
Appendix for *Boosting Resilience of Large Language Models through Causality-Driven Robust Optimization*

Xiaoling Zhou
Peking University
xiaolingzhou@stu.pku.edu.cn

Mingjie Zhang
Peking University
mjzhang0621@stu.pku.edu.cn

Zhemg Lee
Tianjin University
zhemglee@tju.edu.cn

Yuncheng Hua
University of New South Wales
devin.hua@unsw.edu.au

Chengli Xing
Peking University
xingchengli@stu.pku.edu.cn

Wei Ye*
Peking University
weye@pku.edu.cn

Flora D. Salim*
University of New South Wales
flora.salim@unsw.edu.au

Shikun Zhang
Peking University
zhangsk@pku.edu.cn

1 Calculation of Model Calibration Metrics

The two model calibration metrics, namely ECE and Brier Score, are detailed as follows:

- In line with prior research [12, 37, 26], we employ the ECE indicator to measure the discrepancy between a model’s confidence and its actual accuracy. Specifically, model predictions are grouped according to confidence levels, and we compute the accuracy $acc(b_i)$ and the average confidence $conf(b_i)$ within each bin b_i . The ECE is then calculated as $ECE = \sum_i \frac{|b_i|}{M} |acc(b_i) - conf(b_i)|$, where M denotes the number of model generations. A lower ECE signifies better calibration, indicating a closer alignment between the model’s confidence and its actual accuracy.
- The Brier Score [3] is a metric commonly used to evaluate tasks that require assigning probabilities to a set of mutually exclusive discrete outcomes or classes, which can be either binary or categorical. Following [26], we compute the Brier Score as the mean squared difference between the model confidence p_y and the binary correctness $I(y)$ of its predictions $Brier = \frac{1}{M} \sum_y [p_y - I(y)]^2$. This metric offers a direct assessment of the quality of model calibration.

2 Prompt Formulation and Design

This section presents representative examples of prompts employed during the data collection process. Specifically, the prompts utilized for generating counterfactual samples in NLU and NLG tasks are illustrated in Figs. 1 and 2, respectively. For NLG tasks, a negation-based strategy is adopted to generate counterfactual samples in a straightforward and interpretable manner, wherein explicit

*Corresponding authors.

Task Definition: Revise a given sentence with minimal changes to alter its sentiment polarity.

Instruction: This process consists of two steps. The first step is to identify the words in the given sentence that have the highest potential to change the sentiment polarity after substitution, known as the causal words. The second step is to select appropriate replacement words for the causal words that will change the sentiment polarity of the sentence to the desired sentiment polarity. Make sure the given sentence and the revised sentence have opposite sentiment polarities and talk about the same topic. Only output the revised sentence.

Figure 1: Prompt utilized for generating counterfactual instances in NLU tasks, exemplified by the SST-2 dataset.

Task Definition: Revise the given question by adding a negation at a specific point in the sentence while preserving the focus word or concept, and generate a reasonable and correct answer, ensuring that the revised question and the original question are based on exactly the same specific knowledge.

Instruction:

1. Identify the specific knowledge being questioned.
2. Generate a revised question that:
 - Is based on exactly the same specific knowledge as the original question (e.g., the same event, person, entity, or fact).
 - Introduces a reversal or negation by adding a negation at a specific point in the sentence (e.g., 'did' to 'did not') or replacing a word with its antonym or opposite concept (e.g., 'won' to 'lost') to inquire about an opposite or contrary aspect of the knowledge point.
 - Leads to a distinctly different answer compared to the original.
 - Is grammatically valid and factually meaningful.
3. Create a corresponding answer that:
 - Matches the original answer's format type exactly (e.g., a year, a person's name, a percentage, etc.).
 - Is factually correct and distinct from the original answer.

Figure 2: Prompt utilized for generating counterfactual instances in NLG tasks.

negation is introduced into the original question. For instance, given the original query "Who was the first scientist to produce electromagnetic waves in a laboratory?", a corresponding counterfactual variant would be "Who was not the first scientist to produce electromagnetic waves in a laboratory?". In such cases, any scientist other than the ground-truth answer provided in the original instance may be considered a plausible response.

Moreover, the prompt employed for generating paraphrased samples in both NLU and NLG tasks is illustrated in Fig. 3. The prompts utilized to evaluate the counterfactual and paraphrased samples are presented in Figs. 4 and 5, respectively. With changes in the dataset, the aforementioned prompts may undergo some adjustments according to their unique characteristics. For instance, in the case of the CoLA dataset, it is reasonable for augmented samples to exhibit grammatical errors when the target label is "unacceptable." Furthermore, several demonstrations will be incorporated following the given prompt to more effectively guide the output generation of LLMs.

You are a highly skilled paraphrasing agent with an exceptional ability to rephrase sentences while preserving their original meaning and context. Your task is to take the given sentence and transform it into a new version that is both coherent and contextually accurate. Feel free to enhance the sentence with relevant details or insights that align with the original intent, showcasing your ability to enrich the content subtly. Only output the new sentence.

Figure 3: Prompt utilized for generating paraphrase instances in both NLU and NLG tasks.

System: You are an unbiased and professional assistant for evaluating the quality of counterfactual samples.

User: Please evaluate the following counterfactual samples based on the original sample and provide a score between 1 and 10, with 10 being the highest score.

The criteria for evaluation are as follows:

1. **Answer Alteration**: For a counterfactual sample, the answer must be different from the original sample's answer.
2. **Answer Correctness**: The answer of the counterfactual sample should be correct given the input.
3. **Thematic Consistency**: The counterfactual sample should discuss a theme or topic that is the same as the original sample.
4. **Clarity of Expression**: The counterfactual sample should be clear and logically coherent in its expression.
5. **Safety and Privacy**: The counterfactual sample should not include any content that may be deemed unsafe or pose a privacy risk.

Score guidance:

- **10**: The counterfactual sample meets all criteria perfectly: the answer is changed and correct, the theme is consistent, the expression is clear, and there are no safety or privacy issues.
- **5**: The counterfactual sample meets some criteria but has moderate issues.
- **0**: The counterfactual sample fails to meet any criteria, especially if the answer is not changed and is not correct, the theme is not consistent, the expression is unclear, and there are significant safety or privacy issues.

Please provide your score first, followed by a brief explanation of your reasons. Be concise and specific.

Please answer in the following format:

Score: "Your rating score for the counterfactual sample."

Explanation: "Your explanation for your rating."

Figure 4: Prompt utilized for evaluating the generated counterfactual samples.

3 Additional Details on Debiasing Experiments

3.1 More Details about Experimental Settings

Consistent with previous studies, we utilize manually curated lists of stereotype and attribute words to analyze biases in LLMs. Specifically, we employ the gender attribute and target word lists proposed in [19], which are widely used in debiasing research [13, 24]. Additionally, we incorporate the race attribute words and other attribute words provided in [28].

The metrics for measuring debiasing ability we consider include SEAT [29] and CrowS-Pair [31]. An ideally unbiased model should exhibit no variation in relative similarity. In line with previous studies [13, 24, 19], we apply SEAT tests 6, 6b, 7, 7b, 8, and 8b to assess gender bias, and use SEAT tests 3, 3b, 4, 5, and 5b for evaluating racial bias. We report the effect size in the SEAT evaluation, with values closer to 0 indicating lower bias in the model.

System: You are an unbiased and professional assistant for evaluating the quality of paraphrase samples.

User: Please evaluate the following paraphrase samples based on the original sample and provide a score between 1 and 10, with 10 being the highest score.

The criteria for evaluation are as follows:

1. ****Answer Correctness****: The answer of the paraphrased sample should be correct given the input.
2. ****Thematic Consistency****: The paraphrase sample should discuss a theme or topic that is the same as the original sample.
3. ****Clarity of Expression****: The paraphrase sample should be clear and logically coherent in its expression.
4. ****Safety and Privacy****: The paraphrase sample should not include any content that may be deemed unsafe or pose a privacy risk.

Score guidance:

- ****10****: The paraphrase sample meets all criteria perfectly: the answer is correct, the theme is consistent, the expression is clear, and there are no safety or privacy issues.
- ****5****: The paraphrase sample meets some criteria but has moderate issues.
- ****0****: The paraphrase sample fails to meet any criteria, especially if the answer is not correct, the theme is not consistent, the expression is unclear, and there are significant safety or privacy issues.

Please provide your score first, followed by a brief explanation of your reasons. Be concise and specific.

Please answer in the following format:

Score: "Your rating score for the paraphrase sample."

Explanation: "Your explanation for your rating."

Figure 5: Prompt utilized for evaluating the generated paraphrase samples.

Table 1: The SEAT test details derived from [4].

Bias type	Test	Demographic-specific words	Stereotype words
Racial	SEAT-3	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-3b	European-American/African American terms	Pleasant vs. Unpleasant
	SEAT-4	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-5	European-American/African American names	Pleasant vs. Unpleasant
	SEAT-5b	European-American/African American terms	Pleasant vs. Unpleasant
Gender	SEAT-6	Male vs. Female names	Career vs. Family
	SEAT-6b	Male vs. Female terms	Career vs. Family
	SEAT-7	Male vs. Female terms	Math vs. Arts
	SEAT-7b	Male vs. Female names	Math vs. Arts
	SEAT-8	Male vs. Female names	Science vs. Arts
	SEAT-9b	Male vs. Female terms	Science vs. Arts

The SEAT metric generalizes the WEAT metric [1] by substituting word embeddings with simple sentence templates (e.g., "This is the <word>"), and the calculation process for the WEAT metric is outlined as follows. This metric measures bias by comparing two sets of attribute words, W_a (e.g., M and F), and two sets of target words, W_t (e.g., A and B). In the context of gender bias, M represents masculine words such as "he," while F represents feminine words such as "she." Meanwhile, A and B are gender-neutral words (e.g., career-related terms or adjectives) whose embeddings should be equivalent between M and F . Formally, the degree of bias for each word w is defined as follows:

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(w, a) - \frac{1}{|B|} \sum_{b \in B} \cos(w, b), \quad (1)$$

Table 2: Comparison of gender debiasing performance on the SST-2 dataset.

Metric	SEAT (\downarrow)	CrowS (\rightarrow 50%)	Acc. (\uparrow)
BERT	0.29	55.17%	92.4%
CDA	0.47	58.43%	81.3%
Dropout	0.48	44.53%	81.9%
Context-Debias	0.23	58.80%	91.9%
Auto-Debias	0.28	44.92%	92.1%
MABEL	0.35	46.75%	92.2%
Sent-Debias	0.21	55.06%	89.1%
FairFil	0.18	52.92%	91.6%
Causal-Debias	0.11	48.95%	92.9%
PCFR	0.09	50.68%	91.9%
CdRO (Ours)	0.05	50.36%	94.2%

where $\cos(\cdot, \cdot)$ denotes the cosine similarity. Based on Eq. (1), the SEAT effect size is:

$$d_{\text{WEAT}} = \frac{\mu(\{s(m, A, B)\}_{m \in M}) - \mu(\{s(f, A, B)\}_{f \in F})}{\sigma(\{s(t, A, B)\}_{t \in A \cup B})}, \quad (2)$$

where μ and σ denote the mean and standard deviation, respectively. As can be inferred from Eq. (2), an absolute SEAT effect size closer to 0 indicates lower bias. Detailed information about the SEAT tests used in our experiments is provided in Table 1, which is adapted from [4].

Besides SEAT, we also employ CrowS-Pair [31] as an additional metric for evaluating gender bias. CrowS-Pair is a dataset consisting of 1,508 examples that encompass various types of biases. Each example in the dataset is a pair of stereotype and anti-stereotype sentences with minimal semantic differences. A CrowS-Pair score closer to 50% indicates a lower degree of stereotyping, suggesting that the model assigns comparable probabilities to both male- and female-oriented sentences.

We implement all three models, including BERT-base, RoBERTa-base, and ALBERT-large, using the Transformers library [42]. The experimental settings for all baseline methods follow the configurations specified in their original papers. All reported results represent the average performance over five independent runs. For parameter-efficient fine-tuning, we adopt the PiSSA method [30], with the rank hyperparameter set to 8, following the setup described in the original work. For the three models employed in these experiments, only the query, key, and value matrices are designated as parameters that can be tuned. All experiments are conducted on an NVIDIA A100 GPU with 80GB of memory. For the implementation of the proposed CdRO approach, the learning rates for the logistic regression model and the causal parameter components within the LLM are set to 1×10^{-3} and 4×10^{-4} , respectively. The hyperparameters ϵ and α used in the REINFORCE++ algorithm are set to 0.2 and 0.001, in accordance with the default settings in the TRL library. The hyperparameter γ , which governs the strength of the reward ranking component, is set to 0.1. Meanwhile, λ , which balances the relative importance of different reward components, is fixed at 0.5 across all experiments. We set the number of training epochs to 30 and the batch size to 128. Furthermore, in each iteration, weight matrices with predicted probabilities exceeding the average predicted probability across all matrices are selected as those that encode causal relationships.

During the data collection process, we examine the performance of using both the LLaMA-3-70B [11] and GPT-4o [16] models for data generation and evaluation. Furthermore, to assess the quality of the generated samples, we investigate the impact of using the same generative LLMs for both generation and evaluation, compared to using distinct models for these two stages.

3.2 More Details about Compared Baselines

A variety of debiasing approaches are compared with the proposed CdRO approach, including non-task-specific methods: CDA [40], Dropout [40], Context-Debias [19], Auto-Debias [13], and MABEL [14]—as well as task-specific methods—Sent-Debias [24], FairFil [6], Causal-Debias [46], and PCFR [15]. These compared methods are introduced as follows:

- **CDA** [40] augments the training data using controlled perturbations to attributes including names or demographics.

Table 3: The effectiveness of both LLaMA-3-70B and GPT-4o utilized for data generation and evaluation is investigated. The symbol * indicates that the evaluation of the generated samples is performed using a distinct LLM from the one utilized for generation. For instance, when the augmented samples are generated by LLaMA-3-70B, the evaluation is conducted using GPT-4o.

Dataset	SST-2		CoLA		QNLI	
Method	SEAT (\downarrow)	Acc. (\uparrow)	SEAT (\downarrow)	Mcc. (\uparrow)	SEAT (\downarrow)	Acc. (\uparrow)
BERT	0.30	92.4%	0.16	57.6%	0.30	91.3%
Auto-Debias	0.31	92.1%	0.20	52.9%	0.24	91.1%
Causal-Debias	0.11	92.9%	0.06	58.1%	0.11	91.6%
CdRO (LLaMA-3-70B)	0.06	94.2%	0.04	59.4%	0.07	92.8%
CdRO* (LLaMA-3-70B)	0.06	94.1%	0.05	59.5%	0.07	92.9%
CdRO (GPT-4o)	0.06	94.0%	0.04	59.5%	0.07	92.7%
CdRO* (GPT-4o)	0.05	94.1%	0.05	59.4%	0.07	92.7%
ALBERT	0.29	92.6%	0.19	58.5%	0.20	91.3%
Auto-Ddebias	0.39	86.8%	0.18	56.9%	0.36	91.1%
Causal-Debias	0.13	92.9%	0.16	57.1%	0.01	91.6%
CdRO (LLaMA-3-70B)	0.07	93.8%	0.09	59.8%	0.01	92.5%
CdRO* (LLaMA-3-70B)	0.07	93.8%	0.10	59.8%	0.02	92.4%
CdRO (GPT-4o)	0.06	94.1%	0.10	59.7%	0.01	92.4%
CdRO* (GPT-4o)	0.06	93.8%	0.10	59.6%	0.02	92.5%

- **Dropout** [40] is a regularization technique aimed at mitigating overfitting in models. The primary concept behind this method is to randomly "drop" (set to zero) the outputs of neurons during training with a specified probability.
- **Context-Debias** [19] is a fine-tuning technique designed to mitigate bias in pretrained contextualized embeddings. This method can be applied at either the token or sentence level.
- **Auto-Debias** [13] is an automated approach for mitigating biases in LLMs. Unlike previous debiasing methods that rely on external corpora for fine-tuning, this technique directly probes the biases encoded within LLMs using prompts.
- **MABEL** [14] is an intermediate pretraining method designed to mitigate gender bias in contextualized representations. Central to this approach is the application of a contrastive learning objective on counterfactually augmented, gender-balanced entailment pairs derived from natural language inference (NLI) datasets.
- **Sent-Debias** [24] is an extension of Hard-Debias [1], designed to mitigate bias in sentences with respect to both binary and multiclass bias attributes, including gender and religion.
- **FairFil** [6] is the first neural debiasing method designed for pretrained sentence encoders. It transforms the outputs of these encoders into debiased representations through the use of a fair filter network. To train the FairFil model, a contrastive learning framework is utilized that minimizes the correlation between the filtered embeddings and bias-related words, while simultaneously preserving the rich semantic information of the original sentences.
- **Causal-Debias** [46] is a causal invariant debiasing model that integrates debiasing with downstream fine-tuning. This approach fundamentally analyzes the causes and propagation of biases, introducing a structural causal model [32] to address bias mitigation by leveraging the inherent causal mechanisms present in downstream datasets.
- **PCFR** [15] is an innovative disentanglement method that combines prompt learning and contrastive learning to mitigate bias in LLMs. Prompt learning is utilized to represent sensitive information as distinct embeddings, and then contrastive learning is applied to compare these information embeddings, rather than the traditional sentence embeddings.

3.3 More Experimental Results

The results of gender and race debiasing using the SEAT evaluation have been presented in the main text. This section presents more gender debiasing results using CrowS-Pair across various methods. As shown in Table 2, the proposed CdRO approach consistently outperforms existing debiasing

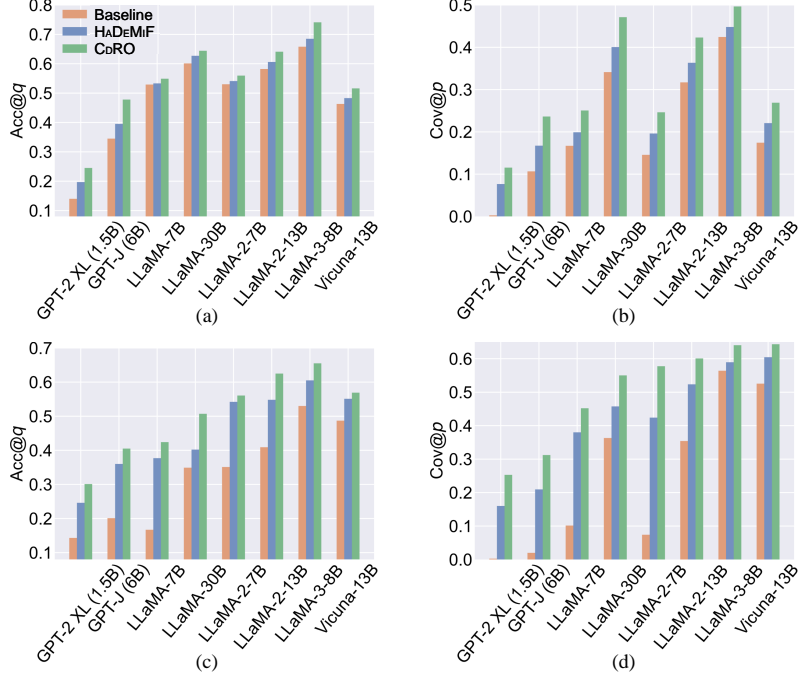


Figure 6: Bar charts illustrating the averaged $\text{Acc}@q$ and $\text{Cov}@p$ scores for the original LLMs, HADEMiF, and CDRO. (a) and (b) summarize the average performance across the NQ, SciQ, and TriviaQA datasets, whereas (c) and (d) present the average performance on the TruthfulQA and WikiQA datasets.

methods in terms of both mitigating gender bias and enhancing downstream task performance. These results underscore the effectiveness of our method in reducing stereotypical associations and enhancing the resilience of the model. Furthermore, the results presented in Table 3 demonstrate that utilizing both GPT-4o and LLaMA-3-70B models during the data collection process yields strong and comparable performance. Nevertheless, to reduce computational and financial overhead, we primarily utilize the LLaMA-3-70B model for both data generation and evaluation in our experiments.

4 Additional Details on Hallucination Mitigation Experiments

4.1 More Details about Experimental Settings

We consider five popular NLG tasks, including NQ², SciQ³, TriviaQA⁴, TruthfulQA⁵, and WikiQA⁶. Following previous research [26, 47], for the first three tasks, 1K samples are utilized for testing and 2K samples for training. For TruthfulQA, which lacks an official training set, 397 instances are randomly sampled from the original test set for training, and the remaining instances are utilized for testing. For the WikiQA dataset, the training set consists of 1,040 instances, while the test set contains 293 instances.

Given that hallucinations refer to incorrect model generations made with unwarranted confidence, we employ two key indicators to evaluate the effectiveness of our approach in mitigating knowledge hallucinations. These indicators include accuracy at coverage ($\text{Acc}@q$) and coverage at accuracy ($\text{Cov}@p$). The first metric, $\text{Acc}@q$, measures model precision by assessing the accuracy of the top- q percent of predictions. The second metric, $\text{Cov}@p$, quantifies recall by determining the maximum proportion of highly confident predictions that exceed a predefined accuracy threshold p . Unlike

²<https://github.com/google-research-datasets/natural-questions>

³<https://huggingface.co/datasets/allenai/sciq>

⁴<https://nlp.cs.washington.edu/triviaqa/>

⁵<https://github.com/sylinrl/TruthfulQA>

⁶https://huggingface.co/datasets/microsoft/wiki_qa

Table 4: Performance comparison between our approach and vanilla fine-tuning on three models: BERT-base, RoBERTa-base, and BART-base. All models were trained on their respective in-distribution datasets and evaluated on the development set comprising OOD examples.

Dataset		SST-2		MNLI		QQP
Model	Method	IMDB-Cont	IMDB-CAD	HANS	AdvNLI	PAWS
BERT-base	Fine-tuning	79.08%	87.00%	56.90%	24.12%	32.80%
	CDRO (Linear)	85.12%	90.98%	66.85%	33.61%	39.18%
RoBERTa-base	Fine-tuning	84.51%	88.39%	67.80%	31.22%	38.45%
	CDRO (Linear)	89.62%	92.65%	77.68%	39.40%	46.01%
BART-base	Fine-tuning	82.48%	86.03%	56.30%	30.51%	32.27%
	CDRO (Linear)	86.24%	91.11%	62.97%	37.21%	39.50%

AUROC [2], which primarily evaluates the quality of confidence scores, these metrics offer a more direct assessment of the model’s ability to filter out incorrect predictions through explicit thresholding. Regarding the hyperparameter configuration, the rank of PiSSA is set to 32 across all datasets, and the LLMs are fine-tuned for 15 epochs with a learning rate of 1×10^{-5} . The learning rate for the logistic regression model is set to 1×10^{-2} . Five types of matrices—namely the query, key, value, up-projection, and down-projection matrices—are designated as tunable parameters. The compared baseline methods are implemented according to the settings outlined in [47] or their respective original papers. The remaining hyperparameters for our proposed approach are consistent with those specified in Section 3.

4.2 More Details about Compared Baselines

We compare the proposed CDRO approach with a range of traditional and advanced methods designed to enhance the reliability and robustness of models, including techniques for model calibration and hallucination detection and mitigation, which are detailed as follows:

- **Temperature Scaling** [25] is a post-hoc calibration method that refines model confidence by introducing a temperature parameter to scale the logits before applying the Softmax function. This adjustment effectively smooths the predicted probability distribution, improving the alignment between predicted confidence scores and true probabilities.
- **Label Smoothing** [36] is incorporated into the fine-tuning process of LLMs using LoRA, where the training labels are adjusted to distribute a small portion of the probability mass across all classes. This technique helps prevent overconfidence in model predictions and enhances generalization performance.
- **LITCAB** [26] utilizes a single linear layer to transform the input text representation and estimate a bias term, which is subsequently incorporated into the output logits of the LLMs.
- **Calibration-Tuning** [20] enhances LLMs by fine-tuning them on a task designed to enable the model to independently assess the consistency of its generations with the ground truth.
- **Verbalization** involves prompting the LLMs to self-assess and report their confidence level for a given output, using the prompt methodology proposed by [37].
- **P(IK)** [17] introduces a linear layer applied to the final hidden state of the LLMs, corresponding to the last token of a given question. This layer is trained to predict the likelihood that the model will provide an accurate answer to the question.
- **Self-Consistency** [37, 43] is based on the principle that responses with higher confidence are more likely to be consistent when sampled repeatedly from the model.
- **R-Tuning** [44] fine-tunes LLMs using refusal-aware datasets, enabling the models to generate responses that account for refusals, thereby reducing the occurrence of hallucinations.
- **HADEMiF** [47] leverages two compact detection networks to identify hallucinations occurring in both the internal representations and output space of LLMs, and refines the final predictions by incorporating the signals produced by these detection modules.

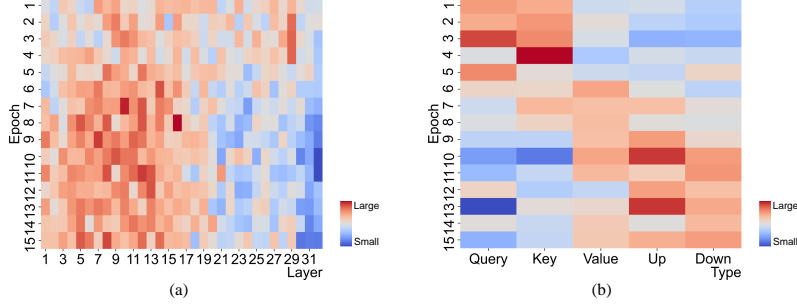


Figure 7: The update patterns of matrices across different layers (a) and across different matrix types (i.e., query, key, value, up, and down) (b) during the training process.

- **ITI** [23], a technique designed to enhance the truthfulness of LLMs, which operates by shifting model activations during inference, following a set of directions across a limited number of attention heads.
- **DoLa** [7] obtains the next-token distribution by contrasting the differences in logits obtained from projecting the later layers versus earlier layers to the vocabulary space, exploiting the fact that factual knowledge in an LLM has generally been shown to be localized to particular transformer layers.
- **SH2** [18] is proposed to help LLMs decode more truthfully, which introduces hesitations to give LLMs more time to understand contexts and answer questions. For LLMs, the tokens assigned with lower probabilities are harder to predict, while more likely to be informative.

4.3 More Experimental Results

Besides LLaMA-2-7B⁷, we conduct experiments on GPT-2 XL (1.5B)⁸, GPT-J (6B)⁹, Vicuna-13B¹⁰, LLaMA-7B¹¹, LLaMA-30B¹², LLaMA-2-13B¹³, and LLaMA-3-8B¹⁴. Fig. 6 presents a comparative analysis of the proposed CdRO approach, HADEMiF, and the original LLMs, evaluated across two performance metrics: $\text{Acc}@q$ and $\text{Cov}@p$. The proposed method consistently outperforms the baseline approaches across various LLMs, achieving the highest values for $\text{Acc}@q$ and $\text{Cov}@p$. These findings underscore the significant effectiveness of CdRO in mitigating knowledge hallucinations, particularly by reducing the generation of content that is produced with high confidence but lacks factual accuracy.

5 Additional Details on Out-of-Distribution Experiments

5.1 More Details about Experimental Settings

To assess the generalization capability of our method under OOD scenarios, we conduct experiments on three tasks from the GLUE benchmark [38]: QQP [39] for paraphrase identification, MNLI [41] for NLI, and SST-2 [35] for sentiment analysis. These tasks are chosen due to the availability of well-established OOD test sets, which facilitate a robust evaluation of model performance under distributional shift.

- SST-2 has been chosen as the in-distribution dataset for the sentiment classification task. For OOD counterparts, we have selected two challenging sentiment classification datasets: the

⁷<https://huggingface.co/meta-llama/Llama-2-7b>

⁸<https://huggingface.co/openai-community/gpt2-xl>

⁹<https://huggingface.co/EleutherAI/gpt-j-6b>

¹⁰<https://lmsys.org/blog/2023-03-30-vicuna/>

¹¹<https://huggingface.co/huggyllama/llama-7b>

¹²<https://huggingface.co/huggyllama/llama-30b>

¹³<https://huggingface.co/meta-llama/Llama-2-13b-hf>

¹⁴<https://huggingface.co/meta-llama/Meta-Llama-3-8B>

IMDB Contrast Set (IMDB-Cont) [9] and the IMDB Counterfactually Augmented Dataset (IMDB-CAD)¹⁵ [21].

- MNLI [41] is a widely used dataset for the NLI task, consisting of approximately 400K premise/hypothesis pairs annotated with textual entailment information, including neutral, entailment, and contradiction labels. MNLI includes two distinct development sets: the matched set (MNLI-m), which is derived from the same sources as the training set, and the mismatched set (MNLI-mm), which contains examples that do not closely resemble those seen during training [41]. Pretrained models were trained on MNLI as an in-distribution dataset and subsequently evaluated on both MNLI development sets (m/mm), as well as two relevant OOD datasets: HANS¹⁶ and Adversarial NLI (AdvNLI)¹⁷.
- QQP [33] is a widely used dataset for the paraphrase identification task, consisting of approximately 400K sentence pairs labeled as either paraphrases or non-paraphrases. The corresponding OOD dataset is PAWS-QQP¹⁸ [45], which features high lexical overlap but different semantic meanings.

The experimental configurations for the baseline methods are aligned with the settings reported in their respective original publications. Three pretrained language models are utilized in these experiments, including BERT-base [8], RoBERTa-base [27], and BART-base [22]. Notably, BART adopts an autoregressive architecture, which differs fundamentally from the encoder-based designs of BERT and RoBERTa. Moreover, the experimental configurations for our proposed CDRO method are consistent with the specifications detailed in Section 3.

5.2 More Details about Compared Baselines

Several methods aimed at enhancing model generalization and robustness, including Span Cutoff [34], HiddenCut [5], IPT-Adapter [10], Causal-Debias [46], and PCFR [15], are included in the comparison. These methods are introduced in detail as follows:

- In **Span Cutoff** [34], a portion of the information within an input sentence is removed to create its restricted views during the fine-tuning stage. Notably, this approach relies solely on stochastic sampling, resulting in minimal computational overhead.
- **HiddenCut** [5] was proposed to enhance model regularization and promote the learning of more generalizable features. Specifically, contiguous spans within the hidden space are dynamically and strategically dropped during training.
- **IPT-Adapter** [10] employs adapter-based fine-tuning to address potential issues by freezing the original transformer parameters and introducing new adapter parameters within the transformer layers.
- **Causal-Debias** [46] is a causal invariant debiasing model that integrates debiasing with downstream fine-tuning. This approach fundamentally analyzes the causes and propagation of biases, introducing a structural causal model [32] to address bias mitigation by leveraging the inherent causal mechanisms present in downstream datasets.
- **PCFR** [15] is a novel disentanglement method that integrates prompt learning with contrastive learning to reduce bias in LLMs. It uses prompt learning to represent sensitive information as separate embeddings, followed by contrastive learning to compare these embeddings instead of traditional sentence embeddings.

5.3 More Experimental Results

This section presents additional comparison results on more LLMs. As shown in Table 4, our proposed CDRO method consistently outperforms vanilla fine-tuning across all three models—BERT-base, RoBERTa-base, and BART-base—in OOD scenarios. Specifically, our approach achieves average improvements of 7.17%, 7.00%, and 5.89% on the BERT, RoBERTa, and BART models, respectively.

¹⁵<https://github.com/acmi-lab/counterfactually-augmented-data>

¹⁶<https://github.com/tommccoy1/hans>

¹⁷<https://github.com/facebookresearch/anli>

¹⁸<https://github.com/google-research-datasets/paws>

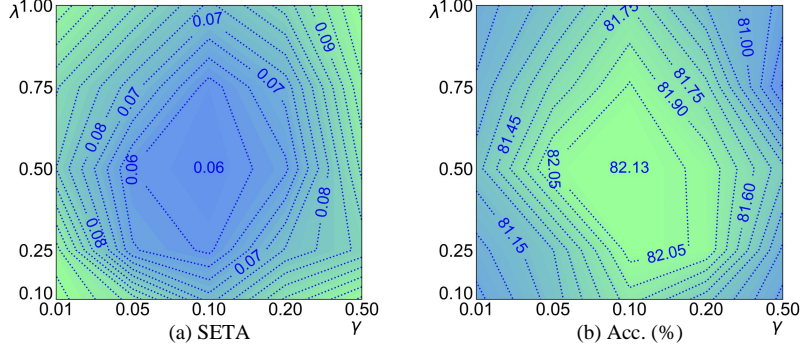


Figure 8: Sensitivity test results for the λ and γ values in NLU tasks, evaluated across SEAT for gender bias and accuracy metrics.

These results underscore the effectiveness of our method in enhancing model resilience by mitigating spurious correlations present in the training data.

6 Training Analysis

In addition, we examine the update behavior of parameter matrices throughout the training process. To facilitate analysis, the matrices are categorized according to their layer positions and functional types. A total of twenty matrices with the highest predicted probabilities are selected to be updated. Fig. 7(a) illustrates the update patterns across different layers. A notable observation is that, after a period of training, the proportion of updates in the deeper layers decreases. This indicates that, as the model converges, the matrices most sensitive to causal relationships tend to concentrate in the earlier and intermediate layers. Fig. 7(b) further illustrates the update patterns of different matrix types. It is evident that the query and key matrices are predominantly updated during the early training stages, while the value, up, and down matrices receive more updates in the later phases.

7 More Sensitivity Analysis

We then conduct sensitivity analyses on two hyperparameters specific to our approach, namely γ and λ . The hyperparameter γ controls the strength of the reward ranking term in the advantage computation, while λ governs the relative weighting between the accuracy reward and the other three reward components. As illustrated in Fig. 8, the model demonstrates stable performance when γ is within the range of $[0.05, 0.2]$. Additionally, the model maintains consistent performance for $\lambda \in [0.25, 0.75]$, further emphasizing the robustness of our approach to variations in these hyperparameters. For practical applications, the values of these two parameters can be conveniently set to 0.1 and 0.5, respectively, to achieve satisfactory performance.

8 Limitations and Future Work

The proposed CDRO framework exhibits strong effectiveness in breaking spurious correlations and mitigating hallucinations within LLMs. Nonetheless, several limitations remain, highlighting promising directions for future research. A primary limitation lies in the reliance on fine-tuning LLMs' parameters, which restricts the applicability of the method to black-box settings. Future work could explore extending our framework to improve the prediction robustness of black-box models, potentially through techniques such as prompt engineering or adapter-based fine-tuning. Additionally, applying our framework to areas beyond enhancing causality in LLMs, such as improving security, privacy, and fairness, presents an intriguing avenue for future research. For instance, by developing alternative knowledge localization strategies, it becomes feasible to identify and selectively optimize parameters associated with specific knowledge domains. This approach enables parameter-efficient and targeted enhancements in particular aspects of model behavior.

9 Ethical Considerations

All models and datasets employed in this study have been meticulously processed and curated by their respective developers to address and mitigate potential ethical concerns. Moreover, safety and privacy are carefully maintained throughout our data collection process.

References

- [1] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 4356–4364, 2016.
- [2] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.
- [3] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.
- [4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
- [5] Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. Hiddencut: Simple data augmentation for natural language understanding with better generalization. *arXiv preprint arXiv:2106.00149*, 2021.
- [6] Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *Proceedings of the International Conference on Learning Representations*, 2021.
- [7] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.
- [9] Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models’ local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1307–1323, 2020.
- [10] Goran Glavaš and Ivan Vulić. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, 2021.
- [11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 1321–1330, 2017.
- [13] Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, 2022.

- [14] Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. Mabel: Attenuating gender bias using textual entailment data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9681–9702, 2022.
- [15] Junheng He, Nankai Lin, Qifeng Bai, Haoyu Liang, Dong Zhou, and Aimin Yang. Towards fair decision: A novel representation method for debiasing pre-trained models. *Decision Support Systems*, 181:114208, 2024.
- [16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [17] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- [18] Jushi Kai, Tianhang Zhang, Hai Hu, and Zhouhan Lin. Sh2: Self-highlighted hesitation helps you decode more truthfully. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4514–4530, 2024.
- [19] Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, 2021.
- [20] Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. Calibration-tuning: Teaching large language models to know what they don’t know. In *Proceedings of the Workshop on Uncertainty-Aware NLP*, pages 1–14, 2024.
- [21] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the International Conference on Learning Representations*, 2020.
- [22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.
- [23] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 41451–41530, 2023.
- [24] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, 2020.
- [25] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the Sixth International Conference on Learning Representations*, 2018.
- [26] Xin Liu, Muhammad Khalifa, and Lu Wang. Litcab: Lightweight language model calibration over short-and long-form responses. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.
- [27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [28] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, 2019.

- [29] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, 2019.
- [30] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 121038–121072, 2024.
- [31] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1953–1967, 2020.
- [32] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.
- [33] Quora. First quora dataset release: Question pairs, 2017.
- [34] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*, 2020.
- [35] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.
- [36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
- [37] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, 2023.
- [38] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.
- [39] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*, 2017.
- [40] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.
- [41] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.
- [42] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.
- [43] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

- [44] Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say ‘i don’t know’. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7106–7132, 2024.
- [45] Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, 2019.
- [46] Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, 2023.
- [47] Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. Hademif: Hallucination detection and mitigation in large language models. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.