Position Matters! Empirical Study of Order Effect in Knowledge-grounded Dialogue

Anonymous ACL submission

Abstract

With the power of large pretrained language models, various research works have integrated knowledge into dialogue systems. The traditional techniques treat knowledge as part of the input sequence for the dialogue system, prepending a set of knowledge statements in front of dialogue history. However, such a mechanism forces knowledge sets to be concatenated in an ordered manner, making models implicitly pay imbalanced attention to the sets during training. In this paper, we first investigate how the order of the knowledge set can influence autoregressive dialogue systems' responses. We conduct experiments on two commonly used dialogue datasets with two types of transformer-based models and find that models view the input knowledge unequally. To this end, we propose a simple and novel technique to alleviate the order effect by modifying the position embeddings of knowledge input in these models. With the proposed position embedding method, the experimental results show that each knowledge statement is uniformly considered to generate responses.

1 Introduction

Transformer-based (Vaswani et al., 2017) pretrained language models are widely used to build dialogue systems (Zhang et al., 2020; Xu et al., 2021; Komeili et al., 2021; Roller et al., 2020; Thoppilan et al., 2022; Rae et al., 2021; Chen et al., 2021; Ham et al., 2020; Hosseini-Asl et al., 2020; Bao et al., 2021). In addition to general-purpose dialogue systems, many specialized dialogue systems have been proposed. Representative examples include personalized dialogue systems (Wolf et al., 2019; Zhang et al., 2018; Wu et al., 2021; Cao et al., 2022; Song et al., 2020), knowledge-grounded dialogue systems (Dinan et al., 2019; Kim et al., 2021; Tao et al., 2021; Cai et al., 2020; Liu et al., 2021), and prompting dialogue systems (Su et al., 2022).

To build specialized dialogue systems, integrating additional information into the input sequence is necessary. Wolf et al. (2019) prepend persona sentences to personalize the history; while Su et al. (2022); Dinan et al. (2020); Keskar et al. (2019); Xu et al. (2020a) prepending task-specific signals to prompt and control the model.

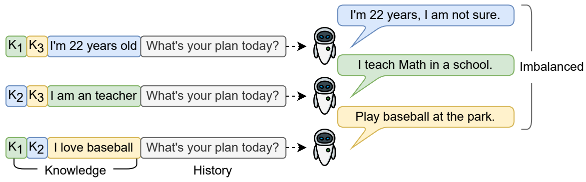


Figure 1: The order effect illustration. Models' responses are influenced by the order of the input knowledge set.

These methods prepend additional information in front of the history as a sequence for models' input. Furthermore, the approach generates an unnecessary order among equal knowledge sets since the knowledge is connected in the sequence. Thus models might be influenced by the order and generate imbalanced responses.

Previous works focus on how perturbations in dialog history affect models' responses (Sankar et al., 2019; O'Connor and Andreas, 2021; Sinha et al., 2021; Lampinen et al., 2022; Webson and Pavlick, 2021; Xu et al., 2020b; Khandelwal et al., 2018). They conduct many experiments and measure the effect of perturbations from the aspect of response quality and information theory to show that these language models are robust and not sensitive to the perturbations in input history. However, dialog history and knowledge are inherently different aspects of a conversation. Dialog history has a temporal property, i.e., the topic and specificity of conversation change as the dialog progresses, whereas knowledge facts are information referenced to generate a response. Although the perturbation in history does not influence the results generated by the model (Sankar et al., 2019; O'Connor and Andreas, 2021), in our early obser-

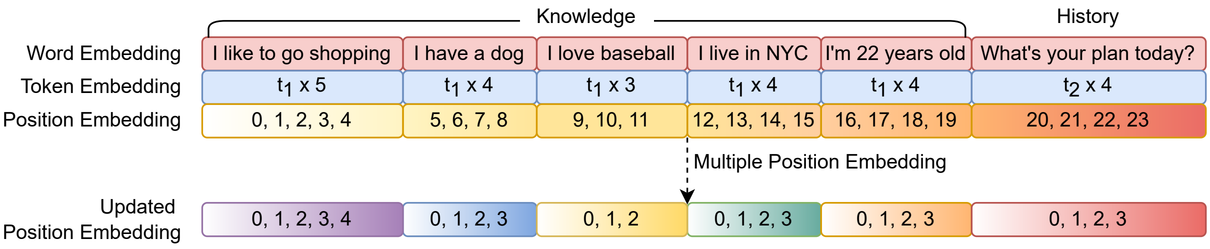


Figure 2: Input format for GPT-series models. The position ids do not treat knowledge equally but as a sequence. The updated position embeddings show our proposed method, where each knowledge statement is encoded with its own position embeddings, hence, models can treat each input sentence equally during training. The same color of blocks indicates using the same layer to generate embeddings.

vation, we found that prepending knowledge influences models' responses. For example, Figure 1 demonstrates an example where the model exhibits imbalanced attention to input knowledge, and the order of knowledge influences the generated responses. This might cause the model to generate inappropriate responses since it attends to knowledge that might not be relevant to a dialog context. The contributions of this work are as follows:

- We conduct experiments across two typical methods and two models on multiple datasets to show that the order of knowledge sentences does affect generated responses.
- We propose a simple approach to alleviate this sentence-level order effect by manipulating the position embedding layers.

2 Knowledge-grounded Dialogue Methods

In this work, we study the order effect in TransferTransfo (Wolf et al., 2019), which is a state-of-the-art knowledge-grounded method. We train TransferTransfo on two datasets and measure the sentence-level order effect on the test datasets.

2.1 TransferTransfo

The TransferTransfo architecture is built on top of GPT-series models, which simply concatenates the knowledge sets and context in a single sequence, putting the reply at the end. To help models distinguish speakers and position of input tokens, it builds three parallel input sequences for word, position, and segments, and fuses them into a single sequence. For the loss function, in addition to a language modeling loss, a next sentence prediction loss is added. The total loss is the weighted sum of the 1) language modeling loss, which is computed as the cross-entropy loss between the predicted logits and the ground truth response and 2) the next-sentence prediction loss, which is a classification loss to distinguish the ground truth response from distractors that are randomly sampled from the dataset.

In the original TransferTransfo implementation, the authors have already pointed out that the order of the knowledge set influences the model's performance. To this end, they augment training data by permuting the knowledge sets several times.

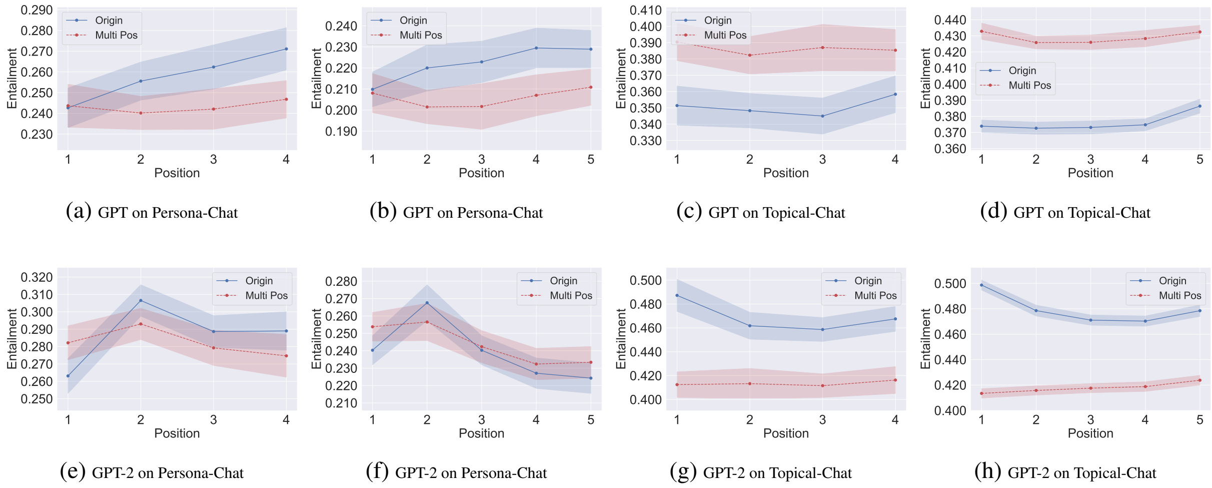
2.2 Experimental Setups

We conduct experiments on two datasets:
Persona-Chat (Zhang et al., 2018): This persona-grounded dialogue dataset consists of crowd-sourced dialogues between a pair of annotators provided with 4-5 persona statements each.
Topical-Chat (Gopalakrishnan et al., 2019): This is a knowledge-grounded dialogue dataset, where the dialogs are constructed by a pair of annotators conversing about specific topics. The annotators are provided with wiki data with 4-5 facts as knowledge sources.

In our experimental setup, we shuffle the knowledge set's order 50 times during testing and implement TransferTransfo on GPT (Radford et al., 2018) and GPT-2 (Radford et al., 2019) models.

3 The Order Effect of the Knowledge Set

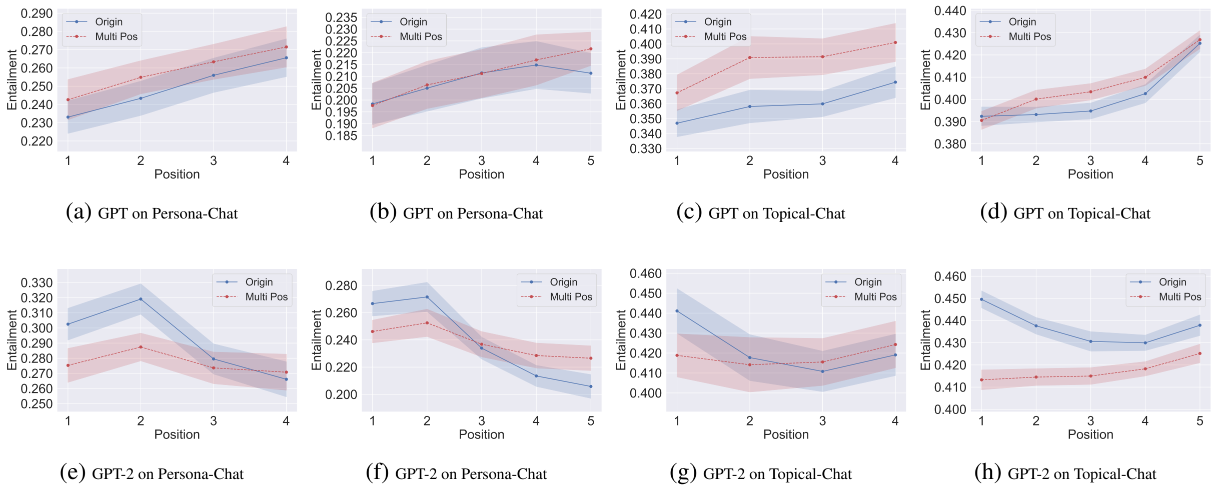
Models are said to have an order effect of input if the generated responses are sensitive and influenced by order of input sequence. Previous works (Sankar et al., 2019; O'Connor and Andreas, 2021; Sinha et al., 2021; Lampinen et al., 2022; Webson and Pavlick, 2021; Xu et al., 2020b; Khandelwal et al., 2018) focus on whether perturbation in dialogue history affect models' responses. In this



(a) GPT on Persona-Chat (b) GPT on Persona-Chat (c) GPT on Topical-Chat (d) GPT on Topical-Chat

(e) GPT-2 on Persona-Chat (f) GPT-2 on Persona-Chat (g) GPT-2 on Topical-Chat (h) GPT-2 on Topical-Chat

Figure 3: Experimental results under TransferTransfo method, the lines indicate the average of 50 times shuffling results with standard deviation represented in the area. The data with 4 and 5 knowledge sets are displayed separately.



(a) GPT on Persona-Chat (b) GPT on Persona-Chat (c) GPT on Topical-Chat (d) GPT on Topical-Chat

(e) GPT-2 on Persona-Chat (f) GPT-2 on Persona-Chat (g) GPT-2 on Topical-Chat (h) GPT-2 on Topical-Chat

Figure 4: Experimental results under LM loss only method, the lines indicate the average of 50 times shuffling results with standard deviation represented in the area. The data with 4 and 5 knowledge sets are displayed separately.

work, to be more specific, we investigate if sentence level change in the order of input knowledge sets will result in substantial semantic differences in the generated responses.

3.1 The Order Effect Measurement

To address the sentence-level order effect of the input knowledge set in models, we aim to measure the semantic difference given different orders of knowledge sentences. It is intuitive to measure if the response content is influenced by knowledge sets order. In other words, we measure the distribution of response-knowledge relationship in different positions. We build a Natural Language Inference (NLI) classifier to evaluate the degree of entailment between responses and each knowledge in the set.

The Natural Language Inference Classifier is built with BERT model (Devlin et al., 2019), trained on the Dialogue NLI dataset (Welleck et al., 2019), which is built on top of Persona-Chat dataset (Zhang et al., 2018). The annotators label the relationship between persona and response in Persona-Chat with entail, neutral, and contradict classes.

3.2 Results and Discussions for Order Effect

Figures 3 and 4 show the entailment scores of the response with each position of knowledge. Figure 3 presents the experiments of TransferTransfo with GPT and GPT-2 models across Persona-Chat and Topical-Chat datasets. Figure 4 shows the results with "LM Loss only Method", which refers to TransferTransfo without the next sentence prediction. We observe that the distribution of data

Model	Method	Persona		Topical	
		TT.	LM.	TT.	LM.
		Entailment Max - Min			
GPT	Origin	.048 / .037	.052 / .035	.037 / .022	.046 / .041
	Multi Pos	**.023 / .028**	.051 / .041	**.031 / .016**	.058 / .044
GPT-2	Origin	.062 / .062	.075 / .085	.052 / .036	.052 / .027
	Multi Pos	**.039 / .044**	**.038 / .045**	**.027 / .018**	**.035 / .021**
		Perplexity ↓			
GPT	Origin	52.29	54.31	39.31	36.80
	Multi Pos	55.47	58.43	42.37	42.98
GPT-2	Origin	61.69	61.80	20.50	18.84
	Multi Pos	60.18	58.91	17.40	17.30
		Coherence			
GPT	Origin	0.633	0.636	0.793	0.770
	Multi Pos	0.644	0.621	0.732	0.744
GPT-2	Origin	0.661	0.667	0.840	0.843
	Multi Pos	0.648	0.662	0.830	0.831
		Diverstiy ↓			
GPT	Origin	0.815	0.822	0.844	0.846
	Multi Pos	0.821	0.833	0.870	0.862
GPT-2	Origin	0.808	0.811	0.833	0.833
	Multi Pos	0.816	0.817	0.843	0.845

Table 1: The results of measurements. The Max-Min of entailment are reported in 4 knowledge / 5 knowledge. The mean of quality across 50 runs are reported and standard deviation are reported in Appendix A.3.

containing only four knowledge statements is very different compared to data containing five knowledge statements. Hence we show them separately.

The NLI classification results are shown with BLUE lines. We can see that the distribution of entailment scores on different positions are imbalanced. In the experiments on the GPT model, (figures 3a, 3b, 3c, 3d, 4a, 4b, 4c, and 4d), it can be observed under both TransferTransfo and LM loss only methods, the entailment score on the last position is always the highest. In fact, there is a huge gap between the entailment scores with the first knowledge and the last knowledge statements. This indicates that GPT model focuses more on the last position of knowledge.

However, the behavior of GPT-2 is very different from GPT model. From Figures 3e, 3f, 3g, 3h, 4e, 4f, 4g, and 4h, we can see that GPT-2 models focus more on the earlier knowledge statements in the sequence rather than the later ones.

These results show that the order effect exists across GPT and GPT-2 models (although different) and is influencing models' responses and this needs to be solved.

4 Alleviate the Order Effect

In this section, we analyse the reason for the order effect in the GPT-series models and propose a method to alleviate the phenomenon. Figure 2 shows the input format of the classic GPT-series. There are three types of embeddings in the model: word embedding to capture the semantic meaning of each word, token embedding to represent the speaker and absolute position embedding that encodes position information of input sequence.

Figure 2 shows that the position ids for each knowledge start from zero with different positional embedding layers. In this naive setting, knowledge of the set are treated equally and not input with the order during training.

4.1 Results and Discussion

In the same Figures 3 and 4, the RED lines demonstrate the entailment result after applying multiple position embedding. We observe that all the red lines, which are the GPT-series applied multiple position embeddings, are much smoother compared to BLUE lines in both figures. Furthermore, we report the difference between maximum and minimum entailment across the positions in Table 1. It shows that the difference is negligible after applying multiple position embeddings. This indicates that we can alleviate the order effect under models trained with with multiple position embedding. However, we also observed that on Figure 4 some red lines are still as steep as before, which means the order effect still exists. We think that the model trained only with LM loss treats knowledge like history and does not ground models on knowledge sets. Under this scenario, the multiple position embedding doesn't work well.

For the measurement of quality, Table 1 shows the perplexity, coherence, and diversity. The details are included in Appendix A.2. We found tiny drops between origin and multiple position embedding. More specifically, our proposed method does not crash the models and can still make models generate plausible responses.

5 Conclusions

In this paper, we investigate whether the order of knowledge set will influence dialogue models' responses. Our experiments across several datasets show that the GPT-series models unfairly pay attention to the knowledge set and are influenced by order of knowledge. To solve this problem, we study the reason for the phenomenon and propose simple method to alleviate the order effect in models. The experimental results show that our approach reduces the order effect and makes the model select the knowledge uniformly.

Limitations

This work has potential limitations:

- We found that on the Figure 3 and 4, The entailment of the methods after applying multiple position embedding (RED lines) are sometimes lower than origin methods(BLUE lines). This is not meet our expectations since we don't want our method to decrease performance. In our opinion, we think the reason might be the embedding method has never been seen before during the pretraining of models, which requires the model's additional efforts to adapt the embedding, thus hurts the performance.. We leave it as future work to be improved.

- We also found that the multiple position embedding does not work very well to alleviate the order effect in the LM loss-only settings4. We have discussed this in previous sections. Since LM loss only does not help the model distinguish which parts in the input sequence are knowledge set and thus treat them the same as history. The multiple position embedding will not be trained finely to help the model distinguish. We also left this as a future work to be improved.

References

Siqi Bao, Bingjin Chen, H. He, Xin Tian, Hanxu Zhou, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, and Yingzhan Lin. 2021. A unified pretraining framework for conversational ai. *ArXiv*, abs/2105.02482.

Yuanyuan Cai, Min Zuo, Qingchuan Zhang, Haitao Xiong, and Ke Li. 2020. A bichannel transformer with context encoding for document-driven conversation generation in social media. *Complex.*, 2020:3710104:1–3710104:13.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *ACL*.

Moya Chen, Paul A. Crook, and Stephen Roller. 2021. Teaching models new apis: Domain-agnostic simulators for task oriented dialogue.

Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. 1998. Evaluation metrics for language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, abs/1811.01241.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking-training with large-scale human feedback data. In *EMNLP*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

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.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *ACL*.

San Kim, Jin Yea Jang, Minyoung Jung, and Saim Shin. 2021. A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems. In *EMNLP*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation.

Andrew Kyle Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *ArXiv*, abs/2204.02329.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021. A three-stage learning framework for low-resource knowledge-grounded dialogue generation. In *EMNLP*.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? 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 851–864, Online. Association for Computational Linguistics.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. *ArXiv*, abs/2004.07672.

Hsuan Su, Pohan Chi, Shih-Cheng Huang, Chung Ho Lam, Saurav Sahay, Shang-Tse Chen, and Hung yi Lee. 2022. Few-shot prompting towards controllable response generation.

Chongyang Tao, Changyu Chen, Jiazhan Feng, Ji rong Wen, and Rui Yan. 2021. A pre-training strategy for zero-resource response selection in knowledge-grounded conversations. In *ACL*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

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.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *ArXiv*, abs/2109.01247.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. 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. *CoRR*, abs/1901.08149.

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.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020a. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020b. A theory of usable information under computational constraints. *ArXiv*, abs/2002.10689.

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, Melbourne, Australia. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

A Appendix

A.1 Experimental Details

- **Hyperparameters**: For the Hyperparameters we use to conduct experiments, we follow TransferTransfo link `https://github.com/huggingface/transfer-learning-conv-ai`. They obtain these Hyperparameters by grid searching. More specifically, They finetuned the model with a batch size of 32 sequences , and finetune the models approximately 2 epochs over training dataset. They used Adam with a learning rate of 6.25e-5, and a coefficient of 2 on the LM loss when summing with the next-sentence prediction loss . The learning rate was linearly decayed to zero over the course of the training.

- **Datasets**: The link to download Persona-Chat `https://parl.ai/docs/tasks.html#persona-chat` and the train/valid/test split is 9907/1000/968 dialogues.. For the link to download Topical-Chat `https://github.com/alexa/Topical-Chat` and the train/valid/test split is 8628/1078/1078 dialogues.

- **Pretrained Models**: For GPT model we use gpt-medium as our pretrain model and use microsoft/DialoGPT-medium as initial checkpoint for GPT-2 model.

A.2 Evaluation Metrics

In addition to entailment, we aimed to employ other metrics that are also important to measure a dialogue system.

Perplexity (Chen et al., 1998): Here we employed the pretrained GPT-2 language model GPT to judge if the output sentence $C(x)$ was an acceptable sentence. The computation of Perplexity (Chen et al., 1998) is shown below.

$$PPL = \prod_{i=1}^{T} \frac{1}{(GPT(C(x,D)_i|x))^{1/T}} \tag{1}$$

Coherence: We employed the DialogRPT (Gao et al., 2020) to calculate the coherence between conversation model's output and the input context. DialogRPT (Gao et al., 2020) is a GPT2-based ranker that finetuned on 133M human feedback data. With the contrastive learning approach that DialogRPT used. The ranker has better understanding on how relevant the response is for the given context. In our evaluation, we take the the probability that output by DialogRPT coherence model (*human_vs_rand*) as our coherence metric.

Diversity: BLEU score (Papineni et al., 2002) is a commonly used metric for automatically evaluating machine translation. However, the Self-BLEU (Zhu et al., 2018) score here was applied to measure the diversity of chatbot responses. Regarding one sentence as the prediction and the others as the reference, we can calculate BLEU score for every sentence, and the average is the Self-BLEU score. A lower Self-BLEU score implies more diversity of the chatbot responses.

A.3 Standard Deviation of Quality Metrics

Model	Method	Persona		Topical	
		TT.	LM.	TT.	LM.
		Perplexity			
GPT	Origin	0.23	0.27	0.20	0.25
	Multi Pos	0.22	0.26	0.27	0.22
GPT-2	Origin	0.31	0.29	0.120	0.09
	Multi Pos	0.28	0.23	0.10	0.110
		Coherence			
GPT	Origin	0.001	0.001	0.002	0.002
	Multi Pos	0.001	0.001	0.002	0.002
GPT-2	Origin	0.002	0.001	0.001	0.001
	Multi Pos	0.001	0.001	0.001	0.001
		Diverstiy			
GPT	Origin	0.002	0.002	0.002	0.002
	Multi Pos	0.002	0.002	0.002	0.002
GPT-2	Origin	0.002	0.002	0.002	0.002
	Multi Pos	0.002	0.002	0.002	0.001

Table 2: The results of quality measurements. The standard deviation across 50 runs are reported.