Towards Robust Personalized Dialogue Generation via Order-Insensitive Representation Regularization

Liang Chen, Hongru Wang, Yang Deng, Wai-Chung Kwan, Zezhong Wang, Kam-Fai Wong
The Chinese University of Hong Kong
MoE Key Laboratory of High Confidence Software Technologies
{lchen,hrwang,wckwan,zzwang,kfwong}@se.cuhk.edu.hk
{dengyang17dydy}@gmail.com

Abstract

Generating persona consistent dialogue response is important for developing an intelligent conversational agent. Recent works typically fine-tune large-scale pre-trained models on this task by concatenating persona texts and dialogue history as a single input sequence to generate the target response. While simple and effective, our analysis shows that this popular practice is seriously affected by ***Order Sensitivity*** where different input orders of persona sentences significantly impact the quality and consistency of generated response, resulting in severe performance fluctuations (i.e., 29.4% on GPT2 and 83.2% on BART). To mitigate the order sensitivity problem, we propose a model-agnostic framework, **OR**der **I**nsensitive **G**eneration (**ORIG**), which enables dialogue models to learn robust representation under different persona orders and improve the consistency of response generation. Experiments on Persona-Chat dataset justify the effectiveness and superiority of our method with two dominant pre-trained models (GPT2 and BART).[1]

1 Introduction

Developing a persona-consistent dialogue model has been one of the key issues and crucial problems in open-domain dialogue systems (Huang et al., 2020). Zhang et al. (2018a) define the problem of personalized dialogue generation, which aims to generate personalized responses based on textually described persona profiles. Many efforts have been made on developing dialogue models that generate responses consistent with the provided persona profile (Song et al., 2019, 2020a,b; Wu et al., 2020a).

The recent development in transformer-based pre-trained models (Vaswani et al., 2017; Devlin et al., 2018; Liu et al., 2019; Chen, 2020) has led to great successes in dialogue systems (Wolf et al., 2019; Wu et al., 2020b; Ham et al., 2020; Kulhánek et al., 2021; Cao et al., 2022; Deng et al., 2022b,c, 2023). Inspired by these successes, previous works incorporate those pre-trained models in persona-based response generation by concatenating the dialogue history and persona as input to generate the response in an auto-regressive manner (Song et al., 2021; Liu et al., 2022). However, a fine-tuned model can generate a high-quality and persona-consistent response in a certain ordering of personas, while varying this order may lead to a generic and even inconsistent response as illustrated by the example in Figure 1. We empirically show that the worst ordering of persona can lead to a 29.4% decline in BLEU score compared with the best ordering.

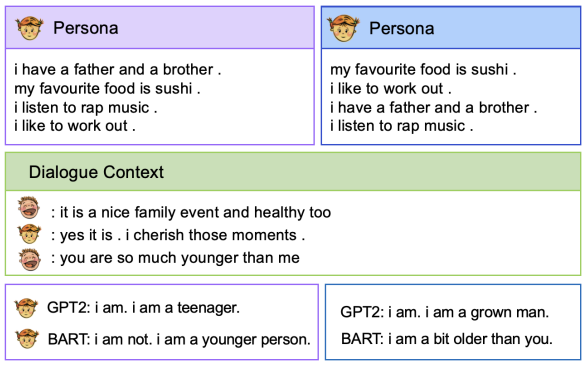

Figure 1: A dialog extract from Persona-Chat showing different orderings of the same persona can lead to different and even inconsistent responses.

Ideally, a well-trained dialogue generation model should be able to generate a persona-consistent response regardless of the ordering of personas in the input. We perform experiments and analyses to identify the cause of the ordering sensitivity. We find that the ordering of persona in the input leads to different representations of context and response. We also show that the model can attend to the appropriate persona and generate high-quality responses under some representations but not under others.

[1]The code is available at https://github.com/ChanLiang/ORIG.

Findings of the Association for Computational Linguistics: ACL 2023, pages 7337–7345
July 9-14, 2023

This leads to instability in response generation.

Motivated by the above findings, we propose **OR**der **I**nsensitive **G**eneration (**ORIG**), which is a simple and effective framework that helps models learn more robust and better representations for different persona orders. More specifically, we formulate ORIG as a constrained optimization problem, which optimizes a persona response generation objective under the constraint: given different orderings of persona, the response representations of the model are the same. Then we optimize it through a stochastic optimization approach.

Experimental results on the Persona-Chat dataset show that ORIG significantly improves the robustness of pre-trained models (GPT2 (Radford et al., 2019) and BART (Lewis et al., 2020)) under different orderings of input persona, as well as advances their generation performance.

In summary, our contributions are threefold: (1) We identify the order sensitivity problem in persona dialogue generation and conduct an empirical analysis to reveal its underlying reasons. (2) We propose a model-agnostic framework, ORIG, that helps different persona dialogue models learn robust representations while achieving better performance. (3) We perform extensive experiments on the Persona-Chat dataset, showing that ORIG outperforms previous models and is more robust and less sensitive to different persona orderings.

2 Related Work

Maintaining a consistent persona is essential for building a human-like dialogue system, where most works regard persona as a set of sentences along with each dialog (Zhang et al., 2018a; Gu et al., 2019; Song et al., 2019; Wu et al., 2021; Cao et al., 2022; Deng et al., 2022a). Song et al. (2021) disentangled the task of persona-based dialogue generation into two sub-tasks: consistency understanding and dialogue generation while Cao et al. (2022) aims to alleviate the problem of limited data by data manipulation methods. Despite satisfactory performance in previous work, the impacts of different orders of personas are still under-explored, resulting in unstable and inconsistent responses.

Our work is also related to work on order sensitivity in prompt-based few-shot learning (Zhao et al., 2021; Lu et al., 2022). Zhao et al. (2021) found that the different order of training examples in the prompt can cause accuracy to vary from near chance to state-of-the-art in the few-shot classification setting. Similarly, order sensitivity for In-context Learning also exists regardless of model size and the prompt format (Lu et al., 2022). Distinguishing from them, we focus on order sensitivity in the language generation task in finetuning setting, especially the impacts of persona orderings to generate persona-consistent responses.

Model	BLEU-1	BLEU-2	ROUGE	CIDEr
GPT2-best	16.79	9.25	18.44	17.56
GPT2-worst	11.85	5.83	11.79	5.51
BART-best.	28.17	18.29	31.07	46.53
BART-worst.	4.73	1.99	4.37	1.34

Table 1: Performance gap between the best case and worst case when changing the ordering of input persona.

3 Order Sensitivity Problem and Analysis

In this section, we first illustrate the seriousness of the order sensitivity problem by showing a huge performance fluctuation in persona dialogue models when fed the same personas in the best and worst orders. Then we analyse why their performance is volatile to different persona orderings.

To illustrate the problem, we finetune PLMs on the Persona-Chat by concatenating the persona and dialogue context together to predict the target response, including GPT2 and BART. After the training converges, we test them on two settings: (1) the best case: for each test sample, we feed the models all possible permutations of persona sentences and keep the maximum score for each sample as the final score; (2) the worst-case: perform the same process as (1), but take the minimum score. Table 1 shows the results for two models. Surprisingly, we find the ordering of input persona has a big impact on the models' performance: GPT2's worst case is 29.4% lower than its best case, while BART's is 83.2% lower.

Moreover, we find that the huge fluctuation in models' performance is closely related to the response representation changes caused by different orderings of input persona sentences. Concretely, we measure the similarity of the responses representation of the same test sample under different input orders of persona. We show their token-level similarity in the Table 2 (persona and context are omitted for brevity), where the bidirectional KL function is employed as the distance function. Ideally, models should have the consistent response representation: KL distance between the same re-

BART	great(0.185) and(0.105) how(0.312) was(0.289) your(0.124) day(0.304) ?

Table 2: The token-level representation of the same response can be very different when the ordering of input persona changes. The value denotes the KL distance of the same tokens representation returned by the models fed with two different orderings of persona.

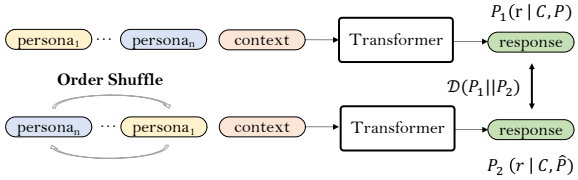

Figure 2: Our proposed framework ORIG

sponse should be zero. However, their distances are significantly higher than zero. It reveals that the models behave more likely a left-to-right language model whose representation is prone to the different orderings of the previous input (e.g. persona sentences). That is highly undesirable for a robust personalized dialogue model. Thus, regularization of representation for the response tokens is necessary to help personalized dialogue models capture order-invariant representation.

4 Method

We introduce the proposed framework, named **ORIG**: **OR**der **I**nsensitive **G**eneration (**ORIG**). As shown in Figure 2, we transform the persona order-sensitivity problem as a constrained optimization problem that optimises a persona dialogue model under the uncertainty of the input persona order.

4.1 Problem Formulation

Given the dialogue context $C = \{u_1, \ldots, u_m\}$ and a set of persona descriptions $P = \{p_1, \ldots, p_n\}$, the goal is to generate a personalized response r. Formally, the generation problem can be formulated as the following chain rule:

$$P(r|C, P; \theta) = \prod_{t=1}^{T} P\left(r_t|r_{1:t-1}, C, P; \theta\right) \quad (1)$$

where θ is the parameters of the dialogue model.

4.2 ORIG Framework

According to the analysis in Section 3, the observation reveals that varying the order of input personas leads to different representations of the dialogue response, thus resulting in fluctuations in performance.

To learn more robust and consistent representations, we propose the ORIG framework that complements the response generation process with a constraint: given the different orderings of a persona, the model's response representations need to be the same.

Then the order-insensitive personalized dialogue generation problem is modelled as the following constrained optimization problem

$$\min_{\theta}[-\log P(r|C, P; \theta)] \quad (2)$$

$$\textit{s.t.} \quad \mathcal{D}[P(r|C, P; \theta), P(r|C, \hat{P}; \theta))] = 0 \quad (3)$$

$$(P, C, r) \sim D \quad (4)$$

$$\hat{P} \sim \text{Shuffle}(P) \quad (5)$$

where $P(r|C, P; \theta)$ are the model's predictions over the dialogue response, D denotes the dialogue corpus, and the function $\mathcal{D}$ is KL divergence to measure the difference between two distributions, and the Shuffle operator samples each persona ordering uniformly from the full permutation of P.

4.3 Optimization

As for optimization, we first apply the Lagrange multipliers strategy to convert the constrained problem into an unconstrained problem

$$\begin{aligned}\mathcal{L}_\theta = &- \log P(r|C, P; \theta) \\ &+\gamma \cdot \mathcal{D}[P(r|C, P; \theta), P(r|C, \hat{P}; \theta)]\end{aligned} \quad (6)$$

where γ is the multiplier corresponding to the equality constraints (3). Then we can update the parameters θ of dialogue models by stochastic gradient descent.

5 Experiments

5.1 Experimental Setups

Datasets We evaluate the models on the Persona-Chat dataset (Zhang et al., 2018a), where each dialogue session has at least 6 turns of interactions. And each interaction is conditioned on a persona that is described with 5 profile sentences.

Baselines To verify the generality of our framework across different architectures, we perform experiments on the two most popular pre-trained architectures: Transformer encoder-decoder (BART) and Transformer decoder (GPT2).

Implementation Details We choose GPT2 base (117M) and BART base (139M) as the base models and compare the base models finetuned with classical max likelihood estimation (MLE) and our

	Automatic Evaluations						Human Evaluations		
Model	BLEU-1	BLEU-2	ROUGE	Entropy	CIDEr	**C**-score	Flu.	Con. Coh.	Per. Cons.
GPT2	13.95	7.22	14.82	6.53	10.10	0.718	1.531	1.281	1.719
GPT2-ORIG	**14.61**	**7.43**	**14.94**	**6.54**	**10.60**	**0.733**	**1.726**	**1.512**	**1.719**
BART	14.19	7.61	15.05	**6.67**	11.07	0.443	1.906	1.312	1.438
BART-ORIG	**14.64**	**7.90**	**15.20**	6.41	**13.27**	**0.446**	**1.938**	**1.332**	**1.457**

Table 3: Automatic and human evaluation results of applying ORIG on two base models in the original test set without any modifications on input persona orders.

proposed ORIG. Our implementation was based on HuggingFace's Transformers library (Wolf et al., 2020). During training, the learning rate is set as 2×10^{-5}, and the batch size for GPT2 and BART is set as 64 and 32, respectively. We trained both models for 10 epochs with Adam (Kingma and Ba, 2015) optimizer until they converged. During decoding, We employ a top-p (p=0.9) (Holtzman et al., 2020) plus top-k (k=50) sampling strategy, which is used to avoid sampling from the unreliable tail of the distribution (only consider a subset of vocabulary composed of k words with the highest probability or some most probable words whose sum of probabilities equals p at each decoding step). The random seed for all experiments is set to 42.

Evaluation Metrics We perform both automatic and human evaluations. (1) Automatic metrics: We adopt BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), Entropy(Zhang et al., 2018b) and CIDEr (Vedantam et al., 2015) for lexical-based measurement. Following previous work, we also adopt the C-score (Madotto et al., 2019) to indicate the consistency between the generated response and the input personas. C-score is calculated by the entailment score of a RoBERTa model finetuned on the DialogueNLI dataset. (2) Human evaluation: We randomly sampled 200 samples from the test set and ask 3 crowdworkers to rate the generated responses in the following three aspects: response fluency, context coherence and persona consistency. The scores $\{0, 1, 2\}$ indicate unacceptable, acceptable and excellent, respectively. The degree of agreement during human evaluation is measured by Fleiss' kappa (Fleiss, 1971).

5.2 Experimental Results

Improves performance in the original test set Table 3 shows different models' performance in the original test set without any modifications (for ORIG, "Shuffle" is used during training but is optional during testing. The Table 3 caption signifies the absence of "Shuffle" during testing. This is to evaluate if ORIG performs well in the normal setting). From automatic metrics, we can see base models trained with our ORIG framework outperform the baselines. It justifies that our framework can be applied to different models to improve their performance. From human evaluation results, models with ORIG are superior to others on almost all metircs, especially on GPT2. This is consistent with the results of automatic metrics. The average kappa value of the annotation is 0.632, indicating

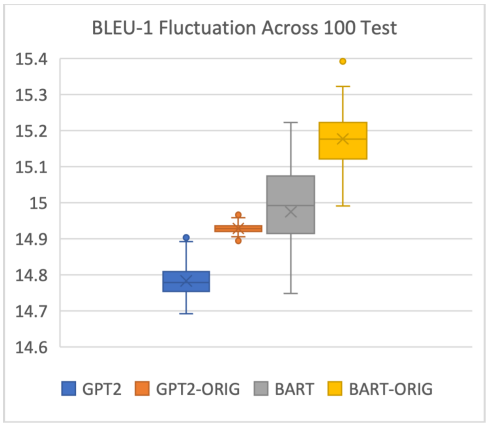

Figure 3: Our ORIG improves the mean performance while reducing the variance of both models. Statistics about running each model 100 times on the test set and randomly shuffling the order of the input persona sentences in each run.

Model	mean	variance	best	worst
GPT2	14.78	0.00193	**16.79**	11.85
GPT2-ORIG	**14.93**	**0.00016**	14.95	**14.25**
BART	15.01	0.01123	**28.17**	4.73
BART-ORIG	**15.18**	**0.00532**	26.44	**5.80**

Table 4: Statistical results of BLEU-1. The mean and variance are obtained by running each model 100 times on the test set and randomly shuffling the order of the input persona sentences in each run. The best case and worst case are obtained by feeding models the best and worst orderings of personas for every test sample.

good agreement during human evaluation.

Reduces variance and improves mean and worst-case performance Figure 3 shows that aside from reducing the variance, ORIG also improves mean and worst-case performance (detailed results in Table 4) across two models consistently, especially in GPT2 (the worst case performance is very close to the best case). We reduce the variance on GPT2 and BART by 91.6% and 51.8%, respectively. Meanwhile, we improve worst-case performance by 20.3% and 22.6% on GPT2 and BART respectively. The only drop is the best case. This is because our distance function $\mathcal{D}$ is unidirectional, which pulls in the two representations in Equation 3 indiscriminately, causing the best case to go down and the worst to go up. We leave more complicated and directional distance constraints for future studies.

6 Conclusion

We show that the current practice of applying pre-trained models to the personalized dialogue generation task is volatile across different input orders of personas. Through the analysis, we find that the problem arises from the representation changes induced by the input changes. Motivated by these, we propose our ORIG, a model-agnostic framework for finetuning the persona dialogue model such that it obtains a persona order-invariant representation. Experiments on two dominant pre-trained dialogue models show that our framework improves performance and reduces order volatility.

Limitations

In this section, we discuss the limitations of this work. First, on the problems side, it's non-trivial to consider the order of all kinds of grounding knowledge, but we have only explored Persona-Chat. We hope to apply our method to more grounded generation tasks such as knowledge-grounded and document-grounded dialogue in the future. Second, on the methods side, our framework is training-based, but we hope more lightweight techniques could be developed to improve the model's robustness even without training the model.

Acknowledgements

We would like to thank Professor Helen Meng and Xixin Wu for their helpful discussion and feedback on the course SEEM5640. We also thank anonymous reviewers for their constructive comments. Thanks to Dr Honshan HO for his support. This research work is partially supported by CUHK under Project No. 3230366.

References

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002, Dublin, Ireland. Association for Computational Linguistics.

Liang Chen. 2020. Variance-reduced language pretraining via a mask proposal network.

Yang Deng, Yaliang Li, Wenxuan Zhang, Bolin Ding, and Wai Lam. 2022a. Toward personalized answer generation in e-commerce via multi-perspective preference modeling. *ACM Trans. Inf. Syst.*, 40(4):87:1–87:28.

Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022b. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In *WWW '22: The ACM Web Conference 2022*, pages 2998–3008.

Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2022c. A unified multi-task learning framework for multi-goal conversational recommender systems. *CoRR*, abs/2204.06923.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *arXiv preprint arXiv:2305.10172*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019. Dually interactive matching network for personalized response selection in retrieval-based chatbots. 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 1845–1854, Hong Kong, China. Association for Computational Linguistics.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. Improving Personality Consistency in Conversation by Persona Extending. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1350–1359.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of ACL 2019*, pages 5454–5459.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. 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 167–177, Online. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020a. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5821–5831.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting Persona Information for Diverse Generation of Conversational Responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5190–5196, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020b. Generating Persona Consistent Dialogues by Exploiting Natural Language Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8878–8885.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

R. Vedantam, C. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, Los Alamitos, CA, USA. IEEE Computer Society.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *arXiv:1901.08149 [cs]*.

Bowen Wu, Mengyuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020a. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 53–65.

Chen Henry Wu, Yinhe Zheng, Xiaoxi Mao, and Minlie Huang. 2021. Transferable persona-grounded dialogues via grounded minimal edits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2368–2382, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020b. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. 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, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *NIPS 2018*, pages 1810–1820.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation Section

☒ A2. Did you discuss any potential risks of your work?
The risk has not yet been identified.

☑ 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?

5.1

☑ B1. Did you cite the creators of artifacts you used?
5.1

☐ 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. 5

☐ 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?
Not applicable. Left blank.

☐ 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.
It's a public dataset.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5.1

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?
5.1

☑ 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?
5.2

☑ 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.)?
5.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
5.1

☒ 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.?
We provide a simple version in 5.1

☒ 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)?
All annotators are students.

☐ 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?
Not applicable. Left blank.

☐ 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.