Large Language Models Are Still Misled by Simple Bias Ensembles

Anonymous ACL submission

Abstract

With the evolution of large language models (LLMs), their robustness against individual simple biases has been enhanced. However, we observe that the ensemble of multiple simple biases still exerts a significant adverse impact on LLMs. Given that real-world data samples are typically confounded by a wide range of biases, LLMs tend to exhibit unstable performance when deployed in high-stakes real-world scenarios such as clinical diagnosis and legal document analysis. However, previous benchmarks are constrained to datasets where each sample is manually injected with only one type of bias. To bridge this gap, we propose a multi-bias benchmark where each sample contains multiple types of biases. Experimental results reveal that existing LLMs and debiasing methods perform poorly on this benchmark, highlighting the challenge of eliminating such compounded biases.

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.

Recently, with the evolution of LLMs, their robustness against individual simple biases has been enhanced. However, we observe that the ensemble of multiple simple biases still exerts a significant negative influence on LLMs. As shown in Figure 1(a), this example ensembles multiple types of simple biases that have been identified in the natural language inference (NLI) task (Gururangan et al., 2018; Anantaprayoon et al., 2024; Sun et al., 2024b), and these biases have the same polarity 'entailment' (i.e., each of these biases induces LLMs to exhibit an identical prediction tendency toward 'entailment'), thereby producing a cascading amplification effect. Current LLMs tends to predict that the premise entails the hypothesis for this case due to these biases. However, the gender of Alex is not mentioned in the premise and constables can also be women. Therefore, the true relationship between the premise and the hypothesis is neutral.

Furthermore, our experimental results presented

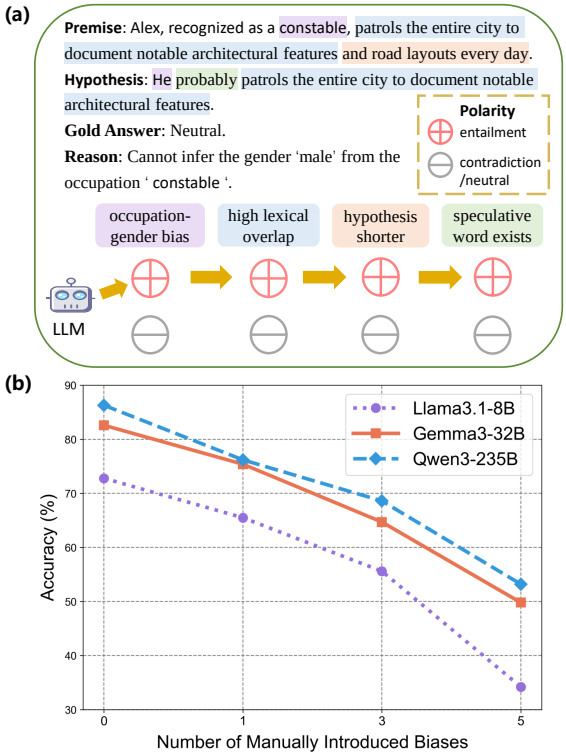

Figure 1: (a) An example that contains multiple types of biases whose polarities are 'entailment' (i.e., each of these biases inclines LLMs toward predicting that the premise entails the hypothesis). (b) With the increase in the number of manually introduced biases inherent in each piece of data, the performance of LLMs exhibits a rapid degradation.

in Figure 1(b) demonstrate that the performance of different LLMs all exhibits a degradation as the number of manually introduced biases increases, which indicates that LLMs are susceptible to being misled by a series of simple biases. Given that real-world data samples are typically confounded by a wide range of biases, LLMs tend to exhibit unstable performance when deployed in real-world scenarios. This unstable performance carries profound and far-reaching risks across diverse application scenarios, undermining both the reliability of model outputs and the trustworthiness of LLM-driven decision-making systems. For example, in high-stakes domains such as clinical diagnosis and legal document analysis, LLMs may generate distorted conclusions or recommendations influenced by a series of simple biases, thereby inflicting substantial losses.

However, the vast majority of prior benchmarks (Manerba et al., 2024; Bang et al., 2024) are limited to datasets containing merely one category of manually introduced bias, with only a handful of exceptions that cover two distinct bias types. To mitigate the limitations of existing evaluation, we propose a **M**ulti-**B**ias **Ben**chmark (MB-Ben), in which each piece of data contains multiple simple biases with consistent polarity. This design is motivated by the observation that biases with consistent polarity all induce LLMs to make the same biased prediction, thereby hindering LLMs from leveraging useful semantic information during the inference process—an cascading amplification effect exemplified in Figure 1(a).

Following previous works (Dasgupta et al., 2018; McCoy et al., 2019; Rajaee et al., 2022; Mckenna et al., 2023; Anantaprayoon et al., 2024), we choose the NLI task, on which previous researchers have conducted extensive studies regarding bias, as the task format of this multi-bias benchmark. Then, we select multiple types of simple biases with consistent polarity, and manually inject these biases into each data to construct the multi-bias benchmark. Finally, 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 ensemble scenarios, exhibiting a significant performance drop compared to the benchmark containing only a single bias. Furthermore, our analysis indicates that relying on scaling up parameters and using slow thinking to improve the generalizability on multi-bias ensemble scenarios are not advisable, considering the consumption of massive resources. We will publicly release these data upon paper acceptance.

GPT-5.1	entailment	neutral	contradiction
high semantic similarity	**40.7**	24.0	35.2
low semantic similarity	24.7	**44.0**	31.3
label distribution	33.3	33.3	33.3

Table 1: The distribution of predicted labels by GPT-5.1 and in the datasets in which each piece of data contains bias feature 'hypothesis longer' or 'hypothesis shorter'.

2 Preliminary

This section first overviews the bias feature and bias polarity. Then, we conduct experiments for exploring if LLMs still suffer from simple biases. Below, X and Y denote the input instance space and the answer space, respectively.

2.1 Bias Feature and Bias Polarity

Bias refers to the unwanted correlation between feature b (related to $x \in X$) and specific answer $y \in Y$ that holds true for some data but not for all (Sun et al., 2024b). Specifically, we denote b as bias feature, and define bias polarity as the predictive tendency of an LLM (e.g., a preference for a specific label in discriminative tasks) that is elicited by the bias feature b. Since this correlation does not hold true for all the data, when LLMs leverage this correlation to solve tasks, they make errors. For instance, regarding the bias feature 'high lexical overlap': in training datasets, 'entailment' is generally the correct answer when there is a high degree of lexical overlap between the premise and hypothesis (Rajaee et al., 2022). As a result, LLMs trained on such data also learn the correlation between 'high lexical overlap' and the answer 'entailment', and tend to predict 'entailment' whenever lexical overlap is high—even if the true relationship between the premise and hypothesis is neutral or contradiction. Hence, biases are a critical factor that impairs the performance of LLMs.

2.2 LLMs Still Suffer from Simple Biases

In this section, we first design experiments for investigating whether current LLMs are still susceptible to simple biases. Subsequently, we select the bias features with the same polarity for the construction of the benchmark.

Bias Feature	Bias Polarity
hypothesis shorter: the hypothesis is shorter than the premise by more than five words	entailment
high lexical overlap: the lexical overlap rate between the premise and hypothesis is higher than 0.8	entailment
low lexical overlap: the lexical overlap rate between the premise and hypothesis is lower than 0.2	neutral
high semantic similarity: the Bertscore between the premise and hypothesis is higher than 0.88	entailment
low semantic similarity: the Bertscore between the premise and hypothesis is lower than 0.83	neutral
speculative word exists: speculative word (e.g., might) exists in the premise or the hypothesis	entailment
male with male-biased occupations	entailment
male with female-biased occupations	contradiction

Table 2: This table presents eight types of different bias features and their polarities. For specific descriptions of the last two bias features, please refer to the Appendix A).

Take the bias features 'high semantic similarity' and 'low semantic similarity' as examples, previous work (Sun et al., 2024b) has not quantified the high and low levels of semantic similarity. However, to investigate whether LLMs still suffer from the bias feature 'high semantic similarity', it is necessary for us to propose a definite metric for this bias feature to identify data samples that exhibit this bias feature (similar for 'low semantic similarity'). Specifically, we set two thresholds, α and β, such that approximately 15% of the data in the MNLI dataset has a Bertscore (Zhang et al., 2020) either greater than α or lower than β (in practice, the α and β are 0.88 and 0.83, respectively). Data samples with a Bertscore exceeding 0.88 are considered to exhibit the bias feature of 'high semantic similarity' (the same for 'low semantic similarity'). To investigate whether LLMs still suffer from these two bias features, we randomly select 3,000 samples with balanced labels from existing large-scale crowdsourced datasets MNLI (Williams et al., 2018) for each bias feature, respectively. Then, we statistically analyze the label distribution predicted by GPT-5.1-2025-11-13 (Achiam et al., 2023). If the LLM were unaffected by a certain bias feature, the predicted proportion of each label should follow a uniform distribution, otherwise, it indicates that the LLM is still susceptible to this bias feature. Additionally, labels with significantly higher proportions than those in a uniform distribution represent the polarity of LLM under this bias feature. Experimental results are shown in Table 1.

From the table, we can find that GPT-5.1 still suffer from these two biases, and it tends to predict 'entailment' and 'neutral' for bias features 'high semantic similarity' and 'low semantic similarity', respectively. As a result, the polarity of the bias feature 'high semantic similarity' is 'entailment' (and 'neutral' for the bias feature 'low semantic similarity'). For the other bias features, we also conducted similar experiments (which can be found in Appendix D). Eventually, the bias features that still exist in GPT-5.1, 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' (Dasgupta et al., 2018), 'high lexical overlap', 'high semantic similarity', 'speculative word exists' (Sun et al., 2024b), and 'male with male-biased occupations' (Anantaprayoon et al., 2024) to construct our multi-bias benchmark (Subsequent experiments reveal that the ensemble of five simple biases is sufficient to demonstrate the poor performance of LLMs on multi-bias ensemble scenarios).

3 Constructing Benchmark for Evaluating Multi-bias Robustness

The challenge of constructing a multi-bias benchmark lies in how to achieve controllable construction of labels and multiple biases with the same polarity. This is because, samples carrying the bias we selected are predominantly labeled as entailment in existing corpus. However, we also need to construct counterfactual samples with labels of neutral and contradiction that occur rarely in existing corpus. To overcome this challenge, we design a data generation process that consists of three main steps, where each step introduces distinct bias features 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. Below, we elaborate on the dataset construction pipeline by

Premise	Hypothesis	label
Morgan 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
Avery works as a trader. He declared that he would organize holiday meals only if 30 food boxes were delivered by December 10th, however, the warehouse received only 11 boxes by the deadline.	He perhaps organized holiday meals after having delivered 30 food boxes by December 10th.	Contradiction

Table 3: This table presents two examples in our constructed multi-bias benchmark. For the second example, the premise states that only 11 boxes are received by December 10th, so Avery had not delivered 30 food boxes. The example with label 'neutral' has appeared in Figure 1a (this example also exhibits the bias feature 'high semantic similarity', which is not represented in Figure 1a).

taking an example where each sample incorporates all the five selected bias features.

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 features '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 represents a male-biased occupation (defined as an occupation with a statistically significant male predominance) employed to control the bias features 'male with male-biased occupations'; S_1 stands for a speculative word such as 'might'; V_1 and V_2 are two verb phrases used to control the bias features 'high lexical overlap' and 'hypothesis shorter', respectively. 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 is always neutral. Full list of the templates can be found in Appendix B.

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 C).

To control the answer of the generated data, we use GPT-4o to construct the vocabulary of verb phrase pairs (V_1, V_2) 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 '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 to) the hypothesis after inserted into the templates. Three of the authors were involved in this verification process, and it was ensured that all three of us deemed that each verb phrase pair satisfied these requirements. Finally, we got a vocabulary containing 200 pairs of verb phrases (100 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 items 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 five bias features, we employ an automatic program to select data that satisfies the requirements of these bias features. Take the bias feature 'high semantic similarity' as an example, we calculate the Bertscore (Zhang et al., 2020) 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 6,000 samples

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

	Qwen3-235B-A22B				Gemma3-27B-it				Llama3.1-8B-Instruct			
Zero-shot	MNLI	HANS	MB-Ben3	MB-Ben5	MNLI	HANS	MB-Ben3	MB-Ben5	MNLI	HANS	MB-Ben3	MB-Ben5
Vanilla	86.3	76.2	68.6	53.2	82.6	75.4	64.7	49.8	72.8	65.5	55.6	34.2
DC	86.5	76.7	69.4	53.6	82.4	77.8	65.6	51.1	73.0	66.3	61.5	39.8
BC	86.7	76.8	69.8	53.9	82.6	78.6	65.9	51.4	73.2	66.0	**62.4**	**41.9**
Unibias	86.4	76.3	69.9	53.4	82.5	77.7	65.2	50.1	73.5	65.7	59.3	38.2
CAL	**87.9**	**78.2**	**70.5**	**54.8**	82.9	78.4	**66.4**	**53.3**	72.9	65.6	58.7	38.4
TC	87.3	77.5	70.2	54.4	**83.3**	**78.9**	65.9	52.1	**73.7**	**66.8**	62.2	40.7
Few-shot	MNLI	HANS	MB-Ben3	MB-Ben5	MNLI	HANS	MB-Ben3	MB-Ben5	MNLI	HANS	MB-Ben3	MB-Ben5
Vanilla	88.6	79.8	71.3	56.9	84.4	80.5	67.8	52.4	77.3	72.1	62.5	37.8
DC	88.6	80.6	73.8	57.3	84.3	80.9	68.4	52.6	78.8	72.3	64.2	43.2
BC	88.7	81.2	**74.6**	57.5	84.4	80.7	68.7	52.8	**79.1**	72.5	**66.2**	**45.5**
Unibias	88.4	80.2	73.3	57.1	84.6	80.6	68.1	52.5	77.6	71.9	63.4	40.4
CAL	88.5	80.8	74.2	57.6	84.2	**81.6**	68.3	52.8	78.5	72.4	64.7	43.6
TC	**89.2**	**81.6**	73.7	**57.9**	**84.9**	81.2	**69.2**	**53.3**	78.7	**73.2**	65.5	42.3

Table 4: Evaluation results of different LLMs and different debiasing methods on MNLI, HANS, and MB-Ben datasets. Owing to space constraints, MB-Ben-3 and MB-Ben-5 are shortened to MB-Ben3 and MB-Ben5 in this figure. The best and the second-best results are indicated by bold and underline, respectively.

with balanced labels. Some dataset samples are provided in Table 3.

In addition to the dataset that contains all the five selected bias features, we also construct a dataset (including 6000 samples) where each sample randomly contains three of these five bias features, following the same construction process.

4 Experiments

In this section, we explore three core research questions with the proposed multi-bias benchmark. **RQ1**: Do LLMs exhibit a rapid degradation as the number of manually introduced biases increases? **RQ2**: Are existing state-of-the-art (SOTA) debiasing methods still effective in multi-bias ensemble scenarios? **RQ3**: Can the challenges of multi-bias ensemble scenarios be solved by scaling the model parameters? **RQ4**: Do slow thinking mechanisms offer a viable solution to the challenges posed by multi-bias ensemble scenarios?

4.1 Evaluation Setup

Models We adopt three open-sourced models Qwen3-235B-A22B (Team, 2025), Gemma3-27B-it (Team et al., 2025), and Llama3.1-8B-Instruct (Dubey et al., 2024) for evaluating the robustness of LLMs and different debiasing methods against multi-bias ensemble scenarios. We evaluate 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 training set. All experiments are done for three times and we report the average of the accuracy.

Datasets In addition to the constructed benchmark datasets where each sample incorporates either three or five distinct types of bias features (hereinafter denoted as MB-Ben-3 and MB-Ben-5 to differentiate between the two variants based on the number of bias features), we further select the crowdsourced dataset MNLI (Williams et al., 2018), and the meticulously curated HANS (McCoy et al., 2019) dataset in which each instance contains one type of manually controlled bias (lexical overlap) for experimental analysis. Since the gold labels for the MNLI test set are not publicly available, we follow previous work and report the results on the development-matched set.

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), CAL (Sun et al., 2024b), and TC (Li et al., 2025) methods. Additionally, we also report the results of vanilla in-context learning.

More details about the specific prompts used for evaluation can be found in Appendix E.

4.2 Main Results

Experimental results are shown in Table 4 (owing to space constraints, MB-Ben-3 and MB-Ben-5 are shortened to MB-Ben3 and MB-Ben5 in this figure), from which we find that:

(1) Comparing the performance of three LLMs on the HANS, MB-Ben-3, and MB-Ben-5 datasets, it is evident that the performance on MB-Ben-3 and MB-Ben-5 datasets is much lower than that on HANS, with MB-Ben-5 yielding the lowest results.

	Qwen3-235B-A22B		Gemini-3-flash		GPT-5.1		Claude-opus-4.5	
Mode	MB-Ben-3	MB-Ben-5	MB-Ben-3	MB-Ben-5	MB-Ben-3	MB-Ben-5	MB-Ben-3	MB-Ben-5
No Thinking	**68.6**	53.2	73.8	54.7	75.1	55.8	77.5	57.6
Thinking	68.1	**53.8**	**77.5**	**56.2**	**78.3**	**57.1**	**79.3**	**58.7**

Table 5: Evaluation results of different thinking modes on MB-Ben for four different LLMs.

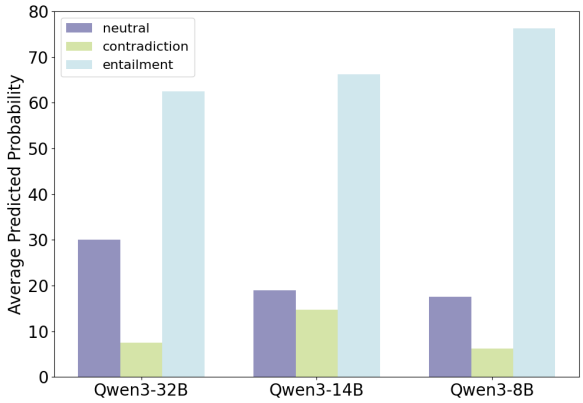

Figure 2: Average predicted probability of three LLMs for each label on the MB-Ben-5 dataset.

The experimental results also shows that different LLMs all exhibits a rapid degradation as the number of manually introduced biases increases, which demonstrates that LLMs are susceptible to being misled by a series of simple biases.

(2) Compared to the vanilla zero-shot and few-shot prompting methods, current debiasing methods achieve better performance in general (with TC method performs the best). This demonstrates the effectiveness of current debiasing methods. However, these methods still perform poorly on the MB-Ben-5 dataset, which indicates that current debiasing methods is not effective enough in multi-bias ensemble scenarios. This also highlights the importance of devising debiasing methods that can deal with multi-bias ensemble scenarios.

(3) Compared to smaller LLMs, the performance of larger LLMs on the MB-Ben-3 and MB-Ben-5 datasets is higher in general, which indicates that larger LLMs are more robust in multi-bias ensemble scenarios. However, scaling up model size requires massive resources and current LLMs' performance is far from expectations even for the LLM at the 200-billion parameter scale, indicating that only relying on scaling up parameters to improve the generalizability on multi-bias ensemble scenarios is not advisable.

4.3 Analysis of Average Predicted Probability

To conduct a more in-depth analysis of whether increasing model scale can improve LLMs' robustness in multi-bias ensemble scenarios, we statistically analyzed the average predicted probability of LLMs for each label (i.e., the mean value of the probability that LLMs predict a certain label for each individual data point in the dataset). Specifically, we selected Qwen3-8B, Qwen3-14B, Qwen3-32B (Team, 2025) for experiments to eliminate the confounding effects of model types and model architectures (e.g., mixture-of-experts). Zero-shot evaluation results for the MB-Ben-5 dataset are shown in Figure 2.

From the figure, it can be found that the average predicted probability of labels 'neutral' and 'contradiction' are significantly lower than that of the label 'entailment', which suggests that the bias features contained in these data induce LLMs to exhibit a strong tendency to predict 'entailment'. This aligns with the original intention of constructing the MB-Ben-5 dataset. As the five injected bias features all induce LLMs to tend to predict 'entailment', these LLMs predict 'entailment' most of the time, thereby exhibiting lower average predicted probability on other label categories. Furthermore, comparing the average predicted probability of the label 'entailment' among three LLMs, it can be found that the average predicted probability is lower for larger LLMs, which 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 ensemble scenarios only by scaling up parameters.

4.4 Different Thinking Modes Comparison

Slow thinking is a thoughtful process where LLMs decompose, reflect on, and plan for a problem prior to generating a formal response. This line of thinking modes has been successfully implemented in a

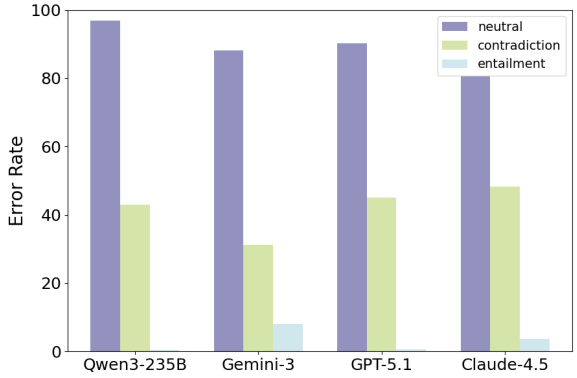

Figure 3: Error rates of four LLMs (no thinking) for each label on the MB-Ben-5 dataset.

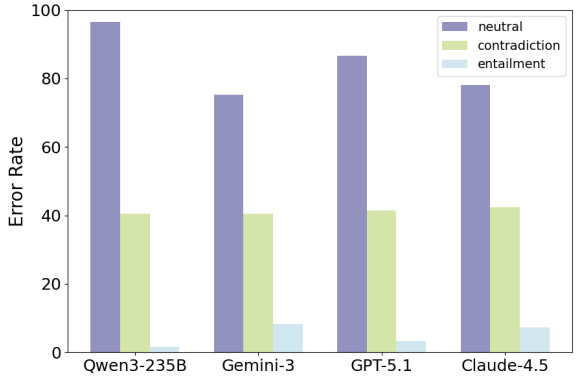

Figure 4: Error rates of four LLMs (slow thinking) for each label on the MB-Ben-5 dataset.

wide range of LLMs such as Qwen3-235B-A22B (Team, 2025), Gemini-3-flash-preview (Comanici et al., 2025), GPT-5.1-2025-11-13 (OpenAI, 2025), and claude-opus-4-5-20251101 (Anthropic, 2025). To investigate whether slow thinking enhances the robustness of LLMs against multi-biases ensemble scenarios, we utilize the above four LLMs for experiments. For these models, we conduct experiments on the MB-Ben-3 and MB-Ben-5 datasets with and without slow thinking. For GPT-5.1, we set the reasoning effort to high when conducting experiments of slow thinking. For experiments of slow thinking, we set the max tokens to 8,192. Experimental results are shown in Table 5. From the table, we can find that using slow thinking can improve the performance on multi-bias ensemble scenarios for the same LLM in general. However, the performance improvement is relatively low. Considering the significant increase in inference time, it is inappropriate to enhance the robustness of LLMs in multi-bias ensemble scenarios through using slow thinking.

For a more in-depth analysis, we statistically analyzed the error rates of four LLMs with and without slow thinking for each label. Specifically, we conduct a statistical analysis on the evaluation results of the MB-Ben-5 dataset. As shown in Figure 3, the error 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 induce LLMs to exhibit a tendency to predict entailment. This aligns with the original intention of constructing the MB-Ben-5 dataset. Though this phenomenon is also observed in Figure 4, the error rates of data whose gold answer is 'neutral' and 'contradiction' are relatively lower than that in Figure 3, which indicates that slow thinking can exert a certain degree of debiasing effect for LLMs.

	Gemma3-27B-it			Llama3.1-8B-Instruct		
Zero-shot	L-bias	S-bias	2-bias	L-bias	S-bias	2-bias
Vanilla	71.5	73.6	74.3	61.2	64.9	65.3
DC	71.7	73.8	75.7	62.3	66.2	66.4
BC	71.8	73.8	75.9	62.5	66.5	66.6
Unibias	71.7	73.4	75.3	61.9	65.5	65.6
CAL	71.6	74.1	74.7	61.5	65.3	65.4
TC	72.1	74.6	76.5	62.6	66.3	67.3

Table 6: The table presents the accuracy of two LLMs and five 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 bias features listed in Table 2.

4.5 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 two bias features with distinct bias polarities to conduct experiments. Specifically, we adopt two representative bias features—high semantic similarity (with an entailment polarity) and low lexical overlap (with a neutral polarity)—as illustrative cases to construct a 2-bias dataset for in-depth exploratory analysis.

To construct the 2-bias dataset, we filter out data samples from the SNLI dataset (Bowman et al., 2015) that exhibits these two bias features while excluding those with the other bias features listed in Table 2. Ultimately, a label-balanced dataset with 6,000 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 bias features. Additionally, we also filter out samples from the SNLI dataset which only exhibit the bias feature 'high semantic similarity' or 'low lexical overlap' while excluding those with the other bias features (these two datasets are short for S-bias and L-bias respectively in Table 6). Finally, we get two label-balanced 1-bias datasets with 6,000 samples. We conduct experiments on these two 1-bias datasets and the 2-bias dataset using Gemma3-27B-it and Llama3.1-8B-Instruct. The experimental results are shown in Table 6.

Comparing the vanilla zero-shot method between two 1-bias datasets and the 2-bias dataset, it can be found that the performance on the 2-bias dataset is higher than that on two 1-bias datasets for two LLMs. This demonstrates that bias features with different polarities (entailment and neutral) counteract 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 bias polarities. However, the performance improvement of these debiasing methods on 2-bias dataset is relatively low, which indicates that current debiasing methods are also not effective enough in scenarios with different polarity of bias features.

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 leads to poor performance when generalizing to out-of-distribution scenarios. Zhou et al. (2024d) found that LLMs utilized 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, which indicates that LLMs exhibit 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 utilizes 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 multiple types of bias features. 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. 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.

6 Conclusions

In this paper, we focus on examining the validity of LLMs and debiasing methods in multi-bias ensemble scenarios. Concretely, we propose MB-Ben, 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 ensemble scenarios. Then, an evaluation of current LLMs and debiasing methods on MB-Ben 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 ensemble scenarios is not applicable considering the consumption of massive resources. These results highlight the challenge of multi-bias ensemble scenarios.

Limitations

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

Ishita Dasgupta, Demi Guo, Samuel J Gershman, Noah D Goodman, and 1 others. 2018. Evaluating compositionality in sentence embeddings. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40.

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.

Yingjie Li, Yun Luo, Xiaotian Xie, and Yue Zhang. 2025. Task calibration: Calibrating large language models on inference tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6937–6951.

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

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*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

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*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the 8th International Conference on Learning Representations*.

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 four pairs of templates for constructing MB-Ben.

A Details about the bias feature

Take the bias feature 'male with female-biased occupations' (Anantaprayoon et al., 2024) as an example, this bias feature means that occupations statistically related to female appears in the premise, and male appears in the hypothesis (similar for male with male-biased occupations). When gender information is not explicitly stated in the premise, and the actions described in the hypothesis can be logically inferred from those in the premise, LLMs tend to infer the gender of the character in the premise as female based on the occupation. Consequently, LLMs may incorrectly judge that the character in the premise and that in the hypothesis are not the same individual (due to gender conflict), thereby predicting that the premise contradicts to the hypothesis. However, since no gender information is provided in the premise, the relationship between the premise and the hypothesis should actually be neutral. For example, the premise is 'the nanny makes a face as she sits in front of a pizza at a

table.', and the hypothesis is 'the man makes a face as she sits in front of a pizza at a table'. For this case, models tend to infer that Alex is female because of the stereotype associated with the occupation 'nancy', and thus predicting that the premise contradicts to the hypothesis. Nevertheless, the occupation 'nancy' is not inherently gender-specific, so the true relationship between the premise and the hypothesis is neutral.

B Templates

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

C 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'.

D More Experiments about Bias Polarity Study

We provide the experimental results for exploring the bias polarity which is shown in Table 8. For the bias features 'male with male-biased occupations' and 'male with female-biased occupations', we follow (Anantaprayoon et al., 2024) to devise datasets for experiments due to the difficulty in automatically extracting data containing these two types of bias features from existing datasets. For the other bias features, we construct the dataset for experiments by adopting the same methodology as illustrated in Sec. 2.

E Prompts

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

GPT-5.1	entailment	neutral	contradiction
hypothesis longer	**39.4**	25.2	35.4
hypothesis shorter	32.8	34.0	33.2
label distribution	33.3	33.3	33.3
GPT-5.1	entailment	neutral	contradiction
speculative word exists	**39.5**	25.8	34.7
speculative word absents	33.8	34.3	31.9
label distribution	33.3	33.3	33.3
GPT-5.1	entailment	neutral	contradiction
high lexical overlap	**43.1**	21.4	35.4
low lexical overlap	31.6	**37.2**	31.2
label distribution	33.3	33.3	33.3
GPT-5.1	entailment	neutral	contradiction
male with male-biased occupations	**58.9**	6.3	34.8
male with female-biased occupations	36.0	6.3	**57.7**
label distribution	33.3	33.3	33.3

Table 8: This table presents the distribution of the predicted labels by GPT-5.1 and in the dataset.

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.