Towards Fewer Hallucinations in Knowledge-Grounded Dialogue Generation via Augmentative and Contrastive Knowledge-Dialogue

Bin Sun[1], **Yitong Li**[2,3], **Fei Mi**[2], **FanHu Bie**[3], **Yiwei Li**[1], **Kan Li**[1*]
[1]School of Computer Science & Technology, Beijing Institute of Technology
[2] Huawei Noah's Ark Lab [3]Huawei Technologies Ltd.
{binsun,liyiwei,likan}@bit.edu.cn
{liyitong3,mifei2,biefanhu}@huawei.com

Abstract

Existing knowledge-grounded open-domain dialogue generation models often face the hallucination problem, i.e. the dialogue generative model will persist in an inappropriate knowledge and generate responses that inconsistent with the facts. We argue that this problem mainly stems from the polarized optimization objectives and weak knowledge generation ability. To mitigate the hallucination, we take inspiration from human communicating that people will replay euphemistic responses for the unclear or unrecognizable knowledge, and propose an Augmentative and Contrastive Knowledge Dialogue Expansion Framework (ACK-DEF). ACK-DEF constructs the augmentative and contrastive knowledge dialogue samples, which consist of the knowledge of different degrees of errors and the response of manual design, to expand the original training set and smooth the polarized optimization objective that enables models to generate ground-truth with or without gold knowledge. Not only the knowledge, ACK-DEF also provides the tactful responses of manual design corresponding to the incomplete correct knowledge. Experimental results on the Wikipedia of Wizard dataset show that employing the ACK-DEF is effective to alleviate the hallucination problem.

1 Introduction

Recently, Knowledge-Grounded Dialogue Generation draws dramatic attentions in artificial intelligence community. Many efforts incorporate knowledge information to improve the performance of dialogue generation models (Zhou et al., 2018; Dinan et al., 2019; Gopalakrishnan et al., 2019; Kim et al., 2020; Zhao et al., 2020a; Zheng et al., 2021; Zhao et al., 2022a; Bao et al., 2022). However, these methods always face the hallucination problem, that is, the dialogue generation model may insist on an inappropriate knowledge and generate responses that inconsistent with the facts (Rashkin et al., 2021; Zhao et al., 2022a; Dziri et al., 2022).

We argue that the hallucination problem primarily caused by two aspects: (1) The optimization objective is usually polarized by the gold knowledge-dialogue samples and general dialogue samples without knowledge in current knowledge-grounded dialogue datasets (Zhou et al., 2018; Gopalakrishnan et al., 2019; Dinan et al., 2019; Wu et al., 2019; Komeili et al., 2022). Few datasets consider teaching models how to respond when dealing with incomplete correct knowledge, which makes the models tend to believe in the given knowledge, regardless of whether the knowledge is appropriate or not, resulting in hallucination problems. In addition, the knowledge retrieval system tends to extract irrelevant knowledge rather than relevant knowledge when the database is large, aggravating the hallucinations (Reimers and Gurevych, 2021; Liu et al., 2022). (2) The generation of knowledge may also face the hallucination problem and obtain the inappropriate knowledge, leading the generation of hallucination responses (Kim et al., 2020; Zhao et al., 2020a; Liu et al., 2022; Adolphs et al., 2021; Bao et al., 2022).

To mitigate the hallucination problem, we propose an Augmentative and Contrastive Knowledge Dialogue Expansion Framework (ACK-DEF), which is inspirited by human communicating that people will replay euphemistic response for the unrecognizable knowledge. ACK-DEF is proposed to smooth the polarized optimization objective by augmenting training set with augmentative and contrastive knowledge-dialogue samples. Not only the knowledge, we also designed the reply patterns for the knowledge with different level of errors. For this, we propose the augmentative knowledge dialogue expansion (AK), and contrastive knowledge dialogue expansion (CK). AK is proposed to boost the generalization ability of models on knowledge with minor noise. On the contrary, inspired from the *contrastive learning* paradigm (Cai et al., 2020; Chen et al., 2020a,b; Sun et al., 2021, 2022), CK

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1741–1750

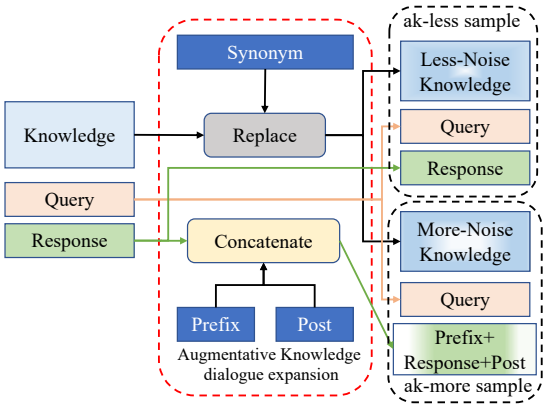

Figure 1: A diagram of our Augmentative Knowledge dialogue expansion method. We replace different proportion of words in the original knowledge with synonyms to construct incomplete correct knowledge, and design response for different knowledge. We also use prompts to guide the dialogue generation process.

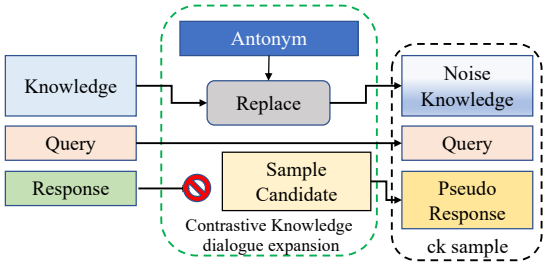

Figure 2: A diagram of our Contrastive Knowledge dialogue expansion method. We use the antonym to reconstruct the knowledge information and design multiple responses for such knowledge. Since antonyms transform the semantics of the original knowledge, the noise knowledge often contains wrong facts. By this, the model can learn a better boundary between correct and incorrect knowledge, and a safety reply pattern for incorrect knowledge.

reconstructs incorrect knowledge and designs euphemistic responses, which aims to push the model learn the reply pattern of incorrect knowledge and a better boundary between correct and incorrect knowledge.

Contributions: We propose an ACK-DEF to construct new knowledge-dialogue samples that consist of knowledge with different level of errors and manual responses, to soften the training optimization objectives of models, which will mitigate the hallucination. Finally, we conduct extension experiments to show the superior performance of ACK-DEF on alleviating the hallucination.

2 Methodology

To mitigate the hallucination problem that caused by the polarized optimization objectives in knowledge grounded dialogue generation, we take inspiration from human communicating, and propose the Augmentative and Contrastive Knowledge Dialogue Expansion Framework (ACK-DEF). Our ACK-DEF aims to soften the polarized training optimization objectives of current knowledge-grounded dialogue generation methods, and guide the dialogue system reply patterns for the knowledge with different level of errors. To achieve this end, we design two effective expansion method, which will be detailed in below.

2.1 Augmentative Knowledge Dialogue

We propose the *Augmentative Knowledge (AK)* dialogue expansion to boost the generalization ability of the dialogue model on the knowledge with similar semantics but different expressions, which can prevent the model from being interfered by partial-relevant knowledge retrieved by the retrieval systems (Lian et al., 2019; Zhao et al., 2020b; Hedayatnia et al., 2020; Zheng et al., 2021; Shuster et al., 2021; Komeili et al., 2022). As shown in Figure 1, we employ the synonym data augmentation tool, which replaces words in the original knowledge with synonyms, to reconstruct the knowledge information (Miller, 1995). Considering that the synonym may disrupt the original semantics of new constructed knowledge, we constrain the replace possibility within [0.1,0.2]. Hence, we can obtain the approximate knowledge. Combining this knowledge and the original dialogue, we obtain the "ak-less sample". In addition, we also replace 30% to 50% words with their synonyms to construct the less similar knowledge. Inspired from *prompt learning* paradigm (Yao et al., 2022; Valvoda et al., 2022; Zhao et al., 2022b), we manually produce some Prefix-prompts and Post-prompts (see Appendix) to (1) make the new response more tactful for the less similar knowledge; (2) regulate and guide the dialogue generation process of the model. We call the sample consist of less-similar knowledge and designed response as "ak-more sample".

2.2 Contrastive Knowledge Dialogue

We propose the *Contrastive Knowledge (CK)* dialogue expansion, inspired from the *contrastive learning* paradigm (Chen et al., 2020b; Cai et al., 2020), not only construct the incorrect knowledge as negative samples for original knowledge, but also build the euphemistic responses as positive

samples for the original response with incorrect knowledge.[1] To help the model learn a boundary between correct and incorrect knowledge, we employ the antonym to make up new incorrect knowledge. For example, given the knowledge "*nintendo was founded on 23 september 1889 ...*", the "founded" will be replaced with "abolish", which greatly changes the semantics but little changes the expression. After that, we random choose an euphemistic response to replace the original response of the dialogue. Finally, The incorrect knowledge and the replaced euphemistic response are combined as the "ck-sample".

3 Experiment and Results

3.1 Experiment Settings

3.1.1 Dataset

We use the Wikipedia of Wizard (WoW) data, a well-established knowledge-grounded open-domain dialogue dataset, for our experiment. We pre-precess the WoW dataset and extract the single-turn knowledge dialogue samples. To evaluate the performance of our method in detail, we perform four test sets: `normal`, `ak-less`, `ak-more` and `ck`. The `normal` set is the original test set. And the `ak-less`, `ak-more` and `ck` are the sets consist of ak-less, ak-more and ck samples, respectively. We also follow the settings of WoW data and divide the test set into two groups (seen test and unseen test): the topic of the knowledge in the unseen test set is missing in the training set.

3.1.2 Baseline

We employ the released PLATO-v1 (Bao et al., 2020) model, a pre-trained dialogue generation model based on UniLM, for our experiment.

Fine-tuning We directly finetune a model on the original WoW training set. By this, the model can only see gold knowledge dialogue samples and general dialogue samples without knowledge. Hence, we call the fine-tuned model PLATO+GOLD.

Fine-tuning with ACK-DEF We finetune the model with the original set and the expansion samples that obtained through ACK-DEF. Thence, we call it PLATO+ACK-DEF.

[1]We manually construct some responses, please see Appendix for the detail.

3.1.3 AutoEvaluation Metrics

Dialogue Metrics Our primary metrics of interest are Distinct-n (Li et al., 2016), Response Length (Len.) (Csaky et al., 2019), BLEU (Papineni et al., 2002), Embedding-based (Greedy (GRE), Average (AVG), Extrema (EXT)) (Liu et al., 2016), and Coherence (COH) (Xu et al., 2018). Distinct-n evaluates the diversity of generated responses, which is calculated through the ratio of distinct n-grams and all generated n-grams. Len. is the average number of words of all generated responses. BLEU validates the degree of the word-overlap between the generated response and the ground-truth, which denotes the consistence between generated response and ground-truth. Embedding-based metrics (GRE, AVG and EXT) are introduced to evaluate the semantic relationship of generated responses and ground-truth responses, illustrating the consistence in semantic level. COH. mainly assesses the relevance between contexts and generated responses.

Knowledge Metrics We follow the PLATO(Bao et al., 2020) and use the knowledge precision, recall and f1 scores. These metrics are used to calculate the ratio of tokens that exist in common in ground-truth knowledge and generated responses to tokens in generated responses. "Recall" is the average ratio of the number of overlapping tokens in response and knowledge to the number of tokens in knowledge. And "Precision" is the average ratio of the number of overlapping tokens to the number of tokens in response. In other words, "Recall" indicates how much knowledge information is contained in the response, while "Precision" indicates the proportion of knowledge information in the response. Even we involve the negative and incorrect knowledge in response generation, we still use the ground-truth knowledge to calculate the metrics in Table 3,4.

3.2 Dialogue Performance Analysis

Table 1 and Table 2 report the automatic results on four test sets and four unseen test sets, respectively. In these Tables, it can be observed that (1) the PLATO+ACK-DEF has a competitive performance with PLATO+GOLD on the normal set, which means that the PLATO+ACK-DEF can recognize the golden knowledge and produce appropriate responses. (2) the PLATO+GOLD perform worse than PLATO+ACK-DEF on ak-less, which means that the robustness of the dialogue model only trained with golden knowledge is very weak.

test set	Distinct-1/2/3		Len.	BLEU-1/2/3/4				GRE	AVG	EXT	COH
normal	**0.1068**	**0.4533**	13.69	0.4280	0.2965	**0.2110**	**0.1529**	**0.7392**	**0.8689**	**0.6361**	0.7808
	0.0902	0.3984	**16.20**	**0.4428**	**0.3017**	0.2109	0.1499	0.7366	0.8683	0.6330	**0.7878**
ak-less	**0.1194**	**0.5024**	13.50	0.3861	0.2574	0.1745	0.1192	0.7160	0.8607	0.6148	0.7755
	0.0823	0.3532	**18.78**	**0.4502**	**0.2982**	**0.2015**	**0.1380**	**0.7307**	**0.8696**	**0.6293**	**0.7948**
ak-more	**0.1234**	**0.5174**	12.81	0.1675	0.1062	0.0680	0.0435	0.6908	0.8551	0.5994	0.7706
	0.0675	0.2946	**21.83**	**0.4358**	**0.3001**	**0.2123**	**0.1542**	**0.7745**	**0.9151**	**0.7093**	**0.8098**
ck	**0.1109**	**0.4779**	13.23	0.2965	0.1779	0.1080	0.0657	0.5838	0.7622	0.5373	0.7712
	0.0652	0.2029	**13.36**	**0.4230**	**0.2705**	**0.1809**	**0.1266**	**0.6572**	**0.8306**	**0.6162**	**0.8049**

Table 1: The automatic results of PLATO+GOLD (up) and PLATO+ACK-DEF (down) on four test seen sets.

test set	Distinct-1/2		Len.	BLUE-1/2/3/4				GRE	AVG	EXT	COH
normal	**0.0503**	**0.2422**	12.43	**0.3516**	**0.2331**	**0.1582**	**0.1090**	**0.6988**	**0.8568**	0.6306	0.8094
	0.0467	0.2311	**13.14**	0.3463	0.2281	0.1536	0.1049	0.6968	0.8541	**0.6338**	**0.8105**
ak-less	**0.0966**	**0.3917**	13.39	0.3871	0.2565	0.1724	0.1164	0.7143	0.8600	0.6122	0.7836
	0.0623	0.2664	**19.18**	**0.4443**	**0.2907**	**0.1936**	**0.1301**	**0.7232**	**0.8663**	**0.6194**	**0.8026**
ak-more	**0.1064**	**0.4440**	12.71	0.1652	0.1046	0.0668	0.0426	0.6888	0.8538	0.5980	0.7797
	0.0561	0.2400	**21.82**	**0.4331**	**0.2968**	**0.2091**	**0.1511**	**0.7697**	**0.9114**	**0.7037**	**0.8197**
ck	**0.0813**	**0.3324**	13.24	0.3011	0.1809	0.1100	0.0669	0.5854	0.7676	0.5479	0.7794
	0.0465	0.1490	**13.52**	**0.4329**	**0.2775**	**0.1861**	**0.1307**	**0.6612**	**0.8334**	**0.6215**	**0.8145**

Table 2: The automatic results of PLATO+GOLD (up) and PLATO+ACK-DEF (down) on four test sets with unseen knowledge.

test set	Recall	Precision	F1	avg. Dec.
normal	0.3607	0.7009	0.4546	–
ak-less	0.2883	0.5585	0.3618	∇ 0.1026
ak-more	0.1752	0.3632	0.2228	∇ **0.2517**
ck	0.3193	0.6133	0.4003	∇ 0.0611
normal	0.3695	0.6538	0.4520	–
ak-less	0.3251	0.5636	0.3927	∇ 0.0647
ak-more	0.2335	0.3983	0.2775	∇ 0.1887
ck	0.1065	0.2041	0.1337	∇ **0.3437**

Table 3: The knowledge correlation results of PLATO+GOLD (up) and PLATO+ACK-DEF (down) on four test sets with seen knowledge.

test set	Recall	Precision	F1	avg. Dec.
normal	0.3732	0.7442	0.4736	–
ak-less	0.2728	0.5475	0.3452	∇ 0.1418
ak-more	0.1665	0.3627	0.2152	∇ **0.2822**
ck	0.3028	0.6068	0.3830	∇ 0.0995
normal	0.3655	0.6882	0.4535	–
ak-less	0.2938	0.5348	0.3579	∇ 0.1069
ak-more	0.2046	0.3714	0.2481	∇ 0.2277
ck	0.0870	0.1847	0.1116	∇ **0.3747**

Table 4: The knowledge correlation results of PLATO+GOLD (up) and PLATO+ACK-DEF (down) on four test sets with unseen knowledge.

Even if the knowledge information only changes by 10% to 20%, the performance of the model will significantly decline, especially consistency metrics (i.e. BLEU, GRE, AVG and EXT). (3) the PLATO+GOLD achieve better Distinct scores but weaker BLEU and embedding-based scores, which means that the PLATO+GOLD is easy to generate responses that are very different from ground-truth responses, that is, the hallucinations.

3.3 Knowledge Correlation Analysis

Table 3 and Table 4 report the knowledge correlation result of PLATO+GOLD and PLATO+ACK-DEF on four test sets and four test unseen sets, respectively. From these table, we can observe that the performance of PLATO+GOLD is reduced when the given knowledge changed, which illustrates the danger that the model generate responses based on incorrect knowledge. In addition to the above findings, we also observed that the recall, precision and f1 scores of PLATO+ACK-DEF are better than PLATO+GOLD on ak-less and ak-more sets, which demonstrates that using ACK-DEF effectively enhance the model's capability for the similar knowledge information. Moreover, the result of PLATO+ACK-DEF on the ck set is significantly reduced, which shows that the model distinguishes the wrong knowledge constructed with antonyms and gives an appropriate response with-

test set	w. GOLD (%)	w. ACK-DEF (%)	kappa
normal	**13.00**	14.00	0.481
ak-less	23.67	**17.33**	0.513
ak-more	33.67	**24.33**	0.479
ck	21.67	**5.67**	0.597
total	23.00	**15.33**	0.552

Table 5: The human evaluation results.

out knowledge (see Table 1 and Table 2 for the effect). These results are inline with our exception that incorporating noised knowledge dialogue samples in training stages can smooth the polarized optimization objective, and mitigate the hallucination problem.

According to the results of test seen sets and unseen sets), we can notice that the PLATO+ACK-DEF achieves a good performance on ground-truth seen knowledge and a weak performance on ground-truth unseen knowledge. This illustrates that the PLATO+ACK-DEF may doubt the authenticity of unseen given knowledge (even if the knowledge is the ground-truth), and will not fully use it to generate responses. This may alleviate the hallucination, and we believe it is caused by (1) the Augmentative knowledge dialogue introduce similar knowledge to improve the generalization of the model; (2) the Contrastive knowledge dialogue introduce knowledge independent responses, which tell the model to generate responses without knowledge; (3) the ACK-DEF smooths the polarized optimization, which ensure the model not to directly use the given knowledge.

3.4 Human Evaluation

To further evaluation the ability of our ACK-DEF on reducing the hallucination problem, we randomly select 400 samples form four test sets, and hire three annotators to do human evaluations by assessing whether the responses generated by PLATO +GOLD and +ACK-DEL have hallucinations. Table 5 reports the results of human evaluation, from which we can notice that the PLATO+ACK-DEF generate less hallucinations than PLATO+GOLD. This shows the effectiveness of our ACK-DEF.

3.5 Case Study

Table 6 shows a generated case of PLATO+GOLD (GOLD) and PLATO+ACK-DEF (ACK-DEF) based on different knowledge information. We can observed that the GOLD is convinced of the given knowledge, regardless of whether the knowledge is appropriate or not, and more easily to copy the knowledge information into responses. Even the GOLD has seen the knowledge topic, it could not remember the knowledge in their parameters. On the contrary, the ACK-DEF has good resistance to incomplete correct knowledge.

Query	They really are super cool looking dogs they are so amazing
Normal Knowledge	golden retriever are well suited to residency in suburban or country environments.
GOLD	I love golden retrievers! They are well suited to residency in suburban or country.
ACK-DEF	They are! They are well suited to suburban or country environments.
AK-Less Knowledge	golden retriever are good suited to residency in suburban operating elbow room land environments.
GOLD	Golden retriever are great suited to residency in suburban operating elbow
ACK-DEF	Golden retriever are great suited to suburban environments.
AK-More Knowledge	gild retriever make up considerably suit to residency in suburban surgery commonwealth environments.
GOLD	gild retrievers make up considerably suit to residency in suburban areas.
ACK-DEF	*Not very clear, maybe* they are! They can be found in suburban areas. *If it is wrong, please tell me.*
CK Knowledge	golden retriever unequal disadvantageously suited to residency in suburban or country environments.
GOLD	They sure are! I have a golden retriever who is unequally suited to residency in suburban environments.
ACK-DEF	*That sounds pretty good.*

Table 6: A case of PLATO +GOLD and +ACK-DEF.

4 Conclusion

This paper focuses on the hallucinations caused by polarized optimization objective in knowledge-grounded dialogue generation (KGDG), and proposes an augmentative and contrastive knowledge dialogue expansion framework (ACK-DEF) to mitigate it. The optimization objective of KGDG is to train the model could generate proper response with or without knowledge, which inevitably weaken the model's ability on unrecognized knowledge and lead hallucinations. Therefore, ACK-DEF constructs multiple level knowledge-dialogue samples to soften the optimization objective of KGDG. Extension experimental results show the superior performance of using our methods on dialogue metrics and knowledge correlations.

Limitations

Our limitations are as follow:

- **Data Scale**: This paper only employ the Wikipedia of Wizard dataset, a small scale and well-established knowledge conversation dataset, and lack of the validation on large-scale dataset.
- **Backbones**: This paper lacks the evaluating of other knowledge dialogue model on the proposed method. Actually, we have two reasons to employ the PLATO. First, the PLATO can better handle the one-to-many phenomenon, which is suitable for learning our expansion samples. Second, the PLATO is a pre-trained dialogue model, and its performance on knowledge dialogue generation task has been proved. We will evaluating the performance of other knowledge dialogue model on our method for our future work.
- **Knowledge Expansion Methods**: This paper only use the synonym and antonym to construct the noised knowledge, which lacks of the comparison of using other data augment method. Indeed, we use two token-level data augmentation methods (synonym and antonym augmentation) to prove our statements on hallucination problem in knowledge-dialogue generation task. Based on this study, we believe that incorporating other data augmentation methods will also mitigate the hallucinations.
- **Manual Prompts and Responses**: This paper designed five prefix prompts, four post-prompts and nineteen euphemistic responses. For *AK-More* method, we simply randomly choose one prefix-prompt and one post-prompt and concatenate them with the ground-truth response. This leads to some irregular responses. As for *CK* method, we randomly select one euphemistic response for the incorrect knowledge. However, we found that the response may not coherent with the query. We will design more smooth expansion ways to construct more human-like training samples for our future work.

Ethics Statement

We acknowledge and ensure that our study is compatible with the provided Code of Ethics. Knowledge-grounded open-domain dialogue generation is crucial for building a knowledgeable dialogue system, which is beyond the wildest dreams in natural language process field. All our experiments are conducted on public available datasets to avoid ethical concerns. All terms for using these datasets are strictly followed in our study. There are no direct ethical concerns in our research.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. This research is supported by Beijing Natural Science Foundation (No.4222037 and L181010) and BIT Research and Innovation Promoting Project (Grant No.2022YCXY021). Kan Li is the corresponding author.

References

Leonard Adolphs, Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2021. Reason first, then respond: Modular generation for knowledge-infused dialogue. *CoRR*, abs/2111.05204.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: pre-trained dialogue generation model with discrete latent variable. In *ACL*, pages 85–96. ACL.

Siqi Bao, Huang He, Jun Xu, Hua Lu, Fan Wang, Hua Wu, Han Zhou, Wenquan Wu, Zheng-Yu Niu, and Haifeng Wang. 2022. PLATO-K: internal and external knowledge enhanced dialogue generation. *CoRR*, abs/2211.00910.

Hengyi Cai, Hongshen Chen, Yonghao Song, Zhuoye Ding, Yongjun Bao, Weipeng Yan, and Xiaofang Zhao. 2020. Group-wise contrastive learning for neural dialogue generation. In *EMNLP*, pages 793–802. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297.

Richard Csaky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *ACL (1)*, pages 5650–5669.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*. OpenReview.net.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar R. Zaïane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *NAACL*, pages 5271–5285. Association for Computational Linguistics.

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

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tür. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *INLG*, pages 412–421. Association for Computational Linguistics.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *ICLR*. OpenReview.net.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *ACL*, pages 8460–8478. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*, pages 110–119.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI*, pages 5081–5087. ijcai.org.

Chia-Wei Liu, Ryan Lowe, Iulian 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. In *EMNLP*, pages 2122–2132.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of ACL*, pages 1317–1337. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *ACL/IJCNLP*, pages 704–718. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2021. The curse of dense low-dimensional information retrieval for large index sizes. In *ACL/IJCNLP*, pages 605–611. Association for Computational Linguistics.

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

Bin Sun, Shaoxiong Feng, Yiwei Li, Jiamou Liu, and Kan Li. 2021. Generating relevant and coherent dialogue responses using self-separated conditional variational autoencoders. In *ACL/IJCNLP*, pages 5624–5637. Association for Computational Linguistics.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2022. Contrastive learning reduces hallucination in conversations. *CoRR*, abs/2212.10400.

Josef Valvoda, Yimai Fang, and David Vandyke. 2022. Prompting for a conversation: How to control a dialog model? In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 1–8, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *ACL*, pages 3794–3804. Association for Computational Linguistics.

Xinnuo Xu, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *EMNLP*, pages 3981–3991.

Yuan Yao, Bowen Dong, Ao Zhang, Zhengyan Zhang, Ruobing Xie, Zhiyuan Liu, Leyu Lin, Maosong Sun, and Jianyong Wang. 2022. Prompt tuning for discriminative pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3468–3473, Dublin, Ireland. Association for Computational Linguistics.

Xueliang Zhao, Tingchen Fu, Chongyang Tao, and Rui Yan. 2022a. There is no standard answer: Knowledge-grounded dialogue generation with adversarial activated multi-reference learning. *CoRR*, abs/2210.12459.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. In *ICLR*. OpenReview.net.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models. In *EMNLP*, pages 3377–3390. Association for Computational Linguistics.

Yingxiu Zhao, Yinhe Zheng, Zhiliang Tian, Chang Gao, Bowen Yu, Haiyang Yu, Yongbin Li, Jian Sun, and Nevin L. Zhang. 2022b. Prompt conditioned VAE: enhancing generative replay for lifelong learning in task-oriented dialogue. *CoRR*, abs/2210.07783.

Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2021. Knowledge-grounded dialogue generation with term-level de-noising. In *Findings of ACL/IJCNLP*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2972–2983. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *EMNLP*, pages 708–713. Association for Computational Linguistics.

Prefix Prompts	*Post Prompts*
I was thinking that perhaps	Maybe i am wrong.
I am not sure, maybe that	If I am wrong, please correct me.
Not very clear, maybe	If I am wrong, please forgive me.
Not very clear, perhaps	If it is wrong, please tell me.
I was thinking that maybe	

Table 7: The designed prefix and post prompts.

Euphemistic Responses
Interesting, do you know that?
That sounds pretty good. Are there any way to visit?
Oh, I had not heard.
Hmm, I have never heard of that. What is that one about?
I have never heard. Can you tell me more about it?
Oh, wow, that is remarkable.
I have never played those, are they fun?
Can I ask you about it?
Please tell me more about that.
Can you tell me more about that?
I have never had that. Anything else you can tell me?
That's really interesting! But I have never heard of that.
I literally know nothing about that!
I have no idea about that.
I have not heard that one. I will have to check it out.
Huh, maybe I will need to check that out then.
Oh, I misunderstood then.
Oh, i do not know about that.
Wow, that's a lot! I haven't heard of those.

Table 8: The designed euphemistic responses.

A Prefix and Post Prompts

We manually design five prefix prompts and four post prompts, which are shown in Table 7. We discuss below about the prefixes and posts.

We designed the prefixes and posts based on the WoW dataset and our daily conversation habits. In WoW dataset, one role is "0_Wizard", and the other is "1_Apprentice". We noticed that the 1_Apprentice will give the sentences such as "*correct my if I am wrong ...*", which is also easy to appear in our daily conversation. Taking inspiration of this, we manually designed the prefixes and posts. Moreover, since the PLATO is pre-trained on conversation datasets, these prefixes may introduce the pre-knowledge that the model learned during the pre-training process.

In fact, we declare the weakness of our manual prefixes and posts, i.e. direct connections of prefixes, responses, and posts do not fit all contexts. Therefore, we are exploring a new way of constructing replies, such as passing the design prefix, response, post, and context into the large-language-model to rewrite the appropriate response. We believe that better prefixes and posts will lead to more benefits in solving the hallucination problem.

B Euphemistic Responses

We manually design nineteen euphemistic responses, which are shown in Table 8.

C Dissuasion about the boundary between ak-less and ak-more

Below we provide an example in our dataset:

- Ground-truth Knowledge: laziness | tesis (" thesis ") is a 1996 spanish thriller film.
- AK-Less Knowledge: acedia | tesis ("thesis") is a 1996 spanish thriller film.
- AK_More Knowledge: laziness | tesis ("thesis") personate a 1996 spanish thriller picture show.

It can be noted that the more synonyms are introduced into a sentence, the semantics of the sentence will become more and more different from the original semantics. Therefore, we suppose that replacing at least 30% of words at once will make a big difference in sentence semantics. Then, we decided the boundary between ak-less and ak-more.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We provide a section of Limitations after the Conclusion and before the Ethics Statement

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We use a publicly well-established dataset.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

3

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We use the released code and checkpoints. We cite the source of our model.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
2

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.