Counterfactual Decoding for Anti-Hallucination Knowledge-grounded Dialogue Generation

Anonymous ACL submission

Abstract

The task of Knowledge-grounded Dialogue (KGD) generation, which intentionally invokes external knowledge resources to produce natural and informative responses, has been a popular topic these years. Empowered by the large-scale pretrained language models, existing methods have demonstrated impressive performance on this task. However, the hallucination problem remains a serious problem, causing unpredictable factual errors in the generated results. Although serious efforts try to alleviate this phenomenon by data pre-processing or fact-checking, these methods still heavily rely on assistance from external tools or resources. Inspired by counterfactual reasoning, we propose a lightweight and independent anti-hallucination mechanism in KGD by conducting a causal effect analysis. Our example implementation's benchmark and human evaluation results show that our method can significantly reduce hallucination without disrupting the model performance. We hope our efforts can call for more attention to utilizing causal inference to solve relevant issues.

1 Introduction

Developing a dialogue system that can communicate with human beings naturally and informatively has always been the goal of the generation of researchers (Weizenbaum, 1966; Colby et al., 1971). The task of Knowledge-grounded Dialogue (KGD) Generation, which aims at grounding the model-generated dialogues on various knowledge sources (Ghazvininejad et al., 2018; Dinan et al., 2018), was proposed with this expectation. Despite the excellent generation ability on fluency and grammaticality, the newly proposed, pretrained language model-based KGD methods (Lian et al., 2019; Xu et al., 2021; Zheng and Huang, 2021) are found to suffer from a severe problem of knowledge hallucination (Honovich et al., 2021; Huang et al., 2021), *i.e.*, models may produce seemingly plausible responses with factual errors. This jeopardizing phenomenon spawns a series of studies on its causes (Ishii et al., 2022), evaluation metrics (Dziri et al., 2021b; Honovich et al., 2021) and mitigation solutions (Shen et al., 2021; Rashkin et al., 2021). According to these investigations, an established opinion is that *one major source of hallucination is the (mis)-use of wrong knowledge to describe factual content* (Kryściński et al., 2020; Dziri et al., 2021a), especially the ill-matched knowledge or the irrelevant contents that may activate parametric knowledge bias of the pretrained models[1] (Longpre et al., 2021; Ishii et al., 2022).

Therefore, one straightforward solution to mitigate the hallucination in KGD is to ensure the quality of the induced knowledge via data pre-processing, *e.g.*, Shen et al. propose a measurement based on seven attributes of the dialogue quality, filtering out the untrustworthy samples. However, such methods heavily rely on the performance of the data cleaning tools and are not so flexible when adapting to new data. Another popular solution is adding fact-checking components in the generation model, such as knowledge path encoding (Shuster et al., 2021) or retriever-in-the-loop training (Dziri et al., 2021a), which also requires the assistance of high-quality external knowledgeable resources. Indeed, how to concisely, independently, and adaptively mitigate hallucination remains an important challenge for the KGD community.

In this paper, we invoke the insights from counterfactual reasoning and causal inference (Pearl et al., 2000; Pearl, 2001) to investigate this question. After formulating the structural causal model of KGD, we analyze that the knowledge resources can be purified by maximizing the natural indirect effect (NIE) of dialogue history on the generation result. Based on the analysis, we propose a counterfactual decoding mechanism to reduce hallucina-

[1] We reproduce and conduct an analysis on these hallucination phenomenon of different causes in Appendix.

tion in KGD generating process without leveraging external knowledge sources. As illustrated in Figure 1, we introduce two scenarios, conventional KGD and counterfactual KGD, to estimate the NIE calculation, which can be defined as follows:

Conventional KGD: *What will the response R be, if machine is given the historical dialogue D, and external knowledge K?*

Counterfactual KGD: *What R would the machine say, if it only knows K, but is not in a chat as D?*

The assumed counterfactual KGD depicts the scenario where dialogue D is unavailable. In this case, we can estimate how the external knowledge itself activates the generative model. By subtracting it from the conventional KGD, we naturally utilize the dialogue history to eliminate the irrelevant or even mismatched knowledge information from the generation. This method only modifies the model inference phase, requires no external resources, and may be appropriately implemented in different generative models.

For evaluation, we employ a large-scale Chinese KGD benchmark KdConv (Zhou et al., 2020b) to estimate our implementation instance, along with fine-grained hallucination human evaluations as (Dziri et al., 2021b). We also conduct online tests to measure the usability of the modified model. Experimental results show this approach can mitigate the hallucination without significantly compromising dialogue quality and generation efficiency while adapting to different models.

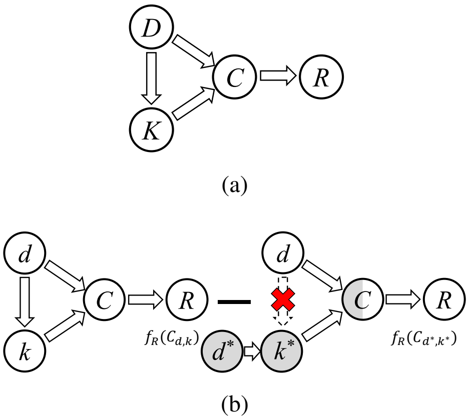
The main contribution of this paper is threefold: 1) an investigation of the task of Knowledge-grounded Dialogue (KGD) from a casual view; 2) a proposal of a counterfactual decoding mechanism for anti-hallucination in this task; 3) an implementation example and its detailed experimental analysis for further exploration in KGD task.

2 Problem Formulation

We begin with formulating the KGD task and introducing its several fundamental concepts.

Dialogue History is a set of conversational utterances between two speakers, formally denoted as $\mathcal{D} = \{U_1, S_1, ..., U_{t-1}, S_{t-1}, U_t\}$, where U_i and S_i are sentences made of words, belonging to the user and the dialogue system respectively. Specially, U_t from the user is also called the *Query*.

External Knowledge contains multiple pieces of information associated with the dialogue topics, which is denoted as $\mathcal{K} = \{k_i\}_{i=1}^{|\mathcal{K}|}$, where k_i is a piece of knowledge information.



(a)

(b)

Figure 1: (a) Casual graph for KGD; (b) Comparison between conventional KGD (left) and counterfactual KGD (right). White nodes are at the value $D = d$ and $K = k$ while gray ones are at $D = d^*$ and $K = k^*$.

Knowledge-grounded Dialogue (KGD) Generation can be formally defined as: given the dialogue history $\mathcal{D}$, the target of this task is to generate a response $\mathcal{R}$ for the t-th round query U_t with the help of the external knowledge pool $\mathcal{K}$.

3 Structural Causal Model of KGD

In this section, we formally construct the structural causal model of the Knowledge-grounded Dialogue Generation (KGD) task, which provides insights into our counterfactual anti-hallucination decoding strategy from a causal view.

Structural Causal Model (SCM) is a fundamental tool of causal inference, which formally describes the interactions among the causal features of a certain task (Pearl et al., 2000; Pearl, 2001; Robins, 2003). To explore the causal story behind the hallucination phenomenon of this task, we construct its SCM in two stages: 1) formulating the causal graph according to the general KGD paradigm; 2) analyzing the potential shortcut of existing methods via causal effect analysis for proposing solution.

Note that in following presentation, we use capitalization for certain random variables (*e.g.*, D for dialogue history $\mathcal{D}$), and lowercase letters for values taken by them (*e.g.*, the value $D = d$).

3.1 Causal Graph

Causal Graph (Pearl et al., 2000) is a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, which indicates how a set of variables $\mathcal{N}$ interact with each others through the causal relational links $\mathcal{E}$, *e.g.*, $X \rightarrow Y$ indicates that variable Y can only obtain value by its prerequisite variable X.

The causal graph of the task of KGD is illustrated in Figure 1a, which is highly general and applicable to a variety of models (Zhou et al., 2018; Ghazvininejad et al., 2018; Gopalakrishnan et al., 2019). Utilizing the language of nodes and links, we study the existing model formulations as:

Node D (Input Dialogue History). As introduced in Section 2, the historical dialogue utterances are basic input features of the KGD task.

Node K (Relevant External Knowledge). The beneficial and relevant knowledge selected from all the external knowledge pool $\mathcal{K} = \{k_i\}_{i=1}^{|\mathcal{K}|}$. Such knowledge is leveraged on adding information to the dialogue models (Zheng and Huang, 2021).

Link $D \rightarrow K$ (Knowledge Selector). This stage is commonly defined as *Knowledge Selection*, which retrieves the most appropriate external knowledge based on the current dialogue state as:

$$K := f_K(D) \tag{1}$$

where node K is causally formed by D, and function f_K refers to the knowledge selection stage.

Node C (Dialogue Context). The final input of the response generation model, which is formed by the combination of dialogue history and selected knowledge via techniques such as encoding integration (Ghazvininejad et al., 2018) and attentional embedding concatenation (Zhou et al., 2018).

Link $D \rightarrow C$ and $K \rightarrow C$ (Contextual Feature Builder). The procedure of the joint modeling of the feature of dialogue D and knowledge K to construct dialogue context feature C as:

$$C := f_C(D, K) \tag{2}$$

where C contains the direct causal effect from both dialogue history and given knowledge, and function f_C refers to the dialogue context modeling stage.

Node R (Response). The feature of the response content corresponds to the current dialogue query, which is the target output of the whole task.

Link $C \rightarrow R$ (Response Generator). The procedure of generating the response according to input context, mostly performs as a decoding process of dialogue models (Zhou et al., 2018) or pretrained language models (Bao et al., 2021).

$$R := f_R(C) \tag{3}$$

where function f_R refers to the generative process, and response R is caused by the input context C.

3.2 Causal Effect Analysis.

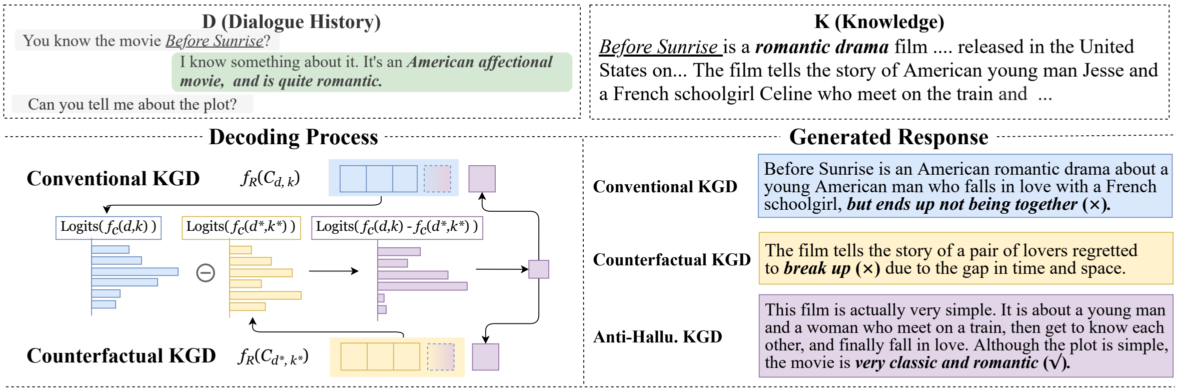
The variables of the KGD task naturally follow the above causal dependencies during the whole training and inference processes. However, the conventional KGD models can only observe the output generation results R of the entire graph when given the input D and K, lacking the understanding of how these elements affect the final generated response. Fortunately, causal inference (Pearl and Mackenzie, 2018) provides analytical tools for opening black-box models, with which we can directly manipulate the values of several nodes and conduct effect analysis. Next, we formally introduce these causal tools and then employ them to analyze the shortcut of the current KGD paradigm.

Counterfactual and ***do*-operation.** *Counterfactual* means "contrary to the facts" (Roese, 1997), which is proposed to analyze the role of variables by assigning them hypothetical values. Such thinking is grounded with the approach ***do*-operation**, denoted as $\boldsymbol{do}(\cdot)$. It cuts off the in-coming links of a variable and requires it to take a certain dummy value, *e.g.*, $do(D = d^*)$ in Figure 1b represents that dialogue is set to a hypothetical value d^*, and its post variable K is correspondingly adjusted[2].

Causal Effect. This term reflects the comparisons between two potential outcomes of the same individual variable from factual and counterfactual scenarios (Rubin, 1978; Robins, 1986), *i.e.*, "*What if I do ..., compared with I had done ...*", and thereby conduct a causal-aware mediation analysis. Note that the causal effect contains a variety of types (Pearl, 2001), and they are selected appropriately by researchers according to the purpose.

Along with these causal approaches, we analyze how to alleviate the hallucination and lift the quality of the generation results. As discussed in Section 1, it is proven that one of the causes of the hallucination phenomenon in KGD is the noises from the supplement external knowledge K, especially when it is *ill-matched knowledge* or it contains several *dialogue-irrelevant contents*. Therefore, it is crucial to filter out irrelevant effects introduced by K to the dialogue D as much as possible. Such an issue can be resolved by considering the dialogue D's Natural Indirect Effect (NIE) on generation result R, which isolates the information flow of $D \rightarrow K \rightarrow C \rightarrow R$ from the total effect and weakens the effect caused by $K \rightarrow C \rightarrow R$

[2] As knowledge node K is only caused by dialogue node D, we take $K := f_K(d^*) = k^*$ for clear presentation.



You know the movie Before Sunrise?

I know something about it. It's an ***American affectional movie, and is quite romantic.***

Can you tell me about the plot?

Before Sunrise is a ***romantic drama*** film .... released in the United States on... The film tells the story of American young man Jesse and a French schoolgirl Celine who meet on the train and ...

Before Sunrise is an American romantic drama about a young American man who falls in love with a French schoolgirl, ***but ends up not being together* (×).**

The film tells the story of a pair of lovers regretted to ***break up* (×)** due to the gap in time and space.

This film is actually very simple. It is about a young man and a woman who meet on a train, then get to know each other, and finally fall in love. Although the plot is simple, the movie is ***very classic and romantic* (√).**

Figure 2: An illustration of counterfactual decoding. Anti-Hallu. is abbreviated from Anti-Hallucination. The hallucination in KGD (such as *Break up*) is probably caused by knowledge bias in pretrained models (*e.g.,* relevant movies in the corpus are mostly bad-ending). This mechanism encourages to improve the NIE of the dialogue D.

alone. This NIE can be formally calculated by following steps: 1) building the counterfactual scenario where the knowledge $K = k^*$ selection and modeling process is conducted without dialogue D, *i.e.*, $d^* = \texttt{Null}$; 2) removing the impact of this dialogue-irrelevant knowledge parts based on the original results, denoted as:

$$\text{NIE} = f_R(C_{d,k}) - f_R(C_{d^*,k^*}) \tag{4}$$

where $C_{d,k}$ and C_{d^*,k^*} correspond to $f_c(d,k)$ and $f_c(d^*,k^*)$. The first term is from the original graph and the second one is from the counterfactual, as illustrated in Figure 1b.

In summary, after formulating the Structural Causal Model of KGD, we invoke the Natural Indirect Effect (NIE) to enhance the information flow $D \to K \to C \to R$ for lifting the generation quality. Such analysis encourages us to conduct an appropriate subtraction with a counterfactual scenario ($D = d^*$), which provides enlightening insights for the upcoming technical exploration.

4 Implementation

Since the optimization direction has been analyzed in a causal view, how to pragmatically implement such NIE in improving the existing KGD paradigm arises as a new challenge. In this section, we propose a simple but effective approach that performs the required subtraction operation of NIE during the model decoding process.

Figure 2 shows a generation sample of our anti-hallucination decoding mechanism, where the original generation process is synchronized with the counterfactual ($D = d^*$), and the subtraction is conducted on the token searching stage. This solution preserves some natural advantages: 1) *Tuning-free*: this method only works on the inference stage, requiring no additional labeled data or attribute model training; 2) *High-Efficiency*: the parallel processes can be performed simultaneously in one batch for avoiding a significant increase of running time. There are several major technical details during our implementation, including:

Knowledge Processing. Calculating above NIE contains two components. $f_R(C_{d,k})$ is from the original generation process while the $f_R(C_{d^*,k^*})$ is under counterfactual setting, where dialogue history D is empty. To maintain the comparability of the two processes, we convert the external knowledge into question-answer pairs, which helps the counterfactual process generate an appropriate dialogue-like response.

Bi-decoding Strategy. Let us denote $R = r$ as the final result of the generated response, where $r = [w_i]_1^{|r|}$ and $w_i \in \mathbb{V}$ is the i^{th} token of it, given the language vocabulary $\mathbb{V}$. For language models (Brown et al., 2020), when given a probability distribution p, the widely-employed approach is to generate text by maximizing the conditional probability. In our decoding process, the $i + 1^{th}$ token w_{i+1} of response is searched as follows:

$$g_i = C_{d,k} \parallel w_{[1:i]} \tag{5}$$

$$\hat{g}_i = C_{d^*,k^*} \parallel w_{[1:i]} \tag{6}$$

$$w_{i+1} = \text{argmax}\left(p\left(g_i\right) - \lambda(i) \cdot p\left(\hat{g}_i\right)\right) \in \mathbb{V} \tag{7}$$

where $\parallel$ is the operation of string concatenation, $w_{[1:i]}$ is the previous generated response tokens at step i. Note that the two processes share the

Table 1: The statistics of the human evaluation protocol, where Avg-length represents the average of utterance length. Hallucination specifically refers to the hallucination annotation on the KdConv generated results.

Annotation	Participants	Label	Avg-length
Hallucination	100	75,000	12.8
Human-Chatbot	50/50	18,000	9.6

generated tokens. The decay function $\lambda(i)$ is empirically designed to prevent the generation ability from being counteracted, as two processes' inputs are increasingly convergent during decoding.

5 Experiment

5.1 Experimental Setting

Previous studies (Liu et al., 2016; Chen et al., 2017) have empirically revealed that the correlation between automatic metrics and human judgments is weak in evaluating dialogue systems. Therefore, except for automatic benchmark evaluation, we conduct manual human evaluation when analyzing detailed method performance. Moreover, we also perform fine-grained hallucination annotation according to (Dziri et al., 2021b) and host online tests to analyze the practical performance of our method further. The experimental settings include:

Model Setup. Our implementation mainly employs the Chinese-versioned GLM with 10B parameters (Du et al., 2022) as the backbone dialogue generation model, which is trained on 302GB raw Chinese data collected from multiple Chinese websites. And the 2.9B pretrained Transformer-XL (CTXL) (Dai et al., 2019) is also utilized for comparison. The decay function is implemented as $\lambda(i) = \alpha^{i-1}$, where α is set to 0.3. The model is deployed with a server of 8 Nvidia V100 GPUs, Intel CPU cores and 376GB Memory. For online tests, we employ the Xlore2 (Jin et al., 2019) as the external knowledge base.

Dataset and Baselines. As our implementation is in Chinese, we employ KdConv (Zhou et al., 2020b), a large Chinese knowledge-grounded dialogue benchmark instead of the widely-used English WoW dataset (Dinan et al., 2018). KdConv preserves over 4.5K conversations and 86K utterances from three topics: *Film*, *Music*, *Travel*.

We reproduce several recent typical pretrained models on this benchmark for comparison, including pretrained general language models and dialogue-specific models. The selected baselines include: • **CPM-2** (Zhang et al., 2021) is the first GPT-architecture Chinese language model, and we employ its 11B version dense model and adapt it to KGD task; • **CDial-GPT** (Wang et al., 2020) is a pretrained dialogue model, trained on LCCC conversation corpus with 95.5M parameters; • **EVA** (Zhou et al., 2021) is an excellent 2.8B parameter Chinese dialogue model based on WDC-dialogue dataset; • **PLATO-KAG**[3]: Huang et al. proposes a joint modeling strategy for KGD task, upon the PLATO-XL (Bao et al., 2021).

Evaluation Metrics. As for the KdConv benchmark, we select its several original metrics for generative dialogue models: (1) **BLEU-1/2/4** (Papineni et al., 2002) is a famous metric to compute the k-gram overlap between a generated sentence and a reference (Sordoni et al., 2015). (2) **Distinct-1/2/4** (Li et al., 2016) is designed to evaluates the diversity of generated responses. Since these metrics only considers the quality of dialogue, we conduct hallucination-aware annotation upon benchmark generation results as (3) **Hallucination**, which includes three subcategories: **Factual**, **Spurious** and **Generic**, corresponding to the ratio of the scenario where the generated content contains factual, spurious or no knowledgeable information.

We also follow existing dialogue researches (Bao et al., 2021; Huang et al., 2021) to conduct human-chatbot dialogue evaluation, which empirically employs utterance-level metrics as: *Coherence*: whether the response is relevant and consistent with the context, *Informativeness*: whether the response is informative, *Inconsistency*↓: whether the response has conflicts with the dialogue context, *Hallucination*↓: whether the response contains any factual errors. Note that the Coherence and Informativeness scale is $[0, 1, 2]$, whose higher score indicates a better performance. Meanwhile, the scale of Inconsistency and Hallucination is $[0, 1]$, whose lower score indicates a better performance.

Human Evaluation Protocol. We recruit 100 people, mostly Chinese university students, to complete our human evaluations: hallucination annotation and human-chatbot dialogue evaluation.

In hallucination annotation, the utterances from different methods are mixed, and each one is la-

[3] https://github.com/sserdoubleh/plato/tree/develop/projects/PLATO-KAG. But the publicly available version is only for English datasets WoW (Dinan et al., 2018) and Holl-E (Moghe et al., 2018). Therefore we translate the KdConv dataset into English and conduct adaptation on it. The results is for reference only.

Method	BLEU			Distinct			Hallucination(%)		
	Metric-1	Metric-2	Metric-4	Metric-1	Metric-2	Metric-4	Factual	Generic	Spurious
CPM	**29.27**	**12.81**	3.99	4.57	29.53	73.25	28.57	30.71	40.71
CDial-GPT	25.06	5.50	0.54	3.49	18.99	54.19	3.60	**60.43**	35.97
EVA	21.46	5.03	0.22	3.80	18.94	44.86	11.35	**64.54**	24.11
PLATO-KAG[3]	28.08	7.87	1.92	3.77	28.01	71.30	10.34	8.28	**81.38**
CTXL w/o AH	25.81	9.56	3.56	3.97	37.07	86.90	23.86	26.40	49.75
CTXL w/ AH	23.77	8.27	2.96	3.93	**37.09**	**87.30**	**36.04**	33.50	**30.45**
GLM w/o AH	23.11	9.98	**4.40**	3.63	31.59	78.94	46.11	20.21	38.95
GLM w/ AH	**24.81**	**10.28**	**4.29**	2.46	25.62	73.36	**61.65**	21.02	**17.33**

Table 2: The overall performance of comparison methods on KdConv. Hallucination metrics are the ratios of factual, generic, and spurious utterances. "w/o AH" and "w/ AH" represents the setting of without and with our anti-hallucination decoding mechanism. The values in bold are emphasized for discussion.

beled by 3 annotators under a double-blind setting, and the label is confirmed by majority voting. Besides, we conduct human-chatbot dialogue evaluation, and this annotation team is divided into two 50-member groups. One group is required to chat with required models for generating real-world dialogues, while the other group aims to triply annotate the utterances produced by the prior one. To ensure the fairness of the evaluation, the topics and initial rounds of conversation are uniformly restricted. We preserve sessions over 5 rounds. Table 1 presents the statistics of our human evaluation.

5.2 Overall Performance

We first analyze the overall performance of our method on the KdConv benchmark. Except for the previously given results, we reproduce the baseline models via Prompting (Liu et al., 2021), and conduct hallucination annotation upon the generation results. Each method generates five responses on the given dialogue session, and Table 2 presents their average values after the significance test. In general, our proposed method can effectively alleviate the hallucination in KGD without significantly degrading the performance of other metrics. Based on these results, our primary observations include:

First, the hallucination of the generated results is indeed reduced after counterfactual decoding, whether employing GLM or CTXL as the backbone. Meanwhile, this mechanism can preserve the original features of the models (*e.g.,* CTXL still performs better at Distinct), which indicates its pretty good adaptability on pretrained models.

Second, the counterfactual KGD decoding is not determined to significantly drop dialogue quality. The experimental results show that the model performance fluctuates only slightly on different metrics, considering the instability of the automatic evaluation of the generated model (Post, 2018). GLM performs better after counterfactual decoding on BLEU, and CXTL performs better on Distinct, suggesting that our adjustments do not inevitably result in specific adverse effects.

Third, our approach tends to affect almost only the hallucination compared with the performance differences brought by the models themselves. Meanwhile, we find that the dialogue models (CDial-GPT, EVA) prefer to generate low-information content (Generic) to prevent mistakes, while our approach not only avoids errors (Spurious), but also prefers factual content, which is a phenomenon worth further investigation.

5.3 Result Analysis

The overall evaluation raises several issues to be further investigated, especially how the hallucination is refined on generated responses. Therefore, we conduct detailed observations that includes:

Fine-grained Annotation. Based on the generation results, we follow (Dziri et al., 2021b) to conduct fine-grained hallucination annotation, which requires each utterance to be annotated to a taxonomy, including *Entailment*, *Generic*, *Off-Topic* and *Contradiction*. According to the discussed future directions in the paper, we add *Experience*, *Erudite* and *Uncooperative* labels. Note that *Entailment* and *Erudite* are both factual. The former is explicitly supported by the given knowledge, while the latter is not directly supported but actually correct. *Uncooperative* and *Experience* are highly related to the dialogue scenario. The details of these metrics and corresponding examples are shown in Table 3.

The comparison results shown in Figure 3 reveal some details of this anti-hallucination mech-

Category	Example	
Entailment	Evidence: Previous turn: Response:	Mervyn LeRoy (1900-1987) was an American director and producer. Do you know about Mervyn LeRoy? You know, he is a famous *American director and producer*. In Chinese terms, it should be regarded as an old-timer of an old artist.
Erudite	Evidence: Previous turn: Response:	Hayao Miyazaki was born in Tokyo, Japan. That's right, he has two jobs, do you know him(Miyazaki)? Of course, he is a master of animation, with many masterpieces, *such as "Princess Mononoke", "Spirited Away"...*
Experience	Evidence: Previous turn: Response:	Stephen Chow's main achievement is Best Director at the 21st Hong Kong Film Awards. Wow, Stephen Chow has taken over all the positions in the film industry, it's awesome. Well, it certainly goes without saying, *I often watch his movies.*
Generic	Evidence: Previous turn: Response:	The Wedding Banquet (1993 American film directed by Ang Lee) starring Zhao Wenxuan. Do you know Zhao Wenxuan? *Yeah, I know.*
Uncooperative	Evidence: Previous turn: Response:	Sean Penn won his first best actor award at the 46th Berlin Film Festival for his prison film "Dead Man Walk" in 1996. He has so many awards, do you know what his first award work is? *Yes, he has acted in many excellent films, I believe you can find him in many films, do you know him?*
Off-topic	Evidence: Previous turn: Response:	Brad Pitt is an American film actor. Yes, do you know where he is from? *Are you abroad now?*
Contradiction	Evidence: Previous turn: Response:	Quentin Tarantino's representative work is Inglourious Basterds. What was his first movie? The first film he directed was Pulp Fiction, *but there is no second one*.

Table 3: Examples of each of the categories of our fine-grained hallucination, which are slightly refined from (Dziri et al., 2021b). Labels are annotated manually. Note that Entailment, and Erudite have a positive effect on dialogue, Off-topic, Contradiction, and Uncooperative are negative, and Experience and Generic are kinds of neutral.

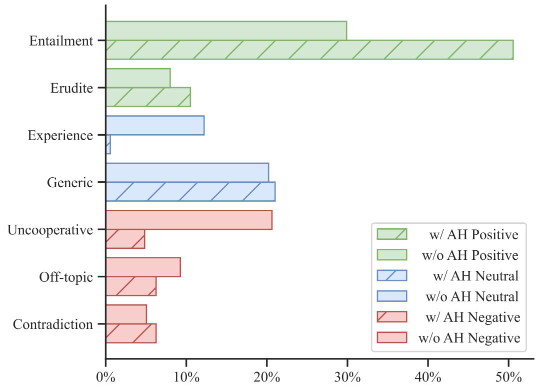


Figure 3: Comparison of the distribution of fine-grained hallucination labels before and after the implementation of the anti-hallucination mechanism. The horizontal axis represents the percentage of hallucination labels.

Metric	w/o AH	w/ AH	Δ (%)
Coherence	1.803	1.795	-0.4
Inconsistency↓	0.137	**0.095**	**-30.6**
Informativeness	1.741	1.721	-1.1
Hallucination↓	0.226	**0.191**	**-15.4**

Table 4: The results of human-chatbot evaluation. The final results presented in the table are the average value of the GLM utterances after filtering out invalid labels.

anism. (1) The proportion of generated responses supported by the given knowledge (*Entailment*) significantly increases, while some of them might be the previous *Off-topic* responses. A plausible explanation is that counterfactual decoding effectively makes the model concentrate on the relevant parts of K, as we discussed in Section 3.2. (2) *Uncooperative* responses surprisingly decrease, which demonstrates that the refined model tends to generate more helpful content related to the conversation. The natural indirect effect (NIE) of D is strengthened. (3) *Erudite* results are stable, and this is because our approach is not working by introducing external knowledgeable fact-checking.

Human-Chatbot Evaluation. As introduced in Section 5.1, we conduct a human evaluation to verify the effect of anti-hallucination decoding on dialogue quality from another perspective. According to the evaluation results in Table 4, we find that after counterfactual decoding, the Coherence and Informativeness keep a high performance

Method	Round		Time	
	Avg	Max	Avg	Max
EVA	**33.8**	34	619.9	715.4
GLM(w/o AH)	30.5	**40**	**675.2**	744.5
GLM(w/ AH)	**32.3**	**38**	624.6	**775.2**

Table 5: The comparison results of online A/B Test. Avg and Max correspond to the average and maximum value after screening outliers.

(with a slight decline of 0.4% and 1.1%), while Inconsistency↓ and Hallucination↓ drop significantly (30.6% and 15.4% correspondingly). This result corroborates with our primary observations in the overall performance, proving that our approach can indeed alleviate the hallucination without destroying the model performance.

Online A/B Test. We deploy the baseline model EVA, pure GLM, and our anti-hallucination GLM in an online environment to collect feedback from users. This evaluation follows the setting of the double-blind A/B test (Man et al., 1999). Table 5 shows the collected statistics of results of over 300 participants from May 18th, 2022 to May 25th, 2022. We observe that the GLM (w/ AH) is competitive in both conversation rounds and duration of conversations, especially it outperforms the other two approaches regarding the longest duration.

6 Related Work

Knowledge-grounded Dialogue. Origin from open-domain dialogue (Ma et al., 2020), the task of knowledge-grounded dialogue (KGD) generation aims to generate more informative responses with the help of various external knowledge (Ghazvininejad et al., 2018), such as knowledge graphs (Dinan et al., 2018) and persona descriptions (Zhang et al., 2018). With the prosperity of pretrained language models (PLMs) (Brown et al., 2020; Du et al., 2022; Bao et al., 2021), recent KGD efforts preferably try to employ prompting (Zheng and Huang, 2021) or adapters (Xu et al., 2021) for exploiting these models in this task. As the features of PLMs are still being investigated (Liu et al., 2021), a fine-grained KGD result analysis is also essential (Roller et al., 2020). Meanwhile, how to fairly evaluate KGD models (Moghe et al., 2018; Dinan et al., 2018; Zhou et al., 2020b) remains a matter to be confirmed.

Hallucination in NLG. Factual errors in generative models have been spotted for some time (Holtzman et al., 2019), and this topic has become conspicuous along with the rising of PLMs (Petroni et al., 2019). The problem of *hallucination* (also termed as *unfaithfulness*, *factual consistency*, etc.) widely exists in NLG tasks including dialogue (Santhanam et al., 2021), summarization (Pagnoni et al., 2021), translation (Zhou et al., 2020a) and data2text (Wiseman et al., 2017). Whatever the intrinsic or extrinsic hallucination, their causes are commonly summarized as *heuristic data collection*, *innate divergence* and *knowledge bias*, inspiring the mitigation methods to conduct data filtering and fact-checking for alleviation. Considering the requirement of external sources of these methods, we hope to exploit the dialogue itself in anti-hallucination KGD.

Causal Inference. Counterfactual thinking and causal inference (Pearl et al., 2000; Pearl and Mackenzie, 2018) have inspired several studies in artificial intelligence. Besides its contributions on Computer Vision (Goyal et al., 2019; Niu et al., 2021), its application on linguistic tasks (Zmigrod et al., 2019) is also a vigorous topic, especially data augmentation (Chen et al., 2021) and model explanation (Wu et al., 2021). Since there is no uniform guidance on causal methods in relevant fields, we hope that our efforts can call for more attention to such insights in NLG tasks.

7 Conclusion and Future Work

In this paper, we propose a lightweight counterfactual decoding mechanism for alleviating hallucination in Knowledge-grounded Dialogue generation. After constructing the causal graph of KGD task, we invoke the natural indirect effect (NIE) to enhance the dialogue quality by utilizing the conversation itself. To evaluate the proposed method, we conduct both benchmark and human evaluations on the example implementation, and experimental results indicate that our counterfactual approach can effectively reduce hallucination while maintaining the good performance of other metrics. Beyond the satisfactory outcomes, the detailed experiments provide several promising directions in the future.

First, it is necessary to further investigate and discuss the fine-grained hallucination phenomenon of KGD task, which may result in a more human-like and friendly dialogue system. Second, constructing a widely acceptable, automatic hallucination evaluation metric is still emergent. Third, we sincerely hope our attempts can inspire more efforts in boosting NLG tasks via counterfactual thinking.

Ethic Consideration

For years the press has been arguing the use of AI and its pros and cons. One advance could be used in various ways and thus lead to different outcomes. In this supplement, we first take a brief on how might our method be used in both good and bad ways, then move on to the discussions about how to appropriately deploy our approach, along with the labor right consideration during our experiment.

Potential Risk Analysis. Chatbot systems, which are one of the destinations of artificial intelligence, have always attracted debate on their ethic issues (Følstad and Brandtzæg, 2017). Once a chatbot is equipped with sufficient knowledge, people are likely to rely more on what it expresses. Therefore, a controllable, explainable, and faithful chatbot system with excellent mitigation of hallucination is an emergent issue to be solved. However, it is worth noting that even employing our anti-hallucination mechanism, the system still occasionally generates factual errors. We argue that any implemented dialogue system with our techniques needs to prominently mark that the system generated this content and prevent the trouble of being mistaken for a trusted human. For critical areas such as healthcare and education, the use of our technology needs to be further licensed to prevent hidden risks.

Annotator Right Consideration. For each recruited annotator, destroy the privacy of the individuals in their conversational logs after evaluation, and we obtain permission from each person for other data to be disclosed. To serve their annotation within reasonable working hours, we developed a platform that allows them to terminate, pause, continue and modify the annotation results at any time. Compensation for each marker is negotiated to meet local salary standards, and the whole progress is supervised by the institution's Ethics and Academic Committee.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, et al. 2021. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *arXiv preprint arXiv:2109.09519*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hao Chen, Rui Xia, and Jianfei Yu. 2021. Reinforced counterfactual data augmentation for dual sentiment classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 269–278.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Nouha Dziri, Andrea Madotto, Osmar R Zaiane, and Avishek Joey Bose. 2021a. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021b. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071*.

Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of hci. *interactions*, 24(4):38–42.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*.

Xinxian Huang, Huang He, Siqi Bao, Fan Wang, Hua Wu, and Haifeng Wang. 2021. Plato-kag: Unsupervised knowledge-grounded conversation via joint modeling. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 143–154.

Yejin Bang Ishii, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Comput. Surv*, 1(1).

Hailong Jin, Chengjiang Li, Jing Zhang, Lei Hou, Juanzi Li, and Peng Zhang. 2019. Xlore2: large-scale cross-lingual knowledge graph construction and application. *Data Intelligence*, 1(1):77–98.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI International Joint Conference on Artificial Intelligence*, page 5081.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Longxuan Ma, Wei-Nan Zhang, Mingda Li, and Ting Liu. 2020. A survey of document grounded dialogue systems (dgds). *arXiv e-prints*, pages arXiv–2004.

Daniel Man, Boris Man, and Harvey Plosker. 1999. A double-blind study. *Plastic & Reconstructive surgery*, 104(7):2261–2266.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2322–2332.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 411–420.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19:2.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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 704–718.

James Robins. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

James M Robins. 2003. Semantics of causal dag models and the identification of direct and indirect effects. *Oxford Statistical Science Series*, pages 70–82.

Neal J Roese. 1997. Counterfactual thinking. *Psychological bulletin*, 121(1):133.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Donald B Rubin. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*.

Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1598–1608.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2021. Retrieval-free knowledge-grounded dialogue response generation with adapters. *arXiv preprint arXiv:2105.06232*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, et al. 2021. Cpm-2: Large-scale cost-effective pre-trained language models. *AI Open*, 2:216–224.

Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020a. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593*.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, et al. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020b. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *arXiv preprint arXiv:2004.04100*.

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

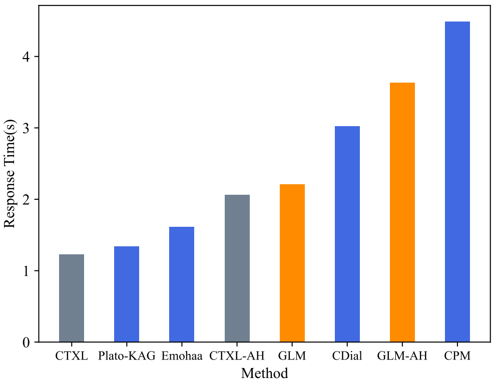
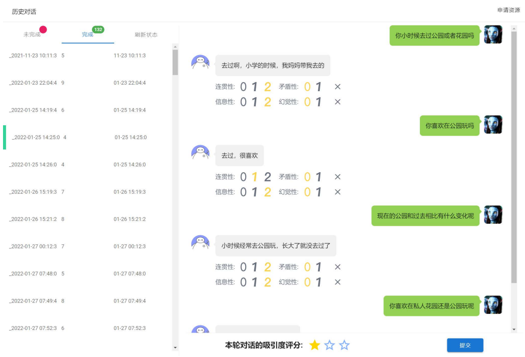


Figure 4: Comparison of the average response time.



_2021-11-23 10:11:3 5 11-23 10:11:3
_2022-01-23 22:04:4 9 01-23 22:04:4
_2022-01-25 14:19:4 6 01-25 14:19:4
_2022-01-25 14:25:0 4 01-25 14:25:0
_2022-01-25 14:26:0 4 01-25 14:26:0
_2022-01-26 15:19:3 7 01-26 15:19:3
_2022-01-26 15:21:2 8 01-26 15:21:2
_2022-01-27 00:12:3 7 01-27 00:12:3
_2022-01-27 07:48:0 5 01-27 07:48:0
_2022-01-27 07:49:4 8 01-27 07:49:4
_2022-01-27 07:52:3 6 01-27 07:52:3

你小时候去过公园或者花园吗

去过啊，小学的时候，我妈妈带我去的

连贯性: 0 1 2 矛盾性: 0 1 ×
信息性: 0 1 2 幻觉性: 0 1 ×

你喜欢在公园玩吗

去过，很喜欢

连贯性: 0 1 2 矛盾性: 0 1 ×
信息性: 0 1 2 幻觉性: 0 1 ×

现在的公园和过去相比有什么变化呢

小时候经常去公园玩，长大了就没去过了

连贯性: 0 1 2 矛盾性: 0 1 ×
信息性: 0 1 2 幻觉性: 0 1 ×

你喜欢在私人花园还是公园玩呢

本轮对话的吸引度评分: ★ ☆ ☆

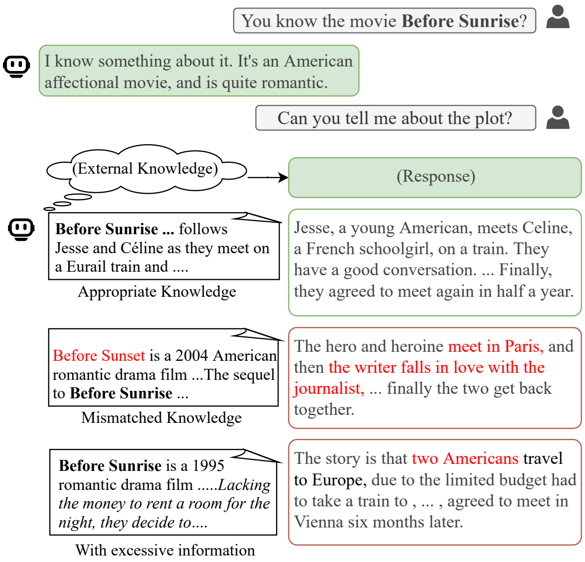
提交

Figure 5: A screenshot of our human evaluation platform, where annotators are grouped for double-blinded evaluation, *i.e.*, not accessible to the source of the assigned conversations.

A Supplement of Experimental Result

A.1 Inference Efficiency

We detect the time efficiency of our method by comparing the average speed with baselines. As shown in Figure 4, the counterfactual KGD on both GLM and CTXL maintain a pretty good generation speed regarding the response time, because its two decoding processes can be simultaneously performed in a batch. Compared with the efficiency gap due to the backbone model scales, such adjustment only slightly increases the latency and does not affect the usability of generation models.



You know the movie **Before Sunrise**?

I know something about it. It's an American affectional movie, and is quite romantic.

Can you tell me about the plot?

Before Sunrise ... follows Jesse and Céline as they meet on a Eurail train and ....

Appropriate Knowledge

Jesse, a young American, meets Celine, a French schoolgirl, on a train. They have a good conversation. ... Finally, they agreed to meet again in half a year.

Before Sunset is a 2004 American romantic drama film ...The sequel to **Before Sunrise** ...

Mismatched Knowledge

The hero and heroine meet in Paris, and then the writer falls in love with the journalist, ... finally the two get back together.

Before Sunrise is a 1995 romantic drama film .....*Lacking the money to rent a room for the night, they decide to....*

With excessive information

The story is that two Americans travel to Europe, due to the limited budget had to take a train to , ... , agreed to meet in Vienna six months later.

Figure 6: The reproduced hallucination examples of GLM (w/o AH). The hallucination occurs even though the provided knowledge content is not erroneous.

A.2 Effect of Decay Function λ

We conduct qualitative analysis on the effect of the decay function as shown in Figure 6. An interesting finding is that the strength of the decay is not monotonically correlated with the quality of the generation; rather, it depends on other factors. From observation, the model performs better with more substantial decay when the knowledge content is highly relevant to the conversation and vice versa. This observation inspires us that we should subsequently optimize our mechanism by exploring better mathematical models to correlate the decay function with knowledge relevance to achieve better KGD generation.

A.3 Case Study: The Cause of Hallucination

We reproduce the previously observed hallucination phenomena (Ishii et al., 2022). Although the mismatched knowledge increases the risk of generating errors, some excessive information surprisingly tends to cause imperceptible hallucination. The hallucination problem cannot be thoroughly solved by data-level fact-checking since the pre-trained models initially hold some short-cutting information and preserve quite a plenty of parameterized knowledge bias.

A.4 Fine-grained Annotation of Baselines

We also conduct fine-grained hallucination annotation on the results of other baselines, which is shown in Figure 8. According to the results, we

Setting	Example	
$\alpha = 0.7$	Evidence:	If someone says the sentence that *You are my lobster* to you, it means he likes you very much, just like Ross to Rachel in Friends.
	Previous turn:	A: What is there to eat in Changsha? B: Stinky tofu, fried dough sticks and spicy *lobster*. A: How is stinky tofu made?
	Response:	First fry the tofu yellow, then add the chili, then add some soy sauce, and finally sprinkle with chopped green onion.
$\alpha = 0.3$	Evidence:	If someone says the sentence that *You are my lobster* to you, it means he likes you very much, just like Ross to Rachel in Friends.
	Previous turn:	A: What is there to eat in Changsha? B: Stinky tofu, fried dough sticks and spicy *lobster*. A: How is stinky tofu made?
	Response:	It means, I like you, you are my lobster.
$\alpha = 0.9$	Evidence:	Miao Yanyu, a native of Ma'an mountain, climbed to the top of *Makalu, the fifth highest mountain in the world*, located on the border between China and Nepal
	Previous turn:	A: Do you often go hiking, especially those famous peaks? B: Yeah, I know a lot about it. A: What is *the fifth highest mountain in the world?*
	Response:	The fifth highest mountain in the world is Makalu, and the altitude of it is 8463 meters.
$\alpha = 0.1$	Evidence:	Miao Yanyu, a native of Ma'an mountain, climbed to the top of *Makalu, the fifth highest mountain in the world*, located on the border between China and Nepal
	Previous turn:	A: Do you often go hiking, especially those famous peaks? B: Yeah, I know a lot about it. A: What is *the fifth highest mountain in the world?*
	Response:	Makalu, located on the border between China and Nepal.

Table 6: Examples of the different α settings. Note that we only implement an example of the decay function λ. An intuitive observation is the effect of decay function can not be independently considered.

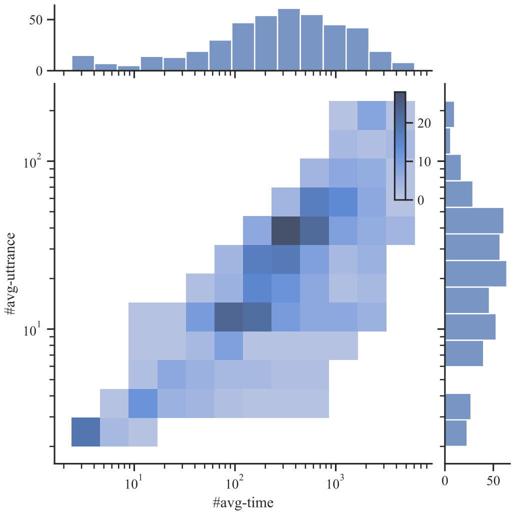


Figure 7: The user involvement distribution of online A/B test of GLM (w/ AH).

observe that even when there are factual errors, the model rarely generates content that conflicts with the given knowledge (Contradiction). This implies that rather than the noise of the input text itself, the deeper cause of the hallucination is more due to parametric knowledge errors in the pretrained models, which are inspired by the input content.

A.5 Human Evaluation Details

The human-chatbot evaluation is conducted in two stages. For the generation stage, we deploy the dialogue service on the WeChat app with the toolkit WeChaty[4]. And for the annotation stage, we develop an online platform, as shown in Figure 5. To prevent annotation bias, the evaluators are divided into groups of three members, and each group is assigned the same dialogues. Meanwhile, the scores given by any annotators are invisible to others. All scoring items can be modified and withdrawn before submission. Annotators can log in and log out, change their labels for already completed problems, or continue evaluation from their current positions freely during the mission period.

A.6 User Involvement Analysis

We summarize the user involvement statistics of the deployed GLM (w/ AH) dialogue service. As shown in Figure 7, the participants' activities follows the normal distribution. It is worth noting that over 52.1% of users chat with our chatbot for more than 10 rounds. Most participants engage in $80 - 90$ rounds of conversation, and these conver-

[4] https://github.com/wechaty/wechaty

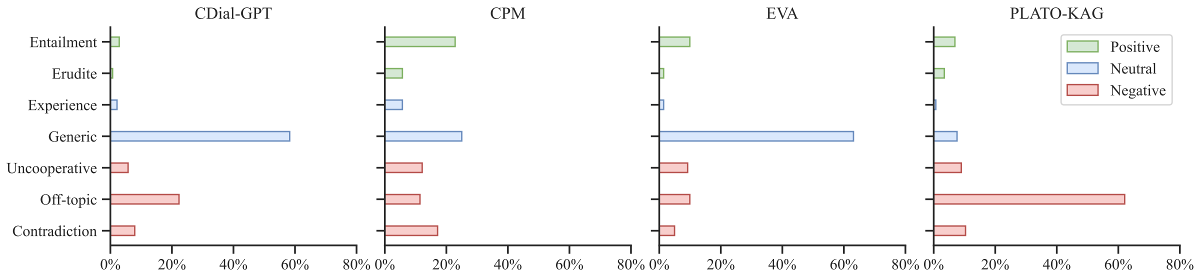


Figure 8: The fine-grained hallucination annotation results of baseline models.

sations last more than 6.3 minutes. This indicates that the quality of conversations is stable and that the anti-hallucination model can maintain a competitive performance in double-blind online tests.