Position Debiasing Fine-Tuning for Causal Perception in Long-Term Dialogue

Shixuan Fan[1,2] , **Wei Wei**[*1,2] , **Wendi Li**[1,2] , **Xian-Ling Mao**[3] , **Wenfeng Xie**[2,4] and **Dangyang Chen**[*2,4]

[1]Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology
[2]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
[3]Department of Computer Science and Technology, Beijing Institute of Technology
[4]Ping An Property & Casualty Insurance Company of China, Ltd
{fanshixuan, weiw, wendili}@hust.edu.cn, maoxl@bit.edu.cn, julian_wind@163.com, chendangyang273@pingan.com.cn

Abstract

The core of the dialogue system is to generate relevant, informative, and human-like responses based on extensive dialogue history. Recently, dialogue generation domain has seen mainstream adoption of large language models (LLMs), due to its powerful capability in generating utterances. However, there is a natural deficiency for such models, that is, inherent position bias, which may lead them to pay more attention to the nearby utterances instead of causally relevant ones, resulting in generating irrelevant and generic responses in long-term dialogue. To alleviate such problem, in this paper, we propose a novel method, named Causal Perception long-term Dialogue framework (CPD), which employs perturbation-based causal variable discovery method to extract casually relevant utterances from the dialogue history and enhances model causal perception during fine-tuning. Specifically, a local-position awareness method is proposed in CPD for inter-sentence position correlation elimination, which helps models extract causally relevant utterances based on perturbations. Then, a casual-perception fine-tuning strategy is also proposed, to enhance the capability of discovering the causal invariant factors, by differently perturbing causally relevant and non-casually relevant ones for response generation. Experimental results on two datasets prove that our proposed method can effectively alleviate the position bias for multiple LLMs and achieve significant progress compared with existing baselines.

1 Introduction

The design of the dialogue systems aspires to generate consistent, controllable, and diverse responses based on dialogue history [Liu *et al.*, 2022b; Lu *et al.*, 2023]. Initial researches on dialogue systems are often constrained by

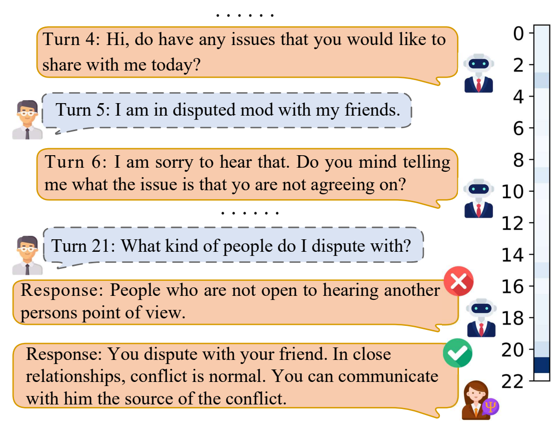

Figure 1: Position bias in large language models (Llama2-7B-chat). Dotted boxes mark relevant utterances. The strip on the right shows the average attention of each turn in the dialogue. Darker colors represent higher attention values.

the length of the input window, with a common assumption that responses are primarily influenced by the last one or a few last turns of the dialogue history [Wei *et al.*, 2021; Liu *et al.*, 2022a]. Recently, with the development of large language models (LLMs), researchers propose long-term dialogue tasks and explore the use of rich information in dialogue history for response generation [Xu *et al.*, 2022b].

Despite the impressive achievements of LLMs in open-domain dialogues, capturing causal relationships within extensive dialogue history remains challenging [Feng *et al.*, 2023]. Some researchers identify that LLMs suffer from severe position bias, focusing only on the final context and disregarding historical information [Liu *et al.*, 2023a; Liu *et al.*, 2023b]. This bias leads models to concentrate on spurious position correlations instead of causally relevant utterances, resulting in irrelevant and generic response generation. As shown in Figure 1, Llama2's attention in the last turn is significantly higher than in other turns, neglecting the relevance utterance in the 5-th turn and resulting in the generation of context-irrelevant and generic responses. To address the above problems, some studies aim

[*]Corresponding authors.

to compress long-term dialogue history using retrieval-based [Feng *et al.*, 2023] or summary-based [Wang *et al.*, 2023; Lee *et al.*, 2023] methods. Although these works achieve certain results, none substantially improve the ability of LLMs to perceive genuinely causally relevant utterances.

To eliminate position bias and enhance the perception of causal relationships in long-term dialogue, two significant challenges must be addressed: (1) Lack of large-scale dialogue datasets with causally relevant utterance annotations. Perturbation-based causal variable discovery methods assume that models can effectively utilize all input variables. They consider the difference in potential outcomes when binary intervention (presence or exclusion) acts on the variable as the treatment effect. However, the inherent position bias in LLMs hinders their ability to fully leverage dialogue history, making perturbation-based methods unable to be directly used to extract causally relevant utterances. (2) Traditional text generation loss is insufficient in guiding models to eliminate spurious correlations between relevant utterances and position distributions. Models tend to learn imbalances in the distribution of relevant utterance positions in datasets. Existing position debiasing methods primarily disrupt positions, posing challenges in maintaining the consistency of semantic structures when applied to dialogue scenarios.

To address the aforementioned challenges, we propose a model-free Causal Perception long-term Dialogue framework, named CPD. Firstly, we extract relevant utterances for two widely used long-term dialogue datasets using causal perturbation. Analyzing the impact of position bias on the model's causal perception, we introduce a local-position awareness method to mitigate this bias by eliminating inter-sentence position information. Through sentence-level perturbations on dialogues, we regard the change in the perplexity [Horgan, 1995] before and after the perturbation as the treatment effect of the perturbed utterance. Validation of sentence-level conditional independence in sets of causally relevant utterances, followed by the application of clustering algorithms to categorize dialogue history into causally relevant or irrelevant parts. Secondly, inspired by invariance learning [Chang *et al.*, 2020], the core of preventing models from fitting position bias is to direct models to concentrate on invariant causal variables. We argue that responses and their corresponding causally relevant utterances should exhibit consistency under perturbations. Specifically, we conduct sentence-level perturbations separately on causally relevant and non-causally relevant parts, encouraging the model's generation to be either consistent or inconsistent with gold responses based on whether the causal variables are perturbed or not. A positional difference sampling strategy, based on the probability of relevant utterance positions, is employed to address the imbalance of relevant utterance positions without compromising the integrity of the dialogue structure.

The contributions of this paper are listed as follows:

- We propose a relevant utterance extraction method based on causal perturbation, which can avoid the interference caused by the position bias of LLMs through local-position awareness.
- We propose a causal perception fine-tuning strategy that effectively alleviates the model's position bias and enhances its ability to perceive correlations in dialogues.
- Experimental results on two benchmark datasets demonstrate that our proposed method can consistently outperform the state-of-the-art baselines in terms of objective and subjective evaluation metrics.

2 Related Work

2.1 Position Bias and Long-Term Dialogue

Position bias in language models has undergone extensive examination. Ko *et al.*[2020] observed that language models introduced position bias due to the imbalance position distribution of relevant sentences in datasets, relying on fixed-position utterances instead of real causal correlation during inference. Wang *et al.*[2021] and Liu *et al.*[2023a] respectively confirmed the existence of position bias in a variety of pre-trained models utilizing different position embeddings and a variety of LLMs of different sizes.

Existing position debiasing methods primarily involved disrupting the position information in datasets and constraining the consistency of the original model during the fine-tuning process. Ko *et al.*[2020] and RPP [Amor *et al.*, 2023] applied random position perturbations at the document and word level separately. Some other works divide context into short segments [Ivgi *et al.*, 2023; Li *et al.*, 2023], or reorder input context based on relevance [Peysakhovich and Lerer, 2023; He *et al.*, 2023]. However, dialogues exhibit temporal structure, and position perturbation can destroy dialogue structures, leading to a substantial decrease in comprehension. ZOE [Liu *et al.*, 2024] required the model to fit both task labels and original model output, alleviating the impact of position imbalance by enforcing consistency between the fine-tuned model and the original model. Unfortunately, in dialogue scenarios, the inherent position bias of the model aligns with the position imbalance in the data, causing the method to fail to a certain extent.

To explore the model's ability to comprehend rich semantics in long-term dialogue history, Xu *et al.* [2022a] constructed a multi-turn long-term dialogue dataset. Existing works primarily focused on enabling LLMs to acquire long-term dialogue awareness by compressing dialogue history to alleviate position bias. RSM [Wang *et al.*, 2023] and Lee *et al.*[2023] employed LLMs to summarize dialogue history as an external memory pool, enhancing LLMs with long-term memory capabilities. CONSTRAIN [Feng *et al.*, 2023] assumed that only two sentences in the dialogue history were relevant to the response, utilizing a trained language model to retrieve relevant utterances. However, these methods suffer from dual challenges of information loss during compression and a lack of substantial improvement in the model's inherent ability to comprehend extended contexts.

2.2 Causal Inference in NLP

Causal inference is a method used to determine whether correlations in data truly reflect cause-and-effect relationships [Pearl *et al.*, 2016; Alaa and Van Der Schaar, 2019]. Works on causal inference in NLP mainly focused on word granular detoxification in pre-trained language models. For instance,

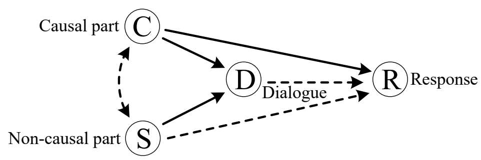

Figure 2: Causal view of response generation, where the solid line represents the causal relationship between two variables, and the dotted line represents the probabilistic dependencies.

Abraham *et al.*[2022] and Madhavan *et al.*[2023] both engaged in word-level language model detoxification, employing average treatment effect and counterfactual enhancement. Wang *et al.*[2023] is similar to our work. They annotate a small test set of causally relevant utterances, named CGDIALOG, and discovered that the causal perception ability of language models is deficient, meaning that language models exhibit similar outcomes to perturbations of causally relevant or non-causally relevant utterances. They simply attributed this problem to overfitting. In our investigation, we delved further and identified position bias as the root cause of this phenomenon.

3 Methodology

3.1 Causal View for Response Generation

Structure Causal Model (SCM) [Shanmugam, 2001] is a method of describing causal correlations among multiple variables. In this paper, we describe the causal structure in dialogue as a causal graph composed of four variables: input dialogue D, response R, causal part C, and non-causal part S. Figure 2 illustrates the SCM of dialogue.

Dialogue D consists of a causal part C and a non-causal part S of the response ($C \rightarrow D \leftarrow S$), where the causal part C is the minimum set of corresponding endogenous cause utterances leading to the response R ($C \rightarrow R$). An outstanding language model should exhibit sensitivity to causal part C while maintaining robustness to non-causal part S. Unfortunately, due to the probabilistic correlation between the causal part C and the non-causal part S ($C \dashleftarrow\dashrightarrow S$), including the imbalanced distribution of causal part utterances across positions, often leads models to overfit position information, fixate on utterances at nearby positions, and lose the ability to perceive causality ($S \dashrightarrow R$).

Treatment Effect (TE) [Austin, 2011] quantifies the impact of a variable on the target outcome. In practice, TE is often assessed through the conditional independence assumption [Dawid, 1980] and counterfactual reasoning [Rafetseder *et al.*, 2013]. For treatment variable u_i, a binary treatment is employed to evaluate the treatment effect of the outcome. The TE of utterance u_i in dialogue D is defined as:

$$\text{TE}(u_i) = f(D) - f(D \setminus u_i), \quad (1)$$

where $D \setminus u_i$ represents the dialogue when utterance u_i is disturbed, $f(\cdot)$ is the perplexity [Horgan, 1995] of the language model to generate correct responses. To eliminate the potential impact of changes in other co-variables after the utterance absence, we use meaningless utterances of similar lengths, such as 'hello,' 'thank you,' *etc.*, to replace u_i to construct the counterfactual condition of the absence treatment.

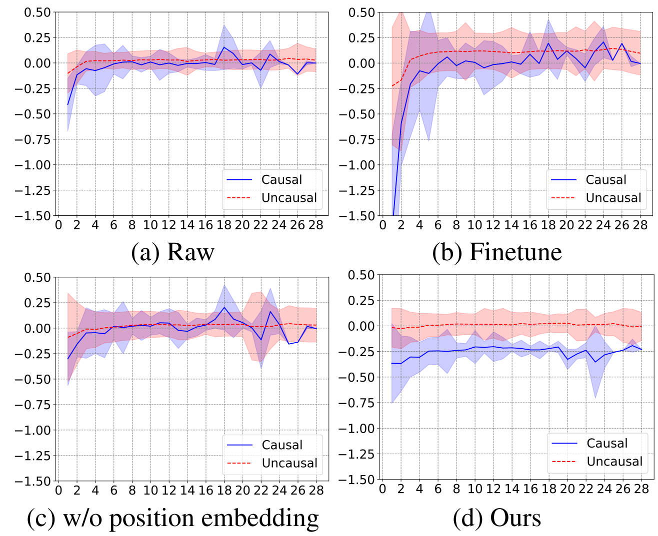

Figure 3: Llama2-7B-chat's ability to identify causally relevant utterances in the CGDIALOG dataset (ESConv part). The x-axis represents the turn distance from the utterance to response, while the y-axis corresponds to TE_{reg}. The solid blue line represents causal utterances, the dashed red line indicates non-causal utterances, and the shaded area represents variance.

3.2 Causal Identification and Position Bias

Outstanding language models should adeptly identify causal correlations in dialogues. Specifically, the TE of causally relevant utterances should be significantly higher than the TE of non-causal utterances. Unfortunately, Abraham *et al.*[2022] observe that language models lack this identification ability. We further observe that the model's causal identification ability is strongly tied to the position of causally relevant utterances in the dialogue. The normalized treatment effect $\text{TE}_{\text{reg}}(u_i)$ is employed to measure the causal correlation between utterances and responses.

$$\text{TE}_{\text{reg}}(u_i) = [f(D) - f(D \setminus u_i)]/f(D). \quad (2)$$

As shown in Figure 3 (a) and (b), we find that: (1) The LLMs can only identify the causal correlation in the last 1 and 2 turns of dialogue. (2) Irrespective of whether the utterances are relevant or not, LLMs consistently exhibit greater sensitivity to perturbations in the last few turns of dialogues. (3) Despite fine-tuning on domain data enhancing the model's sensitivity to causally relevant utterances, the accurate distinction remains challenging in longer dialogue histories.

This indicates that LLMs have the potential to distinguish between relevant and irrelevant utterances. One way to eliminate the influence of position information is to remove all position embedding, as shown in Figure 3 (c). After removing position embeddings, the model no longer focuses on position-specific utterances. Unfortunately, it also results in the damage of semantic information, rendering the LLMs incapable of identifying relevant utterances.

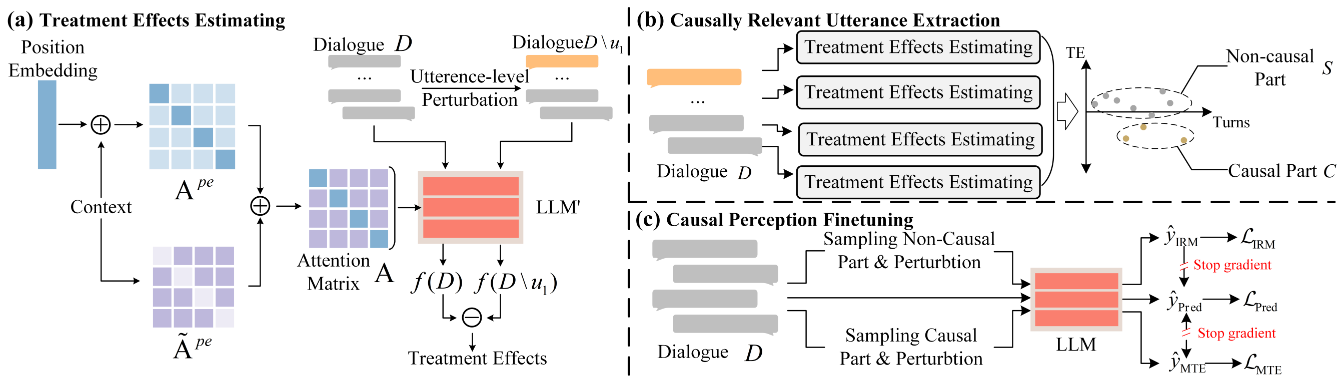

Figure 4: The Framework of our proposed method.

3.3 Causally Relevant Utterance Extraction

To achieve a balance between semantic information and position bias, we propose a sentence-level local-position awareness method for each layer of LLMs. Specifically, we restrict position information within sentences, and inter-sentence attention only uses semantic correlations. To ensure the method's applicability to models employing different positional embeddings, we directly modify the attention matrix. The model utilizes attention with position embeddings $\mathbf{A}^{pe}_{t,s}$ when the input words are in the same sentence. In contrast, when the input words are not in the same sentence, the attention is without position embeddings $\tilde{\mathbf{A}}^{pe}_{t,s}$:

$$\mathbf{A}_{t,s} = \begin{cases} \mathbf{A}^{pe}_{t,s}, & \text{if } x_t, x_s \text{ in the same utterance,} \\ \tilde{\mathbf{A}}^{pe}_{t,s} \times m^{pe}_A / \tilde{m}^{pe}_A, & \text{else ,} \end{cases} \tag{3}$$

where x_t and x_s denote the t-th and s-th input word, m^{pe}_A and $\tilde{m}^{pe}_A$ are the average values of attention matrices $\mathbf{A}^{pe}$ and $\tilde{\mathbf{A}}^{pe}$ respectively, used to balance the difference between two attention matrices.

After fine-tuning, the model's performance is shown in Figure 3 (d). It can be observed that our method effectively mitigates position bias, leading to improved identification of causally relevant utterances. To extract the minimal causally relevant utterance set in dialogues, we individually measure the TE of each utterance in the dialogue, denoted as $[\text{TE}(u_1), \text{TE}(u_2), ..., \text{TE}(u_{|D|})]$, where $|D|$ represents the number of dialogue turns. Due to the difference in TE between causally relevant and non-causally relevant utterances, we employ a simple clustering algorithm, K-means [Krishna and Murty, 1999], to obtain the causally relevant C and the non-causally relevant utterance set S. The initial clustering centers of the two sets are initialized as the minimum and median of the input data, respectively.

After verifying the effectiveness of our method with the 88.6% precision on the CGDIALOG test set, we extract relevant utterances from two long-term dialogue datasets, ESConv [Liu *et al.*, 2021] and MSC [Xu *et al.*, 2022a]. We calculate the position distribution Q of causally relevant utterances, where $q_i \in Q$ represents the frequency of causally relevant utterances with the i-th turn distance to the response. A severe imbalance is observed in the position distribution of causally relevant utterances, which might be the cause of model position bias.

3.4 Causal Perception Finetuning

The fine-tuning of LLMs follows the paradigm of instruction fine-tuning, where instructions and dialogue D are concatenated and fed into the model for generating responses R.

$$\begin{aligned} p(R) &= p(R \mid instruction, D) \\ &= \prod_t p\left(r_{t+1} \mid instruction, D, r_1, r_2, \ldots, r_t\right), \end{aligned} \tag{4}$$

During the fine-tuning process, our objective is for the model to acquire domain knowledge from the data while being sensitive to causal correlations. The loss during the fine-tuning process is divided into two parts: prediction loss and causal perception loss. The prediction loss ensures that the model accurately generates the gold responses and captures domain knowledge in the training data. The causal perception loss is employed to enhance the model's sensitivity to causal correlations.

$$\mathcal{L} = \mathcal{L}_{\text{Pred}} + \underbrace{\alpha\mathcal{L}_{\text{IRM}} + \beta\mathcal{L}_{\text{MTE}}}_{\text{causal perception}}, \tag{5}$$

where α and β are used to scale the three losses to similar orders of magnitude. The prediction loss is to maximize the cross-entropy between model predictions and gold responses.

$$\mathcal{L}_{\text{Pred}} = -\sum_{r_t \in R} log(p(r_t \mid instruction, D)). \tag{6}$$

For the sake of brevity, the *instruction* in the formula is omitted following. The causal perception loss is designed to require the model to focus on the causally relevant utterances while being sensitive to perturbations in environmental variables, such as non-relevant utterances. The causal perception loss consists of two parts: invariant risk minimization $\mathcal{L}_{\text{IRM}}$ and maximizing treatment effect $\mathcal{L}_{\text{MTE}}$.

Invariant risk minimization (IRM) [Arjovsky *et al.*, 2019] drives the model to grasp causal invariance across environments, that is, the model's outcomes should remain consistent across various environments, which are constructed by perturbing non-causally relevant utterances. For dialogue

$D = \{u_1, u_2, \ldots, u_{|D|}\}$ consisting of multiple utterances, we construct counterfactual dialogue $D \setminus u_i$ in which non-causally relevant utterances $u_i \in S$ are replaced. To ensure that the replacements do not significantly alter the semantic structure of dialogues, we randomly select utterances for substitution from non-causally relevant utterances in other dialogues. Invariant risk minimization loss minimizes the KL divergence of responses generated by counterfactual and original dialogues.

$$\mathcal{L}_{\text{IRM}} = \sum_{r_i \in R} KL(p_{sg}(r_i|D)||p(r_i|D \setminus u_i)), u_i \in S, \quad (7)$$

where sg means stopping gradient during backpropagation.

Maximizing treatment effect (MTE) [Marshall *et al.*, 2005] aims to help the model learn the consistency relationship between responses and corresponding causal variables. When causally relevant utterances are replaced with those from other dialogues, the loss expects models to generate utterances that exhibit maximum dissimilarity compared to the original dialogue model output.

$$\mathcal{L}_{\text{MTE}} = -\sum_{r_i \in R} KL(p_{sg}(r_i|D)||p(r_i|D \setminus u_i)), u_i \in C. \quad (8)$$

Sampling strategy. To overcome the imbalance position distribution of relevant utterances, we employ a sampling strategy that involves enforcing more perturbations at positions with the low likelihood of being a relevant utterance in the dataset. At the dialogue level, multiple perturbations are performed for each dialogue when calculating causal perception loss. The number of times n for both auxiliary tasks IRM and MTE is determined through the following formula:

$$n = \lfloor |C| / \sum_{u_i \in C} (q_{|D|-i}) \rfloor, \quad (9)$$

where $\lfloor \cdot \rfloor$ denotes rounding down, $|\cdot|$ represents the number of elements in the set, and $q_{|D|-i}$ represents the frequency of the i-th utterance in the dataset being a causally relevant utterance.

At the utterance level, each time the auxiliary task is calculated, the probability $S(u_i)$ of perturbing each sentence u_i is different. For IRM and MTE, the probability that utterance u_i is perturbed is,

$$S_{\text{IRM}}(u_i) = \begin{cases} 0, \text{ if } u_i \in C, \\ q_{|D|-i}, \text{ if } u_i \in S. \end{cases} \quad (10)$$

$$S_{\text{MTE}}(u_i) = \begin{cases} (1/q_{|D|-i}) / \sum_{u_j \in C} (1/q_{|D|-j}), \text{ if } u_i \in C, \\ 0, \text{ if } u_i \in S, \end{cases} \quad (11)$$

4 Experiments

4.1 Datasets

To evaluate the effectiveness of our proposed method, following previous works [Wang *et al.*, 2023; Feng *et al.*, 2023], we conduct experiments on two widely used benchmark datasets, **ESConv** [Liu *et al.*, 2021] and **MSC** [Xu *et al.*, 2022a], for long-term dialogue. We use the same data preprocessing and train/valid/test splitting strategy as in [Feng *et al.*, 2023].

4.2 Baselines

To demonstrate the effectiveness of our proposed method, we compare it with three kinds of baselines: **(1) Raw and fine-tuned LLMs** have outstanding performance in open-domain dialogue tasks. The model can improve its performance in the tasks during fine-tuning in the corresponding task domain. **(2) Long-term dialogue methods. RSM** [Wang *et al.*, 2023] continuously summarizes long-term dialogues and uses the summary as external memory to alleviate the memory forgetting of long-term dialogues in LLMs. **CONSTRAIN** [Feng *et al.*, 2023] assumes that except for the last sentence in the dialogues, there is only one sentence relevant to the response. Relevant utterances in dialogue history are retrieved through a trained language model and concatenated with the last sentence as input for response generation. **(3) Position debiasing methods. RPP** [Amor *et al.*, 2023] is extended to sentence-level position random perturbation to ensure that the training data no longer has imbalances in position distribution. **ZOE** [Liu *et al.*, 2024] fits both gold response and the suboptimal response generated by the original model to enforce consistency between the fine-tuned and original model. For the fairness of the experiments, the backbones in all baselines are replaced by two widely used LLMs, Llama2-7B-chat and Qwen-14B-chat.

4.3 Evaluation Metrics

Automatic Evaluation. (1) **Word Overlap.** We report *BLEU-n* (n=1, 2) [Papineni *et al.*, 2002] and ROUGE-L [Lin and Och, 2004] to evaluate the coherence and word overlap of generated utterances. (2) **Diversity.** We employ *Distinct-n* (n=1, 2) [Li *et al.*, 2016] to evaluate the diversity of the generated response.

Human Evaluation. We adopt *Relevance*, *Fluency*, and *Informativeness* of the generated utterances with the rating range of [0, 2]. We recruit three experienced annotators to evaluate 100 randomly selected dialogues with a length of more than 20 turns. The Fleiss Kappa is 0.72, indicating consistency in the estimates of annotators.

4.4 Implementation Details

Throughout the experiments, we use Adam optimizer [Kingma and Ba, 2015] with 3e-4 initial learning rate and the 128 batch size. All methods are trained for up to 12 epochs. To improve experimental efficiency, we use lora [Hu *et al.*, 2021] with rank 32 to fine-tune large language models. Both training and inference use 4-bit weight quantization by bitsandbytes [Dettmers *et al.*, 2022].

4.5 Main Result

The main evaluation results are shown in Tabel 1.

Automatic Evaluation. CPD attains optimal performance by effectively mitigating the model's position bias and enhancing its causal perception ability. Long-term dialogue methods alleviate the poison of position bias by compressing long-term dialogue history through summarization and retrieval, respectively. Notably, in the ESConv dataset with shorter dialogue turns, summary-based RSM outperforms CONSTRAIN, while retrieval-based CONSTRAIN excels in

Dataset	Backbone	Method	BLEU-1	BLEU-2	ROUGE-L	Distinct-1	Distinct-2	Relevance	Fluency	Informativeness
MSC	Llama2-7B	(1) Raw LLM	0.0911	0.0250	0.0860	0.0353	0.3051	1.50	1.84	1.43
		(1) Finetuned LLM	0.1037	0.0283	0.0995	0.0371	0.3288	1.56	**1.87**	1.49
		(2) RSM	0.1127	0.0336	0.1117	0.0373	0.3150	1.62	1.82	1.64
		(2) CONSTRAIN	0.1167	0.0342	0.1118	0.0118	0.1410	1.68	1.86	1.69
		(3) RPP	0.0927	0.0294	0.0877	0.0473	0.3650	1.47	**1.87**	1.48
		(3) ZOE	0.1076	0.0328	0.1109	0.0446	0.3416	1.61	1.84	1.65
		CPD (Ours)	**0.1245**	**0.0441**	**0.1214**	**0.0655**	**0.4233**	**1.77**	1.86	**1.78**
	Qwen-14B	(1) Raw LLM	0.0973	0.0321	0.1041	0.0443	0.3199	1.51	1.86	1.49
		(1) Finetuned LLM	0.1199	0.0364	0.1044	0.0486	0.3260	1.54	1.85	1.52
		(2) RSM	0.1175	0.0379	0.1082	0.0526	0.3437	1.64	1.84	1.65
		(2) CONSTRAIN	0.1258	0.0451	0.1233	0.0204	0.1474	1.73	1.86	1.70
		(3) RPP	0.1076	0.0345	0.1059	0.0646	0.4416	1.53	1.85	1.50
		(3) ZOE	0.1124	0.0318	0.1100	0.0632	0.4282	1.63	1.86	1.67
		CPD (Ours)	**0.1462**	**0.0519**	**0.1381**	**0.0887**	**0.5389**	**1.82**	**1.87**	**1.84**
ESConv	Llama2-7B	(1) Raw LLM	0.0713	0.0181	0.0739	0.0578	0.3723	1.54	1.85	1.47
		(1) Finetuned LLM	0.0842	0.0286	0.1047	0.0614	0.4018	1.57	1.85	1.52
		(2) RSM	0.0949	0.0318	0.1098	0.0766	0.4404	1.64	**1.87**	1.68
		(2) CONSTRAIN	0.0919	0.0276	0.1038	0.0118	0.1293	1.63	1.84	1.67
		(3) RPP	0.0670	0.0258	0.0972	0.0870	0.4679	1.50	1.80	1.44
		(3) ZOE	0.0943	0.0297	0.1041	0.0821	0.4532	1.61	1.85	1.67
		CPD (Ours)	**0.1091**	**0.0468**	**0.1324**	**0.0976**	**0.5094**	**1.79**	1.86	**1.81**
	Qwen-14B	(1) Raw LLM	0.1077	0.0326	0.1018	0.0744	0.4068	1.64	1.86	1.49
		(1) Finetuned LLM	0.1145	0.0372	0.1118	0.0825	0.4230	1.66	1.85	1.53
		(2) RSM	0.1237	0.0379	0.1168	0.0874	0.4413	1.74	1.84	1.73
		(2) CONSTRAIN	0.1205	0.0301	0.1122	0.0302	0.2564	1.73	**1.87**	1.71
		(3) RPP	0.1095	0.0307	0.1056	0.0939	0.4782	1.59	1.84	1.54
		(3) ZOE	0.1195	0.0364	0.1147	0.0847	0.4437	1.69	1.86	1.70
		CPD (Ours)	**0.1489**	**0.0591**	**0.1442**	**0.1125**	**0.5459**	**1.84**	**1.87**	**1.88**

Table 1: The performance of our proposed method and all baselines. The results of the best and the second performance in each column are in **boldface** and underlined, respectively.

Backbone	Methods	MSC					ESConv				
		BLEU-1	BLEU-2	ROUGE-L	Distinct-1	Distinct-2	BLEU-1	BLEU-2	ROUGE-L	Distinct-1	Distinct-2
Llama2-7B	CPD	0.1245	0.0441	0.1214	0.0655	0.4233	0.1091	0.0468	0.1324	0.0976	0.5094
	w/o IRM	0.1036	0.0327	0.1102	0.0523	0.4063	0.0906	0.0310	0.1116	0.0895	0.4876
	w/o MTE	0.1165	0.0375	0.1109	0.0423	0.3420	0.0916	0.0369	0.1260	0.0768	0.4412
	w/o sampling	0.1074	0.0334	0.1009	0.0570	0.3882	0.0911	0.0321	0.1194	0.0856	0.4734
Qwen-14B	CPD	0.1462	0.0519	0.1381	0.0887	0.5389	0.1489	0.0591	0.1442	0.1125	0.5459
	w/o IRM	0.1145	0.0431	0.1128	0.0753	0.4883	0.1225	0.0466	0.1239	0.1072	0.5107
	w/o MTE	0.1253	0.0480	0.1271	0.0658	0.3743	0.1381	0.0534	0.1312	0.0946	0.4792
	w/o sampling	0.1179	0.0461	0.1257	0.0784	0.4691	0.1320	0.0496	0.1269	0.1047	0.5079

Table 2: The performance of ablation study.

the longer MSC dataset. These phenomenons underscore that while iterative summarization may retain more information in the short term, it leads to inevitable information loss with multiple iterations. CONSTRAIN neglects the potential significance of multiple causally relevant historical utterances, exhibiting suboptimal performance for shorter dialogues. Additionally, CONSTRAIN performs poorly in diversity indicators, as it overlooks a significant portion of the dialogues' information. Unfortunately, the aforementioned methods, while effectively enhancing LLMs by reducing dialogue history length, fall short of fundamentally improving the models' ability to comprehend extended contexts. Position debiasing methods primarily focus on eliminating position bias but do not address the underlying issue of causal perception in LLMs, resulting in suboptimal performance. RPP, destructing the dialogue structure during fine-tuning, seriously impairs the model's perceptual ability, leading to the poorest performance in word overlap metrics.

Human Evaluation. Overall, the results of manual and automated assessments align consistently. All methods achieved similar performance in fluency metrics, showcasing the remarkable capabilities of LLMs. The raw and finetuned models underperform in relevance and informativeness metrics due to position bias. CPD enhances both relevance and information content indicators, achieving an average improvement of 5% and 7%, respectively, compared to the state-of-the-art baselines. The significant improvement underscores the efficacy of our method in effectively mitigating position bias in LLMs for specific tasks through causal perception finetuning.

4.6 Ablation Study

To demonstrate the effectiveness of our proposed method, we establish three variants: (1) w/o IRM: remove the invariant risk minimization loss (Equation 7); (2) w/o MTE: remove the maximizing treatment effects loss (Equation 8); (3) w/o

Dialogue	. . . Assistant : Hi, do have any issues that you would like to share with me today? User : **I am in disputed mod with my friends** Assistant : I am sorry to hear that. So, you and your friends are not seeing eye to eye? Do you mind telling me what the issue is that you are not agreeing on? . . . User : **They need more money from me, but i couldn't give that much amount** . . . User : Sure, it is my new way of relationship to me with others Assistant : Yes, you're growing and trying to do your best and wanting to expand your boundaries and make them clear. You have taken the right approach and are on the best track to having a successful relationship in the future. Great Job! Thank you for sharing your story with me today. User : **What kind of people do I dispute with?** Assistant :
Qwen-14B-chat	
Raw	It's important to remember that everyone has different opinions and values which may differ from yours. This type of behavior is very common among toxic relationships where one person takes advantage of another's kindness and generosity without reciprocating. If this continues, it will likely damage both parties involved emotionally and mentally over time. Is there anything else I can assist you with?
Finetune	It sounds like you may be disputing with those who are taking advantage of you. This type of behavior can happen when someone does not respect our limits and tries to push their own agenda onto another person. If this continues please reach out to local authorities if needed. Is there anything else I can assist you with?
RSM	**You have a disputation with someone who has a financial confication.** It is advisable to cultivate the skill of setting aside conflicts and exploring more rational solutions. I hope to be of assistance to you.
CONSTRAIN	**You disputed with your friends.**
RPP	That sounds like something which would make anyone worry or be frightened by . Do you have any family around who could support you if worse came to worst?
ZOE	**It sounds like you may just disagree with your friends'** values and beliefs which causes conflict within your relationships.
CPD	**It seems you are in dispute with friends** who have not been responsive to your personal problems and **have asked for more money than you can provide.** These situations can be challenging, but it's important to communicate openly and set boundaries in your relationships. Remember, it's okay to say no when you're uncomfortable with a request.

Table 3: Case study of Qwen-14B-chat. Relevant utterances are marked in boldface.

sampling: remove the positional difference sampling strategy. The experimental results are shown in Table 2.

Experimental results indicate that all ablation variants experienced varying degrees of performance degradation, demonstrating the effectiveness of the components in our method. Furthermore, invariant risk minimization loss can effectively make the model's generation towards gold responses, while maximizing treatment effects loss enhances the diversity of model generation responses. The performance of removing sampling strategy is extremely degraded, proving that it effectively alleviates position imbalance.

4.7 Case Study

Table 3 illustrates an example. The existence of position bias in both the raw and finetuned LLMs leads to generating safe and general but context-irrelevant responses. RSM maintains part of memory capabilities through iterative summarization. CONSTRAIN, despite finding a partially correct answer through retrieval, suffers from substantial information loss in the dialogue, leading to a notable decrease in response diversity. RPP disrupts the dialogue structure by altering sentence-level positions during fine-tuning, impairs the model's ability to identify causally relevant utterances, and results in generating context-irrelevant responses. ZOE eliminates part of position bias of LLMs, but the model's causal perception ability is still weak and unable to identify all relevant utterances for response generation. In contrast, CPD benefits from targeted causal perception fine-tuning, effectively perceiving the correct answer and providing informative answers while considering the entire dialogue history.

5 Conclusion

In this paper, we analyze the deleterious impact effects of position bias in LLMs on long-term dialogue tasks from a causal perspective. To solve the problem, we propose a model-free Causal Perception long-term Dialogue framework (CPD). We extract causally relevant utterances and mitigate position bias through causally perturbed fine-tuning. Specifically, we propose local-position awareness by localizing position information within utterances and further combining it with a perturbation-based method to extract causally relevant utterances. We also propose a causal perception fine-tuning strategy that guides models to focus on causal invariant variables by differently perturbing causally relevant and non-causally relevant utterances in dialogues. A positional difference sampling strategy is employed to address positional imbalances in datasets while maintaining the temporal structure of dialogues. Experiment results demonstrate the effectiveness of our method in alleviating position bias, resulting in informative and human-like response generation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

[Abraham *et al.*, 2022] Eldar D Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems*, 35:17582–17596, 2022.

[Alaa and Van Der Schaar, 2019] Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pages 191–201. PMLR, 2019.

[Amor *et al.*, 2023] Mehdi Ben Amor, Michael Granitzer, and Jelena Mitrović. Technical report on token position bias in transformers, 2023.

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[Austin, 2011] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[Chang *et al.*, 2020] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *International Conference on Machine Learning*, pages 1448–1458. PMLR, 2020.

[Dawid, 1980] A Philip Dawid. Conditional independence for statistical operations. *The Annals of Statistics*, 8(3):598–617, 1980.

[Dettmers *et al.*, 2022] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.

[Feng *et al.*, 2023] Tao Feng, Lizhen Qu, and Gholamreza Haffari. Less is more: Mitigate spurious correlations for open-domain dialogue response generation models by causal discovery. *Transactions of the Association for Computational Linguistics*, 11:511–530, 2023.

[He *et al.*, 2023] Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Yuxin Liang, Hao Wang, Qianguo Sun, Songxin Zhang, Zejian Xie, et al. Never lost in the middle: Improving large language models via attention strengthening question answering. *arXiv e-prints*, pages arXiv–2311, 2023.

[Horgan, 1995] J. Horgan. From complexity to perplexity. *Scientific American*, 272(6):104–109, 1995.

[Hu *et al.*, 2021] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[Ivgi *et al.*, 2023] Maor Ivgi, Uri Shaham, and Jonathan Berant. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299, 2023.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Ko *et al.*, 2020] Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. Look at the first sentence: Position bias in question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1109–1121. Association for Computational Linguistics, 2020.

[Krishna and Murty, 1999] K Krishna and M Narasimha Murty. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439, 1999.

[Lee *et al.*, 2023] Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. Prompted llms as chatbot modules for long open-domain conversation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4536–4554. Association for Computational Linguistics, 2023.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics, 2016.

[Li *et al.*, 2023] Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in large language model based evaluators. *arXiv preprint arXiv:2310.01432*, 2023.

[Lin and Och, 2004] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004.

[Liu *et al.*, 2021] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, 2021.

[Liu *et al.*, 2022a] Jiayi Liu, Wei Wei, Zhixuan Chu, Xing Gao, Ji Zhang, Tan Yan, and Yulin Kang. Incorporating causal analysis into diversified and logical response generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 378–388, 2022.

[Liu *et al.*, 2022b] Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. Improving personality consistency in conversation by persona extending. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1350–1359, 2022.

[Liu *et al.*, 2023a] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

[Liu *et al.*, 2023b] Yijin Liu, Xianfeng Zeng, Fandong Meng, and Jie Zhou. Instruction position matters in sequence generation with large language models. *arXiv preprint arXiv:2308.12097*, 2023.

[Liu *et al.*, 2024] Zhongkun Liu, Zheng Chen, Mengqi Zhang, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. Zero-shot position debiasing for large language models. *arXiv preprint arXiv:2401.01218*, 2024.

[Lu *et al.*, 2023] Zhenyi Lu, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Dangyang Chen, and Jixiong Chen. Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5933–5957, 2023.

[Madhavan *et al.*, 2023] Rahul Madhavan, Rishabh Garg, Kahini Wadhawan, and Sameep Mehta. Cfl: Causally fair language models through token-level attribute controlled generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11344–11358, 2023.

[Marshall *et al.*, 2005] William L Marshall, Tony Ward, Ruth E Mann, Heather Moulden, Yolanda M Fernandez, Geris Serran, and Liam E Marshall. Working positively with sexual offenders: Maximizing the effectiveness of treatment. *Journal of interpersonal violence*, 20(9):1096–1114, 2005.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002.

[Pearl *et al.*, 2016] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[Peysakhovich and Lerer, 2023] Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.

[Rafetseder *et al.*, 2013] Eva Rafetseder, Maria Schwitalla, and Josef Perner. Counterfactual reasoning: From childhood to adulthood. *Journal of experimental child psychology*, 114(3):389–404, 2013.

[Shanmugam, 2001] Ram Shanmugam. Causality: Models, reasoning, and inference : Judea pearl; cambridge university press, cambridge, uk, 2000, pp 384, ISBN 0-521-77362-8. *Neurocomputing*, 41(1-4):189–190, 2001.

[Wang *et al.*, 2021] Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. On position embeddings in BERT. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Wang *et al.*, 2023] Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. Recursively summarizing enables long-term dialogue memory in large language models. *arXiv preprint arXiv:2308.15022*, 2023.

[Wei *et al.*, 2021] Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, Yuchong Hu, and Shanshan Feng. Target-guided emotion-aware chat machine. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–24, 2021.

[Xu *et al.*, 2022a] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5180–5197. Association for Computational Linguistics, 2022.

[Xu *et al.*, 2022b] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, 2022.