

A Cause-Effect Look at Alleviating Hallucination of Knowledge-grounded Dialogue Generation

Jifan Yu[‡] Xiaohan Zhang[♣] Yifan Xu^{‡♯} Xuanyu Lei^{‡♯}
Zijun Yao^{‡♯} Jing Zhang[♣] Lei Hou^{‡♯} Juanzi Li^{‡♯†}
[‡]Institute of Education; [♯]BNRist; [♯]KIRC, Institute for Artificial Intelligence
Tsinghua University, Beijing 100084, China
[♣]Zhipu.AI; [♣]Renmin University
{yujifan,houlei,lijuanzi}@tsinghua.edu.cn

Abstract

Empowered by the large-scale pretrained language models, existing dialogue systems have demonstrated impressive performance conducting fluent and natural-sounding conversations. However, they are still plagued by the *hallucination* problem, causing unpredictable factual errors in the generated responses. Recently, knowledge-grounded dialogue generation models, that intentionally invoke external knowledge resources to more informative responses, are also proven to be effective in reducing hallucination. Following the idea of getting high-quality knowledge, a few efforts have achieved pretty good performance on this issue. As some inevitable knowledge noises may also lead to hallucinations, it is emergent to investigate the reason and future directions for building noise-tolerant methods in KGD tasks. In this paper, we analyze the causal story behind this problem with counterfactual reasoning methods. Based on the causal effect analysis, we propose a possible solution for alleviating the hallucination in KGD by exploiting the dialogue-knowledge interaction. Experimental results of our example implementation show that this method can reduce hallucination without disrupting other dialogue performance, while keeping adaptive to different generation models. We hope our efforts can support and call for more attention to developing lightweight techniques towards robust and trusty dialogue systems.

Keywords: Causal Inference, Language Model Decoding, Knowledge Grounded Dialogue

1. Introduction

The conversation is an exchange of sentiments and ideas between two or more speakers, from which they grow common ground and subsequently sustain long-term relationships (Clark et al., 2019). If a response tends to be unfaithful, *i.e.*, by producing content that conflicts with facts or previous dialogue, it may run the risk of jeopardizing the entire experience of conversation (Dziri et al., 2021a). Unfortunately, even the modern dialogue systems, which are powered by pretrained language models (Brown et al., 2020; Du et al., 2022), still can not fully avoid generating responses with self-contradiction or even factual errors. This phenomenon, as the example shown in Figure 1, has been generally summarized as the issue of *Knowledge Hallucination* (Dziri et al., 2021b; Ishii et al., 2022), which gradually becomes an emergent challenge toward a reliable, robust AI dialogue system (Marcus, 2020).

In recent years, a series of researches have explored that leveraging external knowledge can help lift the informativeness as well as the factual accuracy (Dinan et al., 2018; Huang et al., 2021) of the generated responses. This line of techniques, known as knowledge-grounded dialogue (KGD) generation (Zhou et al., 2020b; Zheng and Huang,

2021), generally retrieve the most relevant knowledge with current dialogue and generate an appropriate response based on it. To further alleviate the hallucination problem, one of the straightforward ideas is to guarantee the quality and matching accuracy of the selected external knowledge. Benefited from this thinking, methods such as augmenting knowledge with web retrievers (Shuster et al., 2021) and conducting refinement by re-querying knowledge bases (Dziri et al., 2021a) have achieved pretty good performance in this task.

Despite these advanced methods that attempt to guarantee the quality of retrieved knowledge, there remain some inevitable knowledge noises in many scenarios. First, real-world knowledge sources still face the challenge of incompleteness and incorrectness (Wang et al., 2021), which occasionally leads to mismatching issues for KGD models. Second, a correct knowledge piece may preserve redundant contents that can possibly cause subsequent hallucinations. As the example illustrated in Figure 1, once the retrieved knowledge is not so relevant to dialogue query, or even contains some extra information, there is a potential risk that the *parametric knowledge bias* (Ishii et al., 2022) of generation model is activated, which eventually leads to hallucinations in generated responses.

Except for the pettish demand on stronger knowledge collection techniques, we hope to explore

[†] Corresponding author.

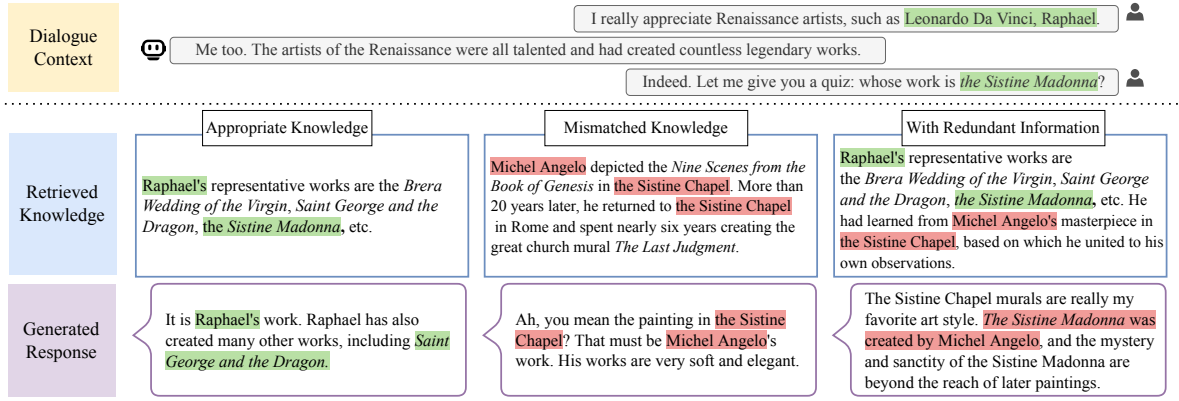


Figure 1: An example of the hallucination of knowledge-grounded dialogue. The figure presents three scenarios where the retrieved knowledge is perfectly matched, mismatched and with some redundant information. The texts with shading are dialogue-relevant knowledge. Green ones means the accurate contents. Red ones are the possible causes and effects of hallucination.

whether there is another direction for alleviating the hallucination of KGD with noise-tolerant methods with the dialogue itself. 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 how the generated responses are causally affected by the input elements. As Figure 2 shows, we introduce two scenarios, conventional KGD and counterfactual KGD, which can be defined as follows:

Conventional KGD: *What will the response R be, if machine is given the 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 the direct effect of dialogue D is reduced. 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 interaction between dialogue and the external knowledge to eliminate the redundant or even mismatched knowledge information from the generation, which enhances the total direct effect of D on responses. This method only modifies the model inference phase, requiring no training stage, thus may be appropriately implemented in different generative models. Based on this analysis, we propose a possible solution with a counterfactual dual-decoding mechanism to alleviate hallucination without training and associated extra data for annotation.

We conduct an example implementation in Chinese language models to primarily validate our suspicions. For evaluation, we conduct a series of human and automatic evaluation on a public knowledge-grounded dialogue dataset (Zhou et al.,

2020b) to measure the features of the proposed method in several dialogue metrics. Experimental results prove our idea can indeed mitigate the hallucination without significantly compromising dialogue quality and generation efficiency while keeping able to adapt to different generation models.

The main contribution of this paper is threefold. First, to our best knowledge, our counterfactual reasoning framework is the first to formulate knowledge bias in KGD as causal effects. Second, we provide a causality-based interpretation for recent anti-hallucination KGD works. Third, we propose an alternative direction with counterfactual decoding, which is suitable for different generation models.

2. Structural Causal Model of KGD

How are the generated responses affected by the input elements? Can we employ the existing information for further exploiting external knowledge? In this section, we formally construct the structural causal model of the Knowledge-grounded Dialogue Generation (KGD) task to answer such questions, which also provides insights into the directions of alleviating the knowledge hallucination 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 of KGD task, thereby clarifying the information flow of the general paradigm of it; 2) detecting the potential shortcut of conventional solutions with causal effect analysis.

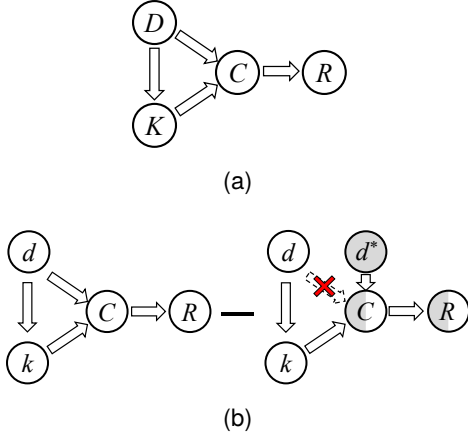


Figure 2: (a) Casual graph for KGD; (b) Comparison between conventional KGD (left) and counterfactual KGD (right). White nodes are at the value $D = d$ and $K = k$ while gray one is $D = d^*$. Node C , R are counterfactual.

2.1. Task Formulation as 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 2a, which is highly general and applicable to a variety of KGD 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). Dialogue History is a set of conversational utterances between two speakers, formally denoted as $D = \{U_1, S_1, \dots, U_{t-1}, S_{t-1}, U_t\}$, where U_i and S_i are sentences made of words, belonging to the user and the dialogue system respectively. Especially, U_t from the user is also called the *Query*.

Node K (Retrieved External Knowledge). External knowledge pool contains multiple pieces of information associated with the dialogue topics, which is denoted as $K = \{k_i\}_{i=1}^{|\mathcal{K}|}$, where k_i is a piece of knowledge information. Such knowledge is injected into the dialogue models (Zheng and Huang, 2021) to add extra information. For most of the existing KGD methods, only the most relevant and beneficial knowledge is needed for subsequent processes.

Link $D \rightarrow K$ (Knowledge Selector). This stage is commonly defined as *Knowledge Selection* (Zheng and Huang, 2021), which retrieves the most appropriate external knowledge based on the current dialogue state as:

$$K := f_K(D) \quad (1)$$

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

Node C (Dialogue Context). This is an intermediate variable of KGD. The dialogue context C is the direct input of the response generation module, which is formed by the combination of dialogue history and selected knowledge via techniques such as encoding integration (Ghazvininejad et al., 2018) or 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:

$$C := f_C(D, K) \quad (2)$$

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 response R corresponds to the content for answering the current dialogue query U_t , which is the target output of the whole KGD 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),

$$R := f_R(C) \quad (3)$$

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

Therefore, Knowledge-grounded Dialogue (KGD) Generation can be formally defined as: given the dialogue history D , the target of this task is to generate a response R for the t -th round query U_t with the help of the external knowledge K . Once the generated R contains information that conflicts with facts, it is considered to be a **hallucinated** response (Dziri et al., 2021a).

2.2. 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),

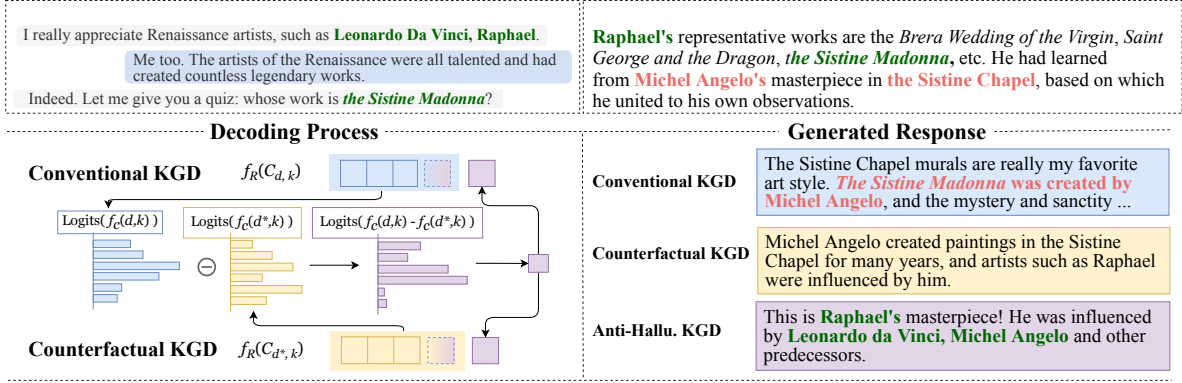


Figure 3: An illustration of counterfactual decoding. Anti-Hallu. is abbreviated from Anti-Hallucination. This mechanism encourages to improve the TDE of the dialogue D . The final generated response is more effected by dialogue contexts.

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 $do(\cdot)$. 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 2b represents that dialogue is set to a hypothetical value d^* . We use the following notations to translate causal graphs into formulas:

$$\begin{aligned}
 C_{d^*, K_{d^*}} &= f_C(do(D = d^*), f_K(do(d = d^*))) \\
 C_{d, k} &= f_C(d, f_K(d)) \\
 C_{d^*, k} &= f_C(do(D = d^*), f_K(d))
 \end{aligned} \tag{4}$$

Causal Effect. This term is broadly defined as a contrast between two *potential outcomes* of the same individual variable under two different *treatments*. It usually involves the comparison between 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 causal effect contains a variety of types (Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2013), and they are selected appropriately by researchers according to the purpose (Nguyen et al., 2021).

In the SCM for KGD (as shown in Figure 2a), dialog history D serves as the treatment, while knowledge K plays the role of the mediator, with the response R as the potential outcome. The *total effect* (TE) of $D = d$ on the response R compares between two situations with and without the dialog history:

$$TE = f_R(C_{d, k}) - f_R(C_{d^*, K_{d^*}}) \tag{5}$$

As suggested by Robins and Greenland (1992), TE can be further decomposed into *total direct effect* (TDE), which captures the effect of D to the response R flows directly via $D \rightarrow C \rightarrow R$, and *pure indirect effect* (PIE) which depicts the effect

indirectly through the mediator K via $D \rightarrow K \rightarrow C \rightarrow R$. Formally:

$$TE = \underbrace{f_R(C_{d, k}) - f_R(C_{d^*, k})}_{\text{TDE}} + \underbrace{f_R(C_{d^*, k}) - f_R(C_{d^*, K_{d^*}})}_{\text{PIE}}$$

To alleviate the hallucination invoked by mismatched knowledge and knowledge with redundant information, we need to choose a response with an appropriate maximum causal effect considering that hallucination resides in the mediator K . The rationale for choosing a suitable causal effect expresses two principles. 1) To prone possible hallucinated responses that are off the topic of the dialogue history, the causal effect must highlight the influence of the dialog history, avoiding solely over-reliance on the mediate knowledge. 2) To keep a plausible knowledgeable response that contains rich supplemented information, the causal effect must be considered under the presence of the supplemented knowledge and pick up the interaction between K and D .

The above two requirements are exactly captured by the *Total Direct Effect* (TDE). In KGD, it reflects the effect of the dialog history on the response if the knowledge is controlled at the level it would take in the presence of the dialog history. It is worth noting that, Robins and Greenland (1992) point out that TE can also be decomposed into *pure direct effect* (PDE) and *total indirect effect* (TIE). We choose the direct effect rather than the indirect effect because we want to highlight the importance of the dialog history rather than neglecting the dialog history and generate the response with knowledge only. The reason why we choose TDE rather than PDE is that TDE preserves knowledge in the presence of the dialog history to pick up the interaction between the knowledge and the dialog history to facilitate inter-evaluation. We show more details on why other causal effects are not suitable for suppressing hallucination in KGD in the Appendix.

Measuring TDE follows 2 steps: 1) Building the counterfactual scenario where the dialog context C is constructed without dialogue D , *i.e.*, $d^* = \text{Null}$; 2) Removing the impact of this dialogue-irrelevant knowledge parts based on factual results, denoted as: $C_{d,k}$ follow the original dialog generation procedure and $C_{d^*,k}$ is from the counterfactual, as illustrated in Figure 2b.

In summary, after analyzing the causal effect of the KGD, we invoke the Total Direct Effect:

$$\text{TDE} = f_R(C_{d,k}) - f_R(C_{d^*,k}) \quad (6)$$

to allow the model to use dialog history to verify the correctness and usefulness of retrieved knowledge. Such analysis encourages us to conduct an appropriate subtraction with a counterfactual scenario ($D = d^*$), which provides enlightening insights for the upcoming technical exploration.

3. Implementation

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

Figure 3 shows a 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 extra model training; 2) *High-Adaptability*: the parallel processes can be performed with different generative language models. There are several major technical details during our implementation, including:

Knowledge Processing. Calculating above TDE contains two components. $f_R(C_{d,k})$ is from the original generation process while the $f_R(C_{d^*,k})$ is counterfactual. The dialogue history D is artificially set to almost empty as d^* . To maintain the comparability of the two processes, we preserve the query of dialogue as the input, which helps the counterfactual process generate an appropriate dialogue-like response.

Dual-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 $i + 1^{\text{th}}$ token

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

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.

w_{i+1} of response is searched as follows:

$$g_i = C_{d,k} \parallel w_{[1:i]} \quad (7)$$

$$\hat{g}_i = C_{d^*,k} \parallel w_{[1:i]} \quad (8)$$

$$w_{i+1} = \operatorname{argmax}(p(g_i) - \lambda(i) \cdot p(\hat{g}_i)) \in \mathbb{V} \quad (9)$$

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

4. Experiment

4.1. Experimental Setting

In this section, we conduct a series of evaluations to check whether our analysis can indeed help alleviate hallucination. Specifically, we perform fine-grained hallucination annotation according to (Dziri et al., 2021b) and host online tests to analyze the practical performance of our method. Moreover, we also conduct several automatic benchmark evaluations to observe whether our implementation will be reduced significantly in other dialogue metrics.

Baseline Language Models. We utilize 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**¹. Huang et al. propose a joint modeling strategy for the KGD task, upon the

¹<https://github.com/sserdouhleh/plato/>

PLATO-XL (Bao et al., 2021). For online tests, we employ the Xlore2 (Jin et al., 2019) as the external knowledge base. All the methods are deployed with XDAI toolkit (Yu et al., 2022) for adapting to the KGD task.

Testbed Dataset. 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, and Travel.

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.

Human Evaluation Protocol. Previous studies (Liu et al., 2016; Chen et al., 2017) have empirically revealed that human judgments are still the most reliable for evaluating dialogue systems. In this paper, we mainly employ human evaluation for both hallucination annotation and dialogue quality evaluation. We recruit 100 people, mostly university students to build our annotation team. For collecting the user online feedback, we deploy the dialogue service as a public demo with the toolkit WeChaty². Moreover, we develop an online evaluation platform for convenient annotations. Annotators can chat with the online service, label the utterances for different tasks, easily log in and log out, change their labels for already completed problems, or continue evaluation from their current positions freely during the mission period.

To guarantee the annotation quality, the evaluators are divided into groups of three members, and each group is assigned the same dialogues in certain annotation tasks. Meanwhile, the scores given by any annotators are invisible to others. We also ensure fairness by restricting the topics and initial rounds of conversation uniformly during annotation. Table 1 presents the overall statistics of our annotation.

¹[tree/develop/projects/PLATO-KAG](https://github.com/projects/PLATO-KAG). But the public 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 an adaptation on it. The results is for reference only.

²<https://github.com/wechaty/wechaty>

Method	Hallucination (%)		
	Factual↑	Generic	Spurious↓
CPM	28.57	30.71	40.71
CDial-GPT	3.60	60.43	35.97
EVA	11.35	64.54	24.11
PLATO-KAG ¹	10.34	8.28	81.38
CTXL w/o AH	23.86	26.40	49.75
CTXL w/ AH	36.04	33.50	30.45
GLM w/o AH	46.11	20.21	38.95
GLM w/ AH	61.65	21.02	17.33

Table 2: The hallucination annotation results of the comparison methods. The values are the ratios of Factual, Generic, and Spurious utterances. “w/o AH” and “w/ AH” correspondingly represent the settings of without and with our anti-hallucination decoding mechanism.

4.2. Result Analysis

Here we analyze a series of experimental results, which indicates the features of this anti-hallucination idea.

Hallucination Annotation. Each method generates five responses on the given dialogue session. For each generated response, three annotators are required to classify it into three categories: **Factual↑**, **Spurious↓** and **Generic**, corresponding to the ratio of the scenario where the generated content is factual, spurious, or not knowledgeable. In general, users hope a KGD system can generate more factual responses. The system may reply a generic perfunctory, but the responses with spurious information are resolutely resisted. Table 2 presents the ratios of each model’s generated responses in terms of hallucination, where more factual and less spurious labels indicate better performance in alleviating hallucination. The presented values are after the significance test, which provides two main observations:

(1) Our proposed method can indeed help alleviate the hallucination. After adding the method, GLM’s hallucinated responses sharply reduce. One of the most interesting results is that the “Generic” ratio is not significantly lifted, which indicates that the rationale behind the improvement is less about erasing all information and more about enabling the model to activate the correct knowledge.

(2) The counterfactual decoding strategy is able to adapt to different generation models. Whether on the main implementation model GLM or the comparison model CTXL, the hallucinations are effectively alleviated. We believe it is promising to further optimize the performance and adapt this thinking to other advanced language models.

Fine-grained Annotation. To further investi-

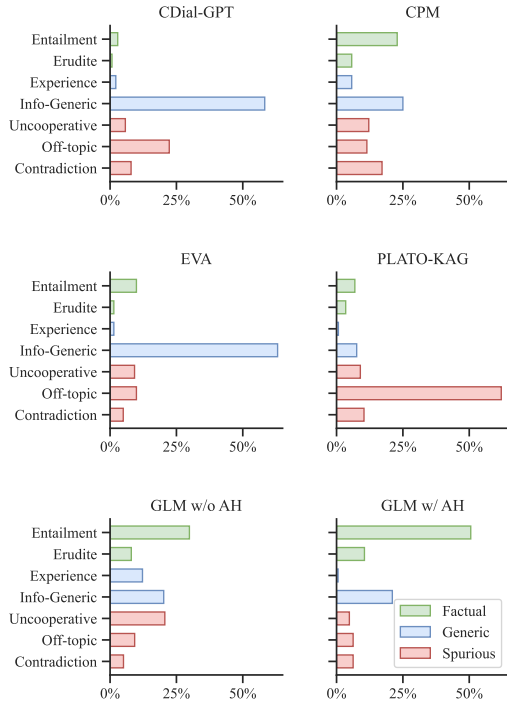


Figure 4: Comparison of the distribution of fine-grained hallucination labels of all comparison methods. The horizontal axis represents the percentage of hallucination labels.

gate how the hallucinations are alleviated, we follow the insights of (Dziri et al., 2021b) to conduct fine-grained hallucination annotation, which requires each utterance to be annotated to a taxonomy³, including *Entailment*, *Info-Generic*, *Off-Topic*, and *Contradiction*. According to the discussed future directions in the paper, we further 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. We require 100 annotators to annotate the utterances based on these tags, and each result is also voted by 3 annotators.

The results shown in Figure 4 reveal some details of this anti-hallucination mechanism. (1) For all methods, the factual responses are more *Entailment* than *Erudite*. The former is explicitly supported by the given knowledge, while the latter is not directly supported but correct. This shows that most of the factual results indeed require a high-quality knowledge context, indicating the reason why existing methods (Shuster et al., 2021) can

³Detailed introduction and corresponding utterance examples of the fine-grained hallucination labels are in Appendix.

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

Table 3: 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.

Setting	Round		Duration(s)	
	Avg	Max	Avg	Max
GLM w/o AH	30.5	34	619.9	715.4
GLM w/ AH	32.3	38	624.6	775.2

Table 4: The statistical results of online A/B Test.

help reduce hallucination. (2) For the comparison within GLM implementation, we find the proportion of 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 makes the model concentrate on the relevant parts of K .

Human-Chatbot Evaluation. To investigate the performance of dialogue quality, we divide the annotation team into two 50-member groups. One group is required to chat with GLM w/ and w/o AH for generating dialogues, while the other group aims to annotate the utterances produced by the prior one. We follow existing dialogue researches (Bao et al., 2021; Huang et al., 2021) to employ metrics. *Coherence*: whether the response is consistent with the context, *Informativeness*: whether the response is informative, *Inconsistency \downarrow* : whether the response has conflicts with the dialogue context, *Hallucination \downarrow* : whether the response contains any factual errors (this metric is generally employed and can also doubly check the performance of the previous annotation). Note that the Coherence and Informativeness scale is $[0, 1, 2]$, whose higher score indicates a better performance while the scale of Inconsistency and Hallucination is $[0, 1]$, whose lower score is better.

According to the evaluation results in Table 3, we find that after counterfactual decoding, the Coherence and Informativeness keep a high performance (with a slight decline of 0.4% and 1.1%), while Inconsistency \downarrow drop significantly (-30.6%). As many previous dialogue context contain factual contents, the alleviating of hallucination can also help reducing the Inconsistency. The drop of Hallucination \downarrow (-15.4%) doubly proves the effectiveness of our strategy on reducing factual errors. These results

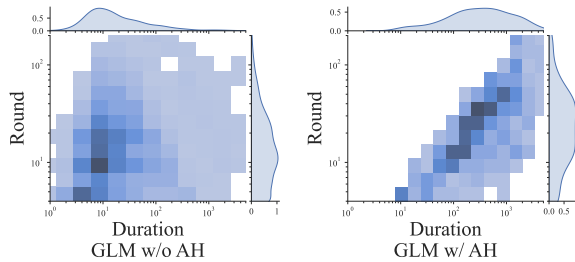


Figure 5: The comparison of the distribution on user involvement distribution of GLM w/ and w/o AH mechanism. We show the number of conversation rounds (y-axis) and the total time involved in the conversation (x-axis).

can further show that our approach can indeed alleviate the hallucination without significantly destroying the model’s other performance.

Online A/B Test & User Involvement Analysis. We deploy the GLM with and without anti-hallucination decoding 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 4 and Figure 5 shows the collected results of over 300 participants from May 18th, 2022 to May 25th, 2022.

From the statistics, GLM w/ AH outperforms the GLM w/o AH both in terms of dialogue round and duration, especially the maximum of these cases. From the distribution view, the distribution of user activities for the model with the AH strategy is more skewed towards sustaining longer conversations. It is worth noting that over 52.1% of users chat with GLM w/ AH for more than 10 rounds. These records indicate that: (1) Our dual-decoding strategy does not harm the generation quality of the dialogue, such as the conversation engagingness and fluency, which is reflected by the high engagement of real users; (2) GLM w/ AH, that is more “faithful” with less hallucinated responses, is more competitive on establishing long-term trust with human beings.

5. Related Work

This work is highly relevant with three mainstream topics. They include:

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 and faithful responses with the help of various external knowledge (Ghazvininejad et al., 2018), such as knowledge graphs (Dinan et al., 2018), knowledge documents (Yao et al., 2023; Yu et al., 2023), and persona descriptions (Zhang et al., 2018). With the prosperity of pretrained large language models

(LLMs) (Brown et al., 2020; Du et al., 2022; Bao et al., 2021), recent KGD efforts preferably try to employ prompting (Zheng and Huang, 2021; Yu et al., 2022) or fine-tuning (Xu et al., 2021; Zhang et al., 2023) for exploiting these models in this task. The problem of hallucination in KGD recently attracts the attention of researchers (Dziri et al., 2021b). However, as the features of LLMs are still being investigated (Liu et al., 2021), a fine-grained KGD result analysis is also essential (Roller et al., 2020) to guide the solution of hallucination.

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 LLMs (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). In this paper, we employ this thinking as a theoretical tool for analyzing existing KGD methods and discussing future directions. 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.

6. Conclusion and Future Work

In this paper, we study the causal story of alleviating hallucination in knowledge-grounded dialogue generation. After constructing the causal graph of the KGD task and analyzing how the generation results are affected by inputs, we invoke the total direct effect (TDE) to enhance the dialogue quality by utilizing the interaction of dialogue history and external knowledge. Based on this finding, we propose an alternative solution for existing models, which employs a dual-decoding strategy during model inference. We conduct a series of experiments on the example implementation, and the results indi-

cate that this counterfactual decoding strategy can effectively reduce hallucination while maintaining the good performance of other metrics.

As a primary exploration, this work preserves some limitations that require future investigations. First, it is emergent to apply this idea to more languages and dialogue models, thereby fully validating its features. Second, building a lightweight and practical method for hallucination in dialogue is still promising, especially employing mechanisms such as human-in-the-loop and continuous learning. Third, we sincerely hope our attempts can inspire more efforts in boosting NLG tasks via counterfactual thinking.

Acknowledgement

This work is supported by Zhipu AI, a grant from the Institute for Guo Qiang, Tsinghua University (2019GQB0003), and National Natural Science Foundation of China (Grant No.62277034).

7. Bibliographical 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.
- Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.
- 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*.
- 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 32-1.
- 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, Ziyang 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*.
- 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.
- 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*.
- 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.
- Gary Marcus. 2020. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.
- 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.
- Trang Quynh Nguyen, Ian Schmid, and Elizabeth A Stuart. 2021. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*, 26(2):255.
- 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*.
- 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: Cambridge University Press*, 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.
- 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.
- James M Robins and Sander Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155.
- Neal J Rouse. 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*.
- 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.
- Tyler J VanderWeele. 2013. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, pages 224–232.
- Meihong Wang, Linling Qiu, and Xiaoli Wang. 2021. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3):485.
- 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.
- 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*.
- Zijun Yao, Yantao Liu, Xin Lv, Shulin Cao, Jifan Yu, Juanzi Li, and Lei Hou. 2023. Korc: Knowledge oriented reading comprehension benchmark for deep text understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11689–11707.
- Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.
- Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, Xinyu Guan, Jing Zhang, Lei Hou, Juanzi Li, and Jie Tang. 2022. Xdai: A tuning-free framework for exploiting pre-trained language models in knowledge grounded dialogue generation. In *Proceedings of the 2022 Conference on Knowledge Discovery and Data Mining (KDD)*.
- Jing Zhang, Xiaokang Zhang, Daniel Zhang-Li, Jifan Yu, Zijun Yao, Zeyao Ma, Yiqi Xu, Hao-hua Wang, Xiaohan Zhang, Nianyi Lin, et al. 2023. Glm-dialog: Noise-tolerant pre-training for knowledge-grounded dialogue generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5564–5575.
- 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, Xiacong 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*.