analyzed the error rates of six LLMs with and without slow thinking for each label. Specifically, we conduct a statistical analysis on the evaluation results of the 5-bias dataset. As shown in Figure 3, the er-

ror rate of data whose gold answer is 'neutral' and 'contradiction' are significantly higher than those with the gold answer 'entailment', which suggests that the bias features contained in these data causes LLMs to exhibit a tendency to predict entailment. This aligns with the original intention of constructing the 5-bias dataset. However, this phenomenon is also observed in Figure 4, which indicates that using slow-thinking cannot mitigate multiple types of biases simultaneously. Additionally, we find that slow-thinking exerts a certain degree of debiasing effect on GPT-5. However, slow thinking causes GPT-5 to exhibit a stronger tendency toward predicting the neutral label, thereby increasing the error rate on data points labeled as 'entailment' or 'contradiction'.

5 Related Work

Bias Evaluation Benchmarks Some recent studies (Fei et al., 2023; Zhou et al., 2024c; Sun et al., 2024a; Zhou et al., 2024a) have revealed that LLMs still utilize biases during the inference stage, and found that this phenomenon may lead to poor performance when generalizing to out-of-distribution scenarios. Zhou et al. (2024d) found that LLMs utilizes lexical and style biases in sentiment analysis tasks and propose a benchmark for verification. Anantaprayoon et al. (2024) proposed the NLI-CoAL benchmark to examine the gender bias within LLMs. Their research indicates that LLMs exhibits gender bias even in the NLI task. Steen and Markert (2024) focused on the news summarization task and proposed a dataset to explore the biases exploited by LLMs. They found that LLMs exhibit entity hallucination bias on this dataset, thus decreasing the performance of LLMs. Lin et al. (2025) also showed that premise order bias exists in the mathematical reasoning task. Sun et al. (2024b) proposed a causal-guided active learning framework to automatically identify biases without the prior knowledge, and found that LLM may utilize the bias of semantic similarity and speculative word on the NLI task.

Though there are many benchmarks proposed to investigate the generalizability of LLMs, each piece of data contains only one type of controlled bias in most of these benchmarks (few of them contains two types). However, a single piece of data may simultaneously contain multiple types of biases in practical applications. To mitigate this gap, we propose a multi-bias benchmark in which each piece of data contains five types of biases. By ensuring that these biases share the same polarity, it can be very challenging even for the most powerful LLMs.

Debiasing Methods for LLMs There are also some works aiming at mitigating biases of LLMs. Domain-context calibration method (Fei et al., 2023) uses random words sampled from the unlabeled evaluation dataset as the content-free text, and then calibrates the outputs based on the content-free text to mitigate label bias. Batch calibration (Zhou et al., 2024a) method models the bias from the prompt context by marginalizing the LLM scores in the batched input. Lv et al. (2025) introduces a historical bias calibration strategy that leverages deviations in the model's past response logits to mitigate cognitive biases in its current knowledge boundary assessment, and this methods works well in retrieval-augmented generation scenarios. Causal-guided active learning method (Sun et al., 2024b) first automatically identifies biased instances and induces explainable biases. Subsequently, this method utilizes these biased instances and explainable biases to debias LLMs. UniBias (Zhou et al., 2024b) identifies and eliminates biased feed-forward network layer (FFN) vectors and attention heads within LLMs.

Though there are some methods proposed to debias LLMs, they cannot deal with multi-bias scenarios as shown in our experiments. This highlight the necessary of devising methods that can eliminate multiple types of biases simultaneously.

6 Conclusions

In this paper, we focus on examining the validity of LLMs and debiasing methods in multi-bias scenarios. Concretely, we propose 5-bias, the first multi-bias benchmark where each piece of data contains 5 types of biases, for probing into the performance boundary of LLMs and debiasing methods in multi-bias scenarios. Then, an evaluation of current LLMs and debiasing methods on 5-bias is conducted. Evaluation results show that current LLMs and debiasing methods exhibit unsatisfied on this benchmark. Additionally, our analysis shows that relying on scaling up parameters or using slow thinking to improve the generalizability on multi-bias scenarios is not applicable considering the consumption of massive resources. These results highlight the challenge of eliminating multiple types of biases simultaneously.

Limitations

Although benchmarking the robustness of LLMs under multi-Bias scenarios, further exploration is still needed to improve the robustness of LLMs in multi-bias scenarios.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. 2024. Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6395–6408.

Anthropic. 2025. The claude 3 model family: Opus, sonnet, haiku.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jiali Cheng and Hadi Amiri. 2024. Fairflow: Mitigating dataset biases through undecided learning for natural language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21960–21975.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7659–7666.

Li Du, Xiao Ding, Zhouhao Sun, Ting Liu, Bing Qin, and Jingshuo Liu. 2023. Towards stable natural language understanding via information entropy guided debiasing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2868–2882.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings Of The 61St Annual Meeting Of The Association For Computational Linguistics (Acl 2023): Long Papers, Vol 1*, pages 14014–14031. Assoc Computational Linguistics-Acl.

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

Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. 2024. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34(5):1777–1834.

Lvxue Li, Jiaqi Chen, Xinyu Lu, Yaojie Lu, Hongyu Lin, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Zhongyi Liu, Xianpei Han, and 1 others. 2024. Debiasing in-context learning by instructing llms how to follow demonstrations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7203–7215.

Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. 2025. Implicit reasoning in transformers is reasoning through shortcuts. *arXiv preprint arXiv:2503.07604*.

Bo Lv, Nayu Liu, Yang Shen, Xin Liu, Ping Luo, and Yue Yu. 2025. Whether llms know if they know: Identifying knowledge boundaries via debiased historical in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19516–19528.

Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social bias probing: Fairness benchmarking for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14653–14671.

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

Nick Mckenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association*

for Computational Linguistics: EMNLP 2023, pages 2758–2774.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*.

OpenAI. 2025. Gpt-5 system card.

Sara Rajaee, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. Looking at the overlooked: An analysis on the word-overlap bias in natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10605–10616.

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

Julius Steen and Katja Markert. 2024. Bias in news summarization: Measures, pitfalls and corpora. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5962–5983.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024a. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 6883–6893.

Zhouhao Sun, Li Du, Xiao Ding, Yixuan Ma, Yang Zhao, Kaitao Qiu, Ting Liu, and Bing Qin. 2024b. Causal-guided active learning for debiasing large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14455–14469.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. 2025. Razor: Sharpening knowledge by cutting bias with unsupervised text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2024a. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2024b. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. In *Advances in Neural Information Processing Systems*, volume 37, pages 102173–102196. Curran Associates, Inc.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024c. Explore spurious correlations at the concept level in language models for text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 478–492.

Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. 2024d. Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2586–2614.

Premise	**Hypothesis**
N_1 is a P_1, V_1.	He S_1 V_2.
N_1 is a P_1. He V_1.	He S_1 V_2.
N_1, a P_1 by trade, V_1.	He S_1 V_2.
N_1, a P_1 by trade. he V_1.	He S_1 V_2.
N_1 works as a P_1, V_1.	He S_1 V_2.
N_1 works as a P_1. He V_1.	He S_1 V_2.
N_1, recognized as a P_1, V_1.	He S_1 V_2.
N_1 is recognized as a P_1. He V_1.	He S_1 V_2.

Table 7: This table presents six examples in our constructed benchmark.

A Templates

We provide the templates that are used for constructing out benchmark in Table 7.

B Speculative Vocabulary

Here we provide the full list of the speculative vocabulary that are used for constructing our benchmark, including “could", “might", “probably", “presumably", “must", “may".

C More Experiments about Bias Polarity Study

We provide the experimental results for exploring the bias polarity which is shown in Table 8. For the

GPT-4o	entailment	neutral	contradiction
low lexical overlap	30.3	**37.5**	32.2
high lexical oerlap	**40.9**	24.6	34.5
label distribution	33.3	33.3	33.3
GPT-4o	entailment	neutral	contradiction
speculative word exists	**40.1**	27.7	32.2
speculative word absents	32.8	34.8	32.4
label distribution	33.3	33.3	33.3
GPT-4o	entailment	neutral	contradiction
high semantic similarity	**40.3**	27.2	32.5
low semantic similarity	30.6	**37.6**	31.8
label distribution	33.3	33.3	33.3
GPT-4o	entailment	neutral	contradiction
male with male-biased occupations	**53.8**	11.5	34.7
male with female-biased occupations	33.7	16.1	**51.2**
label distribution	33.3	33.3	33.3

Table 8: This table presents the distribution of the predicted labels by GPT-4o-2024-11-20 and in the dataset.

bias type of gender-occupation, we follow (Anantaprayoon et al., 2024) to devise datasets for experiments due to the difficulty in automatically extracting data containing this type of bias from existing datasets. For the other types of biases, we construct the dataset for experiments by adopting the same methodology as illustrated in Sec. 2.

D Prompts

We provide the specific prompts used for zero-shot and few-shot experiments in Figure 5.

Zero-shot: Examine the pair of sentences and determine if they exhibit entailment, neutral, or contradiction. Answer with either 'entailment', 'neutral', or 'contradiction':
Few-shot 1: Examine the pair of sentences and determine if they exhibit entailment, neutral, or contradiction. Answer with either 'entailment', 'neutral', or 'contradiction': Premise: I am asserting my membership in the club of Old Geezers. Hypothesis: I am proclaiming that I am now a member of the club of Old Geezers. Answer: 'entailment' Premise: Jon's feeling of age and weariness must have shown. Hypothesis: Jon had traveled longer than his body could handle. Answer: 'neutral' Premise: oh really it wouldn't matter if we plant them when it was starting to get warmer. Hypothesis: It is better to plant when it is colder. Answer: 'contradiction'
Few-shot 2: Examine the pair of sentences and determine if they exhibit entailment, neutral, or contradiction. Answer with either 'entailment', 'neutral', or 'contradiction': Premise: uh i don't know i i have mixed emotions about him uh sometimes i like him but at the same times i love to see somebody beat him. Hypothesis: I like him for the most part, but would still enjoy seeing someone beat him. Answer: 'entailment' Premise: The new rights are nice enough. Hypothesis: Everyone really likes the newest benefits. Answer: 'neutral' Premise: This site includes a list of all award winners and a searchable database of Government Executive articles. Hypothesis: The Government Executive articles housed on the website are not able to be searched. Answer: 'contradiction'
Few-shot 3: Examine the pair of sentences and determine if they exhibit entailment, neutral, or contradiction. Answer with either 'entailment', 'neutral', or 'contradiction': Premise: New York Times columnist Bob Herbert asserts that managed care has bought Republican votes and that patients will die as a result. Hypothesis: Managed care bought Republican votes and patients will end up dead because of this. Answer: 'entailment' Premise: Dirt mounds surrounded the pit so that the spectators stood five or six people deep around the edge of the pit. Hypothesis: The hole is seven feet deep. Answer: 'neutral' Premise: There are many homes built into the hillsides; some have been converted into art galleries and shops selling collectibles. Hypothesis: All of the homes in the hillside have been converted into art galleries and shops selling collectibles. Answer: 'contradiction'

Figure 5: Prompts utilized in out experiments.