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
wye@pku.edu.cn

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

Shikun Zhang
Peking University
zhangsk@pku.edu.cn

Abstract

Large language models (LLMs) have achieved remarkable achievements across diverse applications; however, they remain plagued by spurious correlations and the generation of hallucinated content. Despite extensive efforts to enhance the resilience of LLMs, existing approaches either rely on indiscriminate fine-tuning of all parameters, resulting in parameter inefficiency and lack of specificity, or depend on post-processing techniques that offer limited adaptability and flexibility. This study introduces a novel **C**ausality-**d**riven **R**obust **O**ptimization (CDRO) approach that selectively updates model components sensitive to causal reasoning, enhancing model causality while preserving valuable pretrained knowledge to mitigate overfitting. Our method begins by identifying the parameter components within LLMs that capture causal relationships, achieved through comparing the training dynamics of parameter matrices associated with the original samples, as well as augmented counterfactual and paraphrased variants. These comparisons are then fed into a lightweight logistic regression model, optimized in real time to dynamically identify and adapt the causal components within LLMs. The identified parameters are subsequently optimized using an enhanced policy optimization algorithm, where the reward function is designed to jointly promote both model generalization and robustness. Extensive experiments across various tasks using twelve different LLMs demonstrate the superior performance of our framework, underscoring its significant effectiveness in reducing the model's dependence on spurious associations and mitigating hallucinations.

*Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

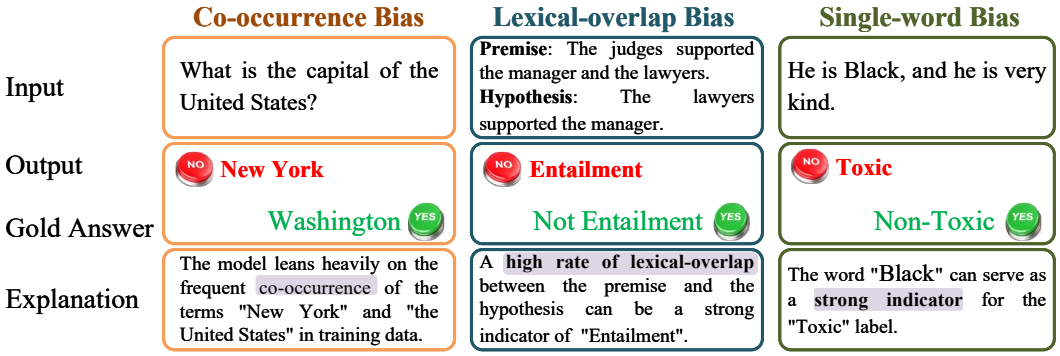


Figure 1: Examples of prediction errors caused by spurious associations due to various biases.

1 Introduction

Large language models (LLMs) have demonstrated remarkable and unprecedented capabilities across a wide range of applications [71, 1, 108, 101]. However, they continue to face substantial challenges concerning the prediction robustness and reliability[115, 50, 111]. Specifically, models are prone to relying on spurious correlations for prediction, as they tend to overfit superficial statistical patterns in the training data rather than learning the underlying causal relationships. This over-reliance not only undermines model generalization but also constitutes a key factor contributing to knowledge hallucination—where models generate factually incorrect outputs with unwarranted confidence [97, 77, 15, 6]. Notably, spurious associations are especially pervasive and often stem from various biases in the data, such as co-occurrence bias, lexical overlap bias, and single-word bias [113, 77, 12, 110], as illustrated in Fig. 1. As LLMs are increasingly deployed in high-stakes domains like healthcare, law, and journalism, addressing these challenges to enhance their resilience and trustworthiness has become an urgent priority [107, 112, 103, 39].

Numerous studies have attempted to enhance the robustness and reliability of LLMs by mitigating their dependence on spurious correlations and reducing hallucinations [82, 87, 6, 52]. Among these, causal learning methods have emerged as a promising direction for disentangling spurious correlations from true causal relationships. For example, Causal-Debias [106] generates counterfactual sentences with non-causal variations but identical semantic meanings. These counterfactual sentences, alongside the original ones, are fed into an invariant optimization function to balance model performance on downstream tasks and debiasing effectiveness. Moreover, Causal Effect Tuning [105] leverages causal inference to identify and preserve valuable pretrained knowledge during fine-tuning, while simultaneously uncovering missing causal effects in the pretrained data that contribute to knowledge forgetting. In parallel, a growing body of research has focused on mitigating hallucinations, with methods ranging from data-related techniques to modeling and inference strategies [66, 111, 9, 80, 31]. For instance, LITCAB [49] is a lightweight calibration mechanism that employs a single linear layer to process input text representations and predict a bias term, which is subsequently utilized to adjust the logits. Furthermore, the self-reflective approach [31] generates relevant background knowledge for a given query, followed by a factual consistency check; if inconsistencies are detected, the model leverages its internal reflective capability to revise its response accordingly. While effective, existing methods either involve indiscriminate fine-tuning of all model parameters [106, 105, 10], leading to parameter inefficiency and a lack of specificity. This can cause the model to forget valuable pretrained knowledge and become susceptible to overfitting, or they rely on post-processing techniques that offer limited adaptability and flexibility, thus hindering fundamental progress in the model's causal reasoning and understanding capabilities [49, 31, 61, 24].

In response, this study proposes a novel **C**ausality-**d**riven **R**obust **O**ptimization (CDRO) framework, aimed at enhancing the causal reasoning abilities of LLMs by accurately identifying and selectively optimizing the parameters that capture causal relationships. Initially, we leverage the instruction-following and textual understanding capabilities of state-of-the-art (SOTA) LLMs to automate the generation of counterfactual and paraphrased variants of the training data. The parameters encoding causal relationships are then identified by analyzing their training dynamics across different sample

types. Specifically, we compare loss gradient and activation patterns of parameter matrices and feed the comparisons into a logistic regression model to automatically identify and predict the components sensitive to causal relationships. In contrast to previous knowledge localization strategies, which focus on causal influence at the layer level with predefined matrix types (e.g., those in feed-forward networks), our approach performs localization at the matrix level, offering greater precision and flexibility [58, 55]. Subsequently, we optimize the localized causal components within LLMs using an enhanced REINFORCE++ algorithm, where the reward signals are designed to simultaneously promote model generalization and robustness; meanwhile, the logistic regression model is updated in real time based on the performance of the LLMs during the optimization process, facilitating the adaptive and dynamic localization of causal components.

Extensive experiments have been conducted on both natural language understanding (NLU) and natural language generation (NLG) tasks, leveraging twelve different LLMs with varying parameter sizes. The results demonstrate that CDRO consistently outperforms existing approaches in reducing stereotypical associations and mitigating hallucinations. Furthermore, it demonstrates superior performance in out-of-distribution (OOD) settings, highlighting its efficacy in reducing the model's reliance on spurious correlations within the training data.

In summary, the primary contributions of our work are as follows:

- We introduce a novel approach for localizing causal knowledge in LLMs by comparing the training dynamics of model parameters across varying instance types and utilizing a logistic regression model to autonomously capture the relationship between these comparisons and the predictions of causal components.
- We propose a collaborative optimization framework wherein the causal components within LLMs are optimized using an enhanced REINFORCE++ algorithm, while the logistic regression model for knowledge localization is simultaneously updated in real-time, driven by the performance of the evolving LLMs.
- We conduct comprehensive experiments on both NLU and NLG tasks to assess the effectiveness of our approach in model debiasing, hallucination mitigation, and OOD prediction. The results consistently demonstrate the superiority of our method across all evaluated scenarios.

2 Related Work

Causality for LLMs. Despite their remarkable success, LLMs often rely on statistical correlations rather than true causal relationships, making them susceptible to demographic biases, social stereotypes, and hallucinations [97, 18, 20]. To address this, various methods have been proposed across different stages. Pretraining methods include debiased embeddings [91, 106], counterfactual corpora [116, 37], and causal foundation models [88, 70]. Fine-tuning approaches such as Causal-Debias [106] and Causal Effect Tuning [106] aim to inject causal awareness into model parameters [105]. Alignment techniques reduce harmful outputs by aligning models with human values [62, 48, 4], while inference-time methods utilize causal prompts to elicit more grounded responses [2, 83, 64]. However, most existing methods either optimize all model parameters uniformly, which results in parameter inefficiencies and an increased risk of overfitting, or rely only on inference, thereby offering limited performance improvements [97, 20]. In contrast, our method first localizes causal knowledge and then applies targeted reinforcement-based fine-tuning, striking a better balance between preserving pretrained capabilities and enhancing downstream task performance.

Knowledge Localization. Prior studies have proposed various methods to localize knowledge within LLMs, aiming to identify components responsible for encoding factual or causal information. Parameter-based approaches such as Knowledge Neurons [16, 109] and DEPN [98] trace model updates to locate key parameters for factual recall. Activation-based methods investigate saliency in hidden states and attention heads via gradients or concept erasure [19, 16, 59, 26, 57]. Moreover, causal probing techniques [84, 5] reveal causal relationships within the model via counterfactual or mediation analysis. We extend causal probing by comparing model behaviors across diverse sample types, including original, counterfactual, and paraphrased instances, and utilizing these comparisons to train a logistic regression model for automated and adaptive knowledge localization in LLMs.

Policy Optimization. To align LLMs with human intent, reinforcement learning methods such as RLHF [76, 65] and PPO [72] are commonly employed. However, these methods typically involve significant computational overhead due to the training of reward models. To address this, more efficient alternatives have emerged. DPO [68] bypasses reward modeling by directly optimizing preferences using a cross-entropy (CE) loss, while GRPO [73] reduces reliance on external evaluators through group-based assessments. Additionally, REINFORCE++[29] enhances both stability and effectiveness by incorporating PPO techniques into the traditional REINFORCE framework [95], leading to improved performance. In this study, we propose an enhanced version of REINFORCE++, which incorporates reward ranking information to refine advantage estimation and optimize LLMs' behavior more effectively.

3 Methodology

To enhance the causal reasoning abilities of LLMs in a parameter-efficient and targeted manner, we propose CDRO, with its overall framework depicted in Fig. 2. This method leverages reinforcement learning-based optimization to selectively update the model components that are most pertinent to modeling causal relationships. Specifically, we first prompt SOTA LLMs to generate counterfactual and paraphrased variants of the training data. By analyzing the training dynamics of parameter matrices across different types of samples, we identify components that exhibit high sensitivity to causal reasoning. These identified components are then optimized using an enhanced REINFORCE++ algorithm, wherein rewards are assigned based on the model's performance on both the original and the augmented counterfactual and paraphrased samples.

3.1 Counterfactual and Paraphrastic Data Collection

Counterfactual and paraphrased variants of the training data are first generated to facilitate the localization of causal knowledge within LLMs. All steps in this process are performed by prompting off-the-shelf LLMs without requiring manual annotation.

To ensure high-quality generation, we utilize SOTA LLMs, such as LLaMA-3-70B [23] and GPT-4o [30], for data collection. Counterfactual samples are generated by minimally modifying original instances to change their labels (in NLU tasks) or answers (in NLG tasks), while preserving thematic consistency [44, 93]. Similar to counterfactual generation, we prompt SOTA LLMs to generate paraphrased samples from the original data, preserving the original semantics to maintain consistent labels or answers [96]. The inclusion of relevant details is permitted to enrich the paraphrased content. We further prompt the LLMs[2] to assess the quality of their generations. Evaluations cover the following dimensions: alignment or divergence between the answers of augmented and original samples, answer correctness, thematic consistency, clarity, and safety and privacy. Each instance is rated three times on a scale from 0 to 10, and the outputs with the highest average scores across eight generations are selected for downstream use. The specific prompts used for both generation and evaluation are provided in the Appendix. After the data collection process, each original sample $\boldsymbol{x}_i$ is paired with a corresponding counterfactual sample $\boldsymbol{x}_i'$, as well as a paraphrased sample $\boldsymbol{x}_i''$.

3.2 Localization of Causality-Sensitive Parameters

Our approach aims to localize the components within LLMs that are sensitive to causal relationships, optimizing only these identified components to enhance the model's causal reasoning capabilities in a targeted manner. This optimization strategy can not only effectively preserve the knowledge gained during pretraining, thereby mitigating the risk of catastrophic forgetting, but also enhance the model's resilience on downstream tasks. To facilitate effective localization, we analyze the learning dynamics of various weight matrices across the original, counterfactual, and paraphrased augmented samples. Specifically, we utilize two indicators of training dynamics—loss gradients and activation maps—to evaluate how causal relationships are encoded within model parameters. The loss gradients capture the model's dependence on and sensitivity to specific matrices during training, while the activation values reveal the model's responses and the information flow across different layers. The

[2]These SOTA LLMs have exhibited strong self-evaluation capabilities [7], and the use of alternative models for evaluation is also assessed, as presented in Appendix 3.

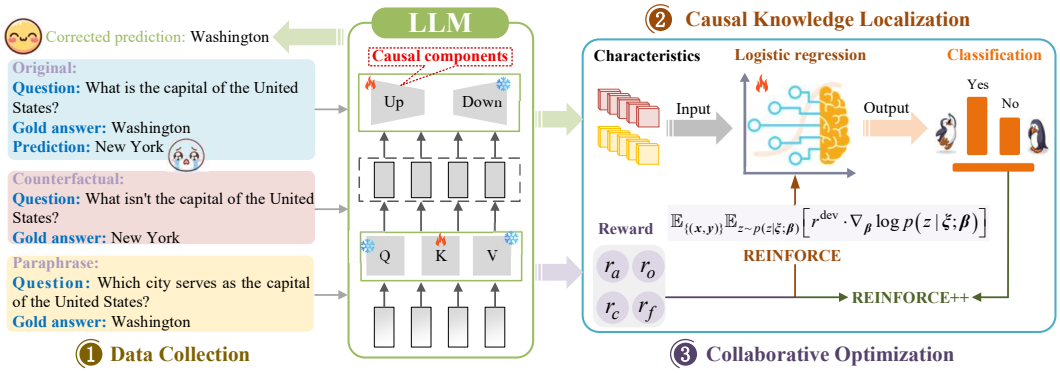


Figure 2: Overview of the proposed CDRO framework. Our approach first prompts SOTA LLMs to generate counterfactual and paraphrased variants of training data, then compares the characteristics of weight matrices of different categories (e.g., query, key, value, up, and down) and layers across different sample types. These comparisons are subsequently fed into a logistic regression model to predict the probability of causal expression. Finally, an enhanced REINFORCE++ algorithm is employed to optimize the identified causal components, while the logistic regression model is concurrently updated in real time using the REINFORCE algorithm.

comparisons in these two indicators are computed between original and counterfactual samples, as well as between paraphrased and counterfactual samples.

For the j-th weight matrix $\boldsymbol{\theta}_j$, we begin by computing the difference in loss gradients between the original and counterfactual samples, defined as $\mathcal{G}_j^1 = \left|\nabla_{\boldsymbol{\theta}_j}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}(\boldsymbol{x}_i)\right) - \nabla_{\boldsymbol{\theta}_j}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}(\boldsymbol{x}'_i)\right)\right|_2$, and the difference between the counterfactual and paraphrased samples as $\mathcal{G}_j^2 = \left|\nabla_{\boldsymbol{\theta}_j}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}(\boldsymbol{x}'_i)\right) - \nabla_{\boldsymbol{\theta}_j}\left(\frac{1}{n}\sum_{i=1}^n \mathcal{L}(\boldsymbol{x}''_i)\right)\right|_2$, where $\mathcal{L}(\cdot)$ denotes the CE loss and n represents the mini-batch size. The second indicator we consider is the activation map, specifically the hidden states of each layer. To facilitate comparisons across different sample types, we compute the cosine similarity of the hidden states between the original and counterfactual samples, as well as between the counterfactual and paraphrased samples: $\mathcal{S}_{i,l_j}^1 = \frac{\boldsymbol{h}_{i,l_j}\cdot\boldsymbol{h}'_{i,l_j}}{|\boldsymbol{h}_{i,l_j}||\boldsymbol{h}'_{i,l_j}|}$ and $\mathcal{S}_{i,l_j}^2 = \frac{\boldsymbol{h}'_{i,l_j}\cdot\boldsymbol{h}''_{i,l_j}}{|\boldsymbol{h}'_{i,l_j}||\boldsymbol{h}''_{i,l_j}|}$, where l_j denotes the layer index of matrix $\boldsymbol{\theta}_j$, and $\boldsymbol{h}_{i,l_j}$, $\boldsymbol{h}'_{i,l_j}$, and $\boldsymbol{h}''_{i,l_j}$ represent the hidden states from the l_j-th layer for the i-th original, counterfactual, and paraphrased samples, respectively. The hidden states of the final token are utilized, as they capture global sentence-level information.

During the optimization process, the two gradient differences, $\mathcal{G}_j^1$ and $\mathcal{G}_j^2$, along with the mean and variance of the two cosine similarities (i.e., $\mathcal{S}_{i,l_j}^1$ and $\mathcal{S}_{i,l_j}^2$) across a batch of samples, are input into a logistic regression model [38]. Specifically, each matrix is associated with a six-dimensional feature vector $\boldsymbol{\xi}_j = \left[\mathcal{G}_j^1, \mathcal{G}_j^2, \bar{\mathcal{S}}_{l_j}^1, \bar{\mathcal{S}}_{l_j}^2, \hat{\mathcal{S}}_{l_j}^1, \hat{\mathcal{S}}_{l_j}^2\right]$, where $\bar{\mathcal{S}}_{l_j}^1$ and $\hat{\mathcal{S}}_{l_j}^1$ represent the mean and variance, respectively, of the values $\mathcal{S}_{i,l_j}^1$ computed over a batch of training data. The symbols for $\mathcal{S}_{i,l_j}^2$ are defined analogously. The logistic regression model subsequently learns the relationship between the input indicators and the predicted probability that a given matrix governs causal reasoning relevant to the downstream task, as formalized in the following:

$$p(z_j \mid \boldsymbol{\xi}_j; \boldsymbol{\beta}) = \frac{1}{1 + \exp\left(-\boldsymbol{\beta}^\top \boldsymbol{\xi}_j\right)}, \tag{1}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_6]$ denotes the parameters of the logistic regression model. The predicted probability $p(z_j \mid \boldsymbol{\xi}_j; \boldsymbol{\beta})$ indicates the likelihood that matrix $\boldsymbol{\theta}_j$ encodes the causal relationship within the model, given its corresponding characteristics $\boldsymbol{\xi}_j$. The use of the logistic regression model provides a simple, efficient, and highly interpretable framework for identifying causal components.

3.3 REINFORCE-Based Collaborative Optimization

The parameter components sensitive to causal reasoning within LLMs and the logistic regression model in our framework are updated in an alternating fashion. Specifically, the LLMs are optimized using an enhanced REINFORCE++ algorithm. Since direct gradient backpropagation from the LLMs to the logistic regression model is not feasible, we employ the standard REINFORCE algorithm [95] to optimize it, taking advantage of its lightweight structure. In this approach, the reward signal is derived from the performance of the LLMs. This collaborative optimization process ensures that the knowledge localization process remains tightly aligned with the evolving learning states of the LLMs.

We define the policy network as the target LLM parameterized by $\boldsymbol{\theta}$, where $\boldsymbol{\theta}^c \subseteq \boldsymbol{\theta}$ denotes the subset of causality-sensitive parameters. To enhance optimization efficiency, we employ the low-rank adaptation method PiSSA [56], which constrains fine-tuning to the principal subspace of the identified causal matrices. In this case, the gradient computation for the weight matrices is also restricted to the top r principal components, thereby improving memory efficiency. During each optimization step, REINFORCE++ samples an output for each input $\boldsymbol{x}$ from the previous policy $\pi_{\boldsymbol{\theta}_{\text{old}}}$. Accordingly, the optimization objective can be defined as follows:

$$\mathcal{J}(\boldsymbol{\theta}^c) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}\mathbb{E}_{\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}^c_{\text{old}}\subseteq\boldsymbol{\theta}_{\text{old}}}(\mathcal{Y}|\boldsymbol{x})}\left[\frac{1}{|\boldsymbol{y}|}\sum_{t=1}^{|\boldsymbol{y}|}\min\left(\rho_t\left(\boldsymbol{\theta}^c\right)\mathcal{A}_t, \text{clip}\left(\rho_t\left(\boldsymbol{\theta}^c\right), 1-\epsilon, 1+\epsilon\right)\mathcal{A}_t\right)\right], \tag{2}$$

where $\rho_t(\boldsymbol{\theta}^c) = \frac{\pi_{\boldsymbol{\theta}^c\subseteq\boldsymbol{\theta}}\left(y_t|\boldsymbol{x},\boldsymbol{y}_{<t}\right)}{\pi_{\boldsymbol{\theta}^c_{\text{old}}\subseteq\boldsymbol{\theta}_{\text{old}}}\left(y_t|\boldsymbol{x},\boldsymbol{y}_{<t}\right)}$ represents the probability ratio between new and old policies, and the hyperparameter ϵ serves as a small constant that limits the extent of permissible ratio variation. Moreover, $\mathcal{A}_t$ denotes the advantage estimation for token t, computed as

$$\mathcal{A}_t = r(\boldsymbol{x},\boldsymbol{y}) - \alpha\sum_{i=t}^{|\boldsymbol{y}|}\log\left[\frac{\pi_{\boldsymbol{\theta}^c_{\text{old}}\subseteq\boldsymbol{\theta}_{\text{old}}}(y_t|\boldsymbol{x},\boldsymbol{y}_{<t})}{\pi_{\text{ref}}(y_t|\boldsymbol{x},\boldsymbol{y}_{<t})}\right] + \gamma\frac{\mathcal{B}-\text{rank}(r(\boldsymbol{x},\boldsymbol{y}))}{\mathcal{B}-1}, \tag{3}$$

where α and γ are two hyperparameters, and π_{ref} denotes the reference policy. Moreover, the rank, $\text{rank}(r(\boldsymbol{x},\boldsymbol{y}))$, represents the position of $r(\boldsymbol{x},\boldsymbol{y})$ within the sorted list of rewards associated with a batch of samples, where $\mathcal{B}$ denotes the batch size. Unlike the standard REINFORCE++ [29] algorithm, we enhance the computation of the advantage function by integrating reward ranking information, which is modeled using a linear decay function based on the rank of the reward. This modification provides a robust and scale-invariant signal that encourages the model to focus on relative performance, thereby fostering more stable and reliable updates. The resulting advantage values are then normalized within each batch to ensure numerical stability.

The design of the reward function $r(\boldsymbol{x},\boldsymbol{y})$ plays a critical role in the training effectiveness of the REINFORCE++ algorithm. Our approach assesses model performance not only on the original samples but also on their augmented counterfactual and paraphrased variants. Specifically, we introduce four types of rewards, each corresponding to a specific dimension of model performance: accuracy, robustness, calibration, and confidence.

- **Accuracy** r_a measures the consistency between the predictions and the ground-truth answers, where GPT-4o is employed to assess prediction correctness by evaluating the semantic equivalence between the model-generated outputs and the reference texts for NLG tasks.
- **Robustness** r_o evaluates the model's ability to maintain consistent and accurate performance under input perturbations. We assess robustness using augmented counterfactual and paraphrased samples. For counterfactuals, robustness is measured by the prediction accuracy on the augmented samples. For paraphrases, it is quantified by calculating the cosine similarity between the hidden states of the model's responses to the original and paraphrased inputs[3]
- **Calibration** r_c measures the extent to which the model's predicted probabilities faithfully represent the true likelihood of outcomes. It is evaluated using two standard metrics: Expected Calibration Error (ECE) and Brier Score [3]. Detailed definitions and computation procedures for these two metrics are provided in Appendix 1.

[3]The metrics for augmented counterfactual and paraphrased samples are rescaled according to their respective evaluation scores, as detailed in Section 3.1.

Table 1: Comparison of gender and race debiasing performance using SEAT and downstream results on three NLU tasks. The best and second-best results are highlighted in bold and underlined. CDRO consistently surpasses previous baselines in both debiasing and downstream task performance.

Dataset	**SST-2**			**CoLA**			**QNLI**		
Metric	**Gender (↓)**	**Race (↓)**	**Acc. (↑)**	**Gender (↓)**	**Race (↓)**	**Mcc. (↑)**	**Gender (↓)**	**Race (↓)**	**Acc. (↑)**
BERT	0.29	0.30	92.4%	0.18	0.16	57.6%	0.37	0.30	91.3%
CDA	0.47	0.39	81.3%	0.29	0.30	53.2%	0.38	0.35	89.1%
Dropout	0.48	0.37	81.9%	0.27	0.31	52.2%	0.44	0.48	90.1%
Context-Debias	0.23	0.20	91.9%	0.47	0.32	55.4%	0.36	0.33	89.9%
Auto-Debias	0.28	0.31	92.1%	0.22	0.20	52.9%	0.24	0.24	91.1%
MABEL	0.35	0.28	92.2%	0.42	0.19	57.8%	0.44	0.30	91.6%
Sent-Debias	0.21	0.17	89.1%	0.22	0.20	55.4%	0.32	0.27	90.6%
FairFil	0.18	0.18	91.6%	0.12	0.14	56.5%	0.22	0.24	90.8%
Causal-Debias	0.11	0.11	92.9%	0.11	0.06	58.1%	0.15	0.11	91.6%
PCFR	0.09	0.13	91.9%	0.08	0.11	55.7%	0.11	0.13	89.2%
CDRO (Ours)	**0.05**	**0.06**	**94.2%**	**0.05**	**0.04**	**59.4%**	**0.07**	**0.07**	**92.8%**
ALBERT	0.22	0.29	92.6%	0.24	0.19	58.5%	0.21	0.20	91.3%
CDA	0.38	0.39	92.4%	0.16	0.18	53.1%	0.31	0.28	90.9%
Dropout	0.28	0.25	90.4%	0.25	0.27	47.4%	0.20	0.24	91.7%
Context-Debias	0.11	0.10	77.3%	0.17	0.14	55.4%	0.20	0.15	91.6%
Causal-Debias	0.08	0.13	92.9%	0.16	0.16	57.1%	0.09	**0.01**	91.6%
PCFR	0.06	0.10	92.3%	0.13	0.11	55.3%	0.08	0.11	89.4%
CDRO (Ours)	**0.04**	**0.07**	**93.8%**	**0.08**	**0.09**	**59.8%**	**0.05**	**0.01**	**92.5%**
RoBERTa	0.41	0.43	94.8%	0.41	0.38	57.6%	0.48	0.49	92.8%
Context-Debias	0.26	0.24	80.3%	0.30	0.35	55.4%	0.37	0.35	91.8%
Causal-Debias	0.09	0.10	93.9%	0.16	0.13	54.1%	0.09	0.05	92.9%
PCFR	0.06	0.09	93.5%	0.15	0.13	55.4%	0.07	0.10	89.4%
CDRO (Ours)	**0.04**	**0.06**	**96.5%**	**0.09**	**0.08**	**58.7%**	**0.05**	**0.03**	**93.7%**

- **Confidence** r_f evaluates the model's prediction confidence in generating a complete sequence from a given input by computing the product of the conditional probabilities of each token in the sequence: $\sqrt[|\boldsymbol{y}|]{\prod_{t=1}^{|\boldsymbol{y}|} p\left(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}\right)}$ [49].

Higher values of accuracy, robustness, and confidence metrics reflect improved model performance, whereas lower values of calibration metrics indicate better prediction reliability. Accordingly, the reward employed during optimization is defined as a weighted sum of the four reward components: $r = r_a + \lambda(r_o - r_c + r_f)$, where the value of λ is fixed as 0.5 in our experiments to maintain the relative dominance of the accuracy-related reward component.

During the optimization process, the logistic regression model is also updated in real-time to ensure dynamic and adaptive knowledge localization. Specifically, the update is performed using the REINFORCE algorithm [95], where the reward quantifies the variation in the LLMs' performance before and after each update. This performance variation is measured on a small validation set and evaluated using the four metrics described earlier. Consequently, the optimization is formulated as

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \tau \mathbb{E}_{\{(\boldsymbol{x},\boldsymbol{y})\}} \mathbb{E}_{z \sim p(z|\boldsymbol{\xi};\boldsymbol{\beta})} \left[r^{\text{dev}} \cdot \nabla_{\boldsymbol{\beta}} \log p\left(z|\boldsymbol{\xi};\boldsymbol{\beta}\right) \right], \tag{4}$$

where r^{dev} represents the computed reward signal and τ denotes the step size of each update.

4 Experiments

Extensive experiments have been conducted to evaluate the effectiveness of the proposed approach. First, we examine its ability to mitigate model biases across various NLU tasks. Next, we assess its effectiveness in reducing hallucinations on multiple NLG tasks. Finally, we evaluate its robustness in OOD scenarios. Due to space limitations, further details regarding the datasets, the compared baselines, and the experimental settings are provided in the Appendix.

Evaluation for Debiasing Ability. Unwanted stereotypical associations are known to degrade model performance [28, 45]. Building on prior research [106, 25], we use human-created stereotypes to investigate and mitigate biases in LLMs, specifically incorporating gender [35] and race [53] word lists. Experiments are conducted on three downstream tasks: SST-2 for sentiment classification [75], CoLA for grammatical acceptability judgment [90], and QNLI for question answering [69], utilizing

Table 2: Performance comparison between CDRO and other baselines across five NLG tasks. The proposed CDRO method consistently outperforms previous baselines in mitigating knowledge hallucinations, achieving the highest Acc@q and Cov@p scores. To ensure a fair comparison, the values of q and p are aligned with those configured in [49, 114].

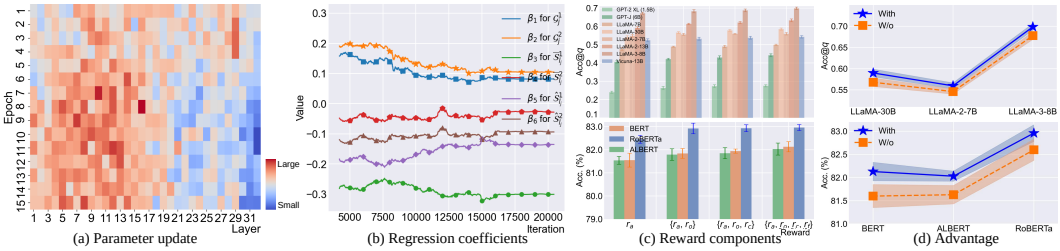
Task	**NQ**		**SciQ**		**TriviaQA**		**TruthfulQA**		**WikiQA**	
Metric	**Acc@50** (↑)	**Cov@50** (↑)	**Acc@50** (↑)	**Cov@90** (↑)	**Acc@50** (↑)	**Cov@60** (↑)	**Acc@50** (↑)	**Cov@40**(↑)	**Acc@50** (↑)	**Cov@50** (↑)
Label Smooth.	0.208	0.061	0.212	0.003	0.302	0.019	0.181	0.000	0.273	0.000
Temp. Scaling	0.288	0.115	0.764	0.211	0.500	0.111	0.314	0.136	0.388	0.012
LITCAB	0.300	0.105	0.762	0.221	0.478	0.201	0.314	0.195	0.397	0.062
Calib. Tuning	0.310	0.115	0.761	0.224	0.482	0.222	0.386	0.393	0.441	0.162
P(IK)	0.286	0.000	0.656	0.004	0.372	0.023	0.267	0.005	0.339	0.004
Verbalization	0.254	0.055	0.660	0.117	0.404	0.053	0.233	0.224	0.372	0.202
Self-Consis.	0.340	0.217	0.744	0.124	0.446	0.079	0.405	0.500	0.628	0.621
ITI	0.297	0.098	0.745	0.213	0.462	0.168	0.300	0.165	0.376	0.058
R-Tuning	0.293	0.084	0.692	0.119	0.400	0.063	0.341	0.332	0.416	0.258
HADEMIF	0.355	0.120	0.766	0.228	0.501	0.240	0.430	0.510	0.653	0.338
DoLa	0.301	0.108	0.759	0.224	0.476	0.205	0.316	0.190	0.400	0.125
SH2	0.322	0.101	0.760	0.221	0.482	0.225	0.352	0.479	0.478	0.297
CDRO (Ours)	**0.376**	**0.228**	**0.781**	**0.245**	**0.520**	**0.258**	**0.456**	**0.529**	**0.665**	**0.627**

three LLMs: BERT-base [17], ALBERT-large [41], and RoBERTa-base [51]. Unless otherwise specified, the training data are augmented using the LLaMA-3-70B [23] model. We report results as the average of five runs for each task. The compared baselines include a range of debiasing approaches, encompassing non-task-specific methods-CDA [91], Dropout [91], Context-Debias [35], Auto-Debias [25], and MABEL [27]—as well as task-specific methods, including Sent-Debias [45], FairFil [11], Causal-Debias [106], and PCFR [28]. Following prior research [106, 25], evaluation metrics consist of accuracy (or Matthew correlation for CoLA) and two bias assessment measures: SEAT [54] for both gender and racial bias, and CrowS-Pair [60] for gender bias. In SEAT, scores closer to 0 indicate lower bias, while in CrowS-Pair, scores approaching 50% reflect reduced stereotyping.

Table 1 presents the gender and race debiasing performance using the SEAT evaluation of various methods, alongside their accuracy on these tasks, with results using the CrowS-Pair evaluation provided in Appendix 3. **The proposed CDRO approach demonstrates superior effectiveness in mitigating gender and race bias, as evidenced by its lowest SEAT scores across all three tasks.** Moreover, while previous debiasing methods often degrade downstream task performance, CDRO not only achieves SOTA debiasing effectiveness but also enhances model performance in downstream applications. This advantage can be largely attributed to the selective and fine-grained optimization of the model parameters that are responsible for encoding causal relationships.

Evaluation for Hallucination Mitigation. We evaluate our approach on five representative NLG benchmarks: Natural Questions (NQ) [40], SciQ [92], TriviaQA [32], TruthfulQA [47], and WikiQA [100]. LLaMA-2-7B [81] is adopted as the primary backbone, given its widespread use in studying knowledge hallucination in LLMs. Additionally, we incorporate seven other popular LLMs with parameters ranging from 1.5B to 30B: GPT-2 XL (1.5B) [67], GPT-J (6B) [86], LLaMA-7B [80], LLaMA-30B, LLaMA-2-13B, LLaMA-3-8B [23], and Vicuna-13B [13]. To ensure a fair comparison, we adhere to the evaluation framework outlined in [49]. Specifically, the model's confidence is computed as the geometric mean of token probabilities. Moreover, GPT-4 [1] is employed to assess the correctness of model outputs by determining the semantic equivalence between the generated text and the reference. Subsequently, two metrics are utilized to evaluate the effectiveness of various approaches in hallucination mitigation: Acc@q and Cov@p. The Acc@q metric measures the precision of the model by evaluating the accuracy of the top-q percent of predictions. The Cov@p metric measures recall by identifying the largest proportion of the most confident predictions where accuracy exceeds a specified threshold p.

We compare CDRO with various approaches designed to enhance prediction reliability. These include model calibration techniques including Temperature Scaling [46], Label Smoothing [78], LITCAB[49], and Calibration Tuning[36], as well as hallucination detection and mitigation methods, including Verbalization [79], P(IK)[33], Self-Consistency[79], R-Tuning [102], DoLa [14], SH2 [34], ITI [43], and HADEMIF [114]. The results for LLaMA-2-7B are summarized in Table 2, with some values referenced from [114]. The evaluation outcomes for other LLMs are presented in Fig. 6 of the Appendix. **Our method demonstrates consistent superiority over existing baselines across all five tasks, attaining the highest Acc@q and Cov@p scores.** These findings highlight the effectiveness of our approach in mitigating hallucinations, which can be attributed to the suppression of spurious correlations and the enhancement of the model's causal reasoning capabilities.



(a) Parameter update (b) Regression coefficients (c) Reward components (d) Advantage

Figure 3: (a) Update of matrices across various layers during the training process. (b) Evolution of the regression coefficients on QNLI. (c) Average performance across ablations of four reward components on NLU and NLG tasks. (d) Ablation studies on reward ranking information for advantage estimation.

Table 3: Performance comparison on the OOD datasets utilizing the RoBERTa-base model. The proposed CDRO framework consistently achieves the highest accuracy among all compared baselines.

Dataset	**SST-2**		**MNLI**		**QQP**
OOD data	**IMDB-Cont**	**IMDB-CAD**	**HANS**	**AdvNLI**	**PAWS**
Fine-tuning	84.51%	88.39%	67.80%	31.22%	38.45%
Span Cutoff	85.53%	89.21%	68.38%	31.14%	38.80%
HiddenCut	87.82%	90.44%	71.16%	32.83%	41.52%
IPT-Adapter	85.01%	88.75%	66.30%	32.54%	38.94%
Causal-Debias	88.45%	91.44%	76.21%	37.53%	44.35%
PCFR	88.51%	91.78%	76.64%	38.01%	44.62%
CDRO (Ours)	**89.62%**	**92.65%**	**77.68%**	**39.40%**	**46.01%**

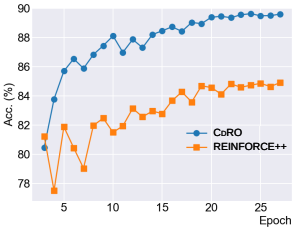


Figure 4: Accuracy comparison between CDRO and vanilla REINFORCE++ on IMDB-Cont using RoBERTa-base model.

Evaluation for OOD Generalization. Models influenced by spurious correlations in training data often exhibit degraded generalization, particularly in OOD scenarios [8]. Accordingly, we conduct experiments on three representative tasks from the GLUE benchmark [85]: SST-2 [75], MNLI [94], and QQP [89], each accompanied by publicly available OOD datasets. For SST-2, the OOD evaluation is conducted on the IMDB-Cont [21] and IMDB-CAD [37] datasets. The OOD datasets for MNLI comprise HANS [99] and AdvNLI [63], while PAWS-QQP [104] serves as the OOD dataset for QQP. We employ three widely utilized pretrained language models—BERT-base [17], RoBERTa-base [51], and BART-base [42]—and report accuracy as the evaluation metric. The compared baselines for improving model generalization in OOD scenarios include Span Cutoff [74], HiddenCut [8], IPT-Adapter [22], Causal-Debias [106], and PCFR [28].

Table 3 reports the comparative results using the RoBERTa-base model, while the comparison results for other models are provided in Appendix 5. As shown, **the proposed CDRO method significantly outperforms all compared baselines across a range of OOD datasets**. In particular, it yields average performance improvements of 1.16% over the strongest baseline and 7.00% over vanilla fine-tuning. These results demonstrate the effectiveness of CDRO in mitigating spurious correlations and enhancing the resilience capability of LLMs even under distributional shifts.

Analysis of Training Process. We investigate the update behavior of parameter matrices during training, categorizing them by layer position and functional type. From Fig. 3(a), the proportion of updates decreases in deeper layers after a period of training, indicating that the parameter matrices sensitive to causal relationships increasingly concentrate in the earlier and intermediate layers as the model converges. The update patterns for different matrix types are shown in Fig. 7 in the Appendix, where query and key matrices are primarily updated during the early stages, while value, up, and down matrices receive more updates later on. Additionally, Fig. 3(b) shows the evolution of the coefficients in the logistic regression model throughout training. The indicators $\mathcal{G}_j^1$ and $\mathcal{G}_j^2$ show positive correlations with predictions, while $\overline{S}_{l_j}^1$, $\overline{S}_{l_j}^2$, $\hat{S}_{l_j}^1$, and $\hat{S}_{l_j}^2$ exhibit negative correlations. These results suggest that layers with lower values of $\overline{S}_{l_j}^1$ and $\overline{S}_{l_j}^2$, and matrices with higher values of $\mathcal{G}_j^1$ and $\mathcal{G}_j^2$, are more sensitive to causal variations, indicating their crucial role in encoding causal signals.

Furthermore, layers with lower variances (i.e., $\hat{S}^1_{l_j}$ and $\hat{S}^2_{l_j}$) are more likely to be selected, as they consistently capture causal information across different samples.

Ablation and Sensitivity Studies. We conduct ablation studies on the four reward components. As shown in Fig. 3(c), the model attains its best performance when all four types of rewards are jointly incorporated, underscoring their complementary contributions. We then assess the impact of incorporating reward ranking information into advantage estimation. From the results presented in Fig. 3(d), the integration of reward ranking consistently leads to performance improvements. Furthermore, Fig. 4 presents the accuracy trajectories during training, demonstrating that CDRO steadily outperforms the vanilla REINFORCE++ in terms of accuracy.

5 Conclusion

This study presents a novel causality-informed robust optimization framework, termed CDRO, aimed at mitigating LLMs' reliance on spurious correlations and enhancing their resilience across diverse tasks. Our approach first identifies parameter components capturing causal relationships by analyzing training dynamics in weight matrices across original, counterfactual, and paraphrased samples. These dynamics are modeled via a logistic regression mechanism, enabling the automatic and adaptive localization of causality-relevant parameters. To further refine the optimization process, we introduce a collaborative reinforcement learning strategy that alternately updates the identified causal parameters and the logistic regression model. Extensive experiments on various NLU and NLG tasks demonstrate that CDRO consistently surpasses the compared baselines in mitigating spurious correlations, suppressing knowledge hallucinations, and enhancing overall model performance.

Acknowledgments

This work was supported by the NSFC under Grant 625B2009 and the 2025 Chinese Institute of Electronics-Tencent PhD Research Incentive Program.

References

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

[2] Alessandro Antonucci, Gregorio Piqué, and Marco Zaffalon. Zero-shot causal graph extrapolation from text via llms. *arXiv preprint arXiv:2312.14670*, 2023.

[3] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[4] Bradley Butcher. Aligning large language models with counterfactual dpo. *arXiv preprint arXiv:2401.09566*, 2024.

[5] Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, and Le Sun. Can prompt probe pretrained language models? understanding the invisible risks from a causal view. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5796–5808, 2022.

[6] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[7] Dongping Chen, Jiawen Shi, Yao Wan, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. Self-cognition in large language models: An exploratory study. *arXiv preprint arXiv:2407.01505*, 2024.

[8] 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.

[9] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[10] Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. Disco: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, 2023.

[11] 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.

[12] Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, and Kuan-Hao Huang. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1013–1025, 2024.

[13] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *LMSYS*, 2023.

[14] 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.

[15] Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. Lm vs lm: Detecting factual errors via cross examination. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 12621–12640, 2023.

[16] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022.

[17] 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.

[18] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120, 2023.

[19] Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. Discovering salient neurons in deep nlp models. *Journal of Machine Learning Research*, 24(362):1–40, 2023.

[20] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Anshita Gupta, Debanjan Mondal, Akshay Sheshadri, Wenlong Zhao, Xiang Li, Sarah Wiegreffe, and Niket Tandon. Editing common sense in transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 8214–8232, 2023.

[27] 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.

[28] 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.

[29] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

[30] 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.

[31] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, 2023.

[32] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[39] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[40] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[41] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations*, 2020.

[42] 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.

[43] 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.

[44] Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 13201–13221, 2024.

[45] 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.

[46] 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.

[47] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022.

[48] Victoria Lin, Eli Ben-Michael, and Louis-Philippe Morency. Optimizing language models for human preferences is a causal inference problem. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 2250–2270, 2024.

[49] 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.

[50] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[51] 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.

[52] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, 2023.

[53] 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.

[54] 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.

[55] Daniel Mela, Aitor González-Agirre, Javier Hernando, and Marta Villegas. Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5831–5847, 2024.

[56] 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.

[57] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 17359–17372, 2022.

[58] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

[59] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2022.

[60] 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.

[61] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

[62] Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. Moca: Measuring human-language model alignment on causal and moral judgment tasks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 78360–78393, 2023.

[63] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, 2020.

[64] Ryusei Ohtani, Yuko Sakurai, and Satoshi Oyama. Does metacognitive prompting improve causal inference in large language models? In *Proceedings of the 2024 IEEE Conference on Artificial Intelligence*, pages 458–459, 2024.

[65] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 27730–27744, 2022.

[66] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 79155–79172, 2023.

[67] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[68] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 53728–53741, 2023.

[69] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, 2018.

[70] Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 31450–31465, 2023.

[71] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Proceedings of the Advances in Neural Information Processing Systems*, pages 55565–55581, 2024.

[72] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[73] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[74] 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.

[75] 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.

[76] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 3008–3021, 2020.

[77] Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 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*, pages 6883–6893, 2024.

[78] 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.

[79] 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.

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

[81] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[82] Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 82235–82264, 2024.

[83] Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

[84] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Proceedings of Advances in Neural Information Processing Systems*, pages 12388–12401, 2020.

[85] 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.

[86] Ben Wang and Aran Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021.

[87] Qianlong Wang, Keyang Ding, Bin Liang, Min Yang, and Ruifeng Xu. Reducing spurious correlations in aspect-based sentiment analysis with explanation from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2930–2941, 2023.

[88] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021.

[89] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*, 2017.

[90] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

[91] 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.

[92] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.

[93] Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, and Minlie Huang. Autocad: Automatically generate counterfactuals for mitigating shortcut learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2302–2317, 2022.

[94] 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.

[95] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[96] Sam Witteveen and Martin Andrews. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, 2019.

[97] Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. Causality for large language models. *arXiv preprint arXiv:2410.15319*, 2024.

[98] Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. Depn: Detecting and editing privacy neurons in pretrained language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, 2023.

[99] Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, 2022.

[100] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, 2015.

[101] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025.

[102] 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.

[103] Shaolei Zhang, Tian Yu, and Yang Feng. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949, 2024.

[104] 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.

[105] Junhao Zheng, Qianli Ma, Shengjie Qiu, Yue Wu, Peitian Ma, Junlong Liu, Huawen Feng, Xichen Shang, and Haibin Chen. Preserving commonsense knowledge from pre-trained language models via causal inference. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9155–9173, 2023.

[106] 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.

[107] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, pages 61–68, 2024.

[108] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 56–66, 2025.

[109] Pengfei Zhou, Jie Xia, Xiaopeng Peng, Wangbo Zhao, Zilong Ye, Zekai Li, Suorong Yang, Jiadong Pan, Yuanxiang Chen, Ziqiao Wang, et al. Neural-driven image editing. *arXiv preprint arXiv:2507.05397*, 2025.

[110] Xiaoling Zhou and Ou Wu. Implicit counterfactual data augmentation for deep neural networks. *arXiv preprint arXiv:2304.13431*, 2023.

[111] Xiaoling Zhou, Wei Ye, Zhemg Lee, Lei Zou, and Shikun Zhang. Valuing training data via causal inference for in-context learning. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[112] Xiaoling Zhou, Wei Ye, Yidong Wang, Chaoya Jiang, Zhemg Lee, Rui Xie, and Shikun Zhang. Enhancing in-context learning via implicit demonstration augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2810–2828, 2024.

[113] Xiaoling Zhou, Wei Ye, Rui Xie, and Shikun Zhang. Mitigating spurious correlations with causal logit perturbation. *Information Sciences*, page 122276, 2025.

[114] 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.

[115] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*, 2023.

[116] Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we provide a comprehensive explanation of the motivation and research direction of this work, along with a detailed summary of the key contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Appendix 8.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All formulas in this manuscript are properly numbered and consistently cross-referenced throughout the text.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental settings are detailed in Section 4 and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets utilized in this study are publicly available and our code will be made publicly available upon acceptance of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide a comprehensive description of the experimental settings—including datasets, hyperparameters, and the optimization process—in Section 4 and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the standard errors of the results in Section 4 and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details regarding the computational resources are provided in Section 4 and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research in this paper complies with NeurIPS ethical standards in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential broader impacts have been discussed in Appendices 8 and 9.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper presents no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have provided appropriate citations and explanations for the papers, models, and datasets referenced in this work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: In Sections 3 and 4, and the Appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.