Attribute Controlled Dialogue Prompting

Runcheng Liu[1,2*], **Ahmad Rashid**[1,2*], **Ivan Kobyzev**[3]
Mehdi Rezagholizadeh[3], **Pascal Poupart**[1,2]
[1]David R. Cheriton School of Computer Science, University of Waterloo
[2]Vector Institute, Canada
[3]Huawei Noah's Ark Lab, Canada
{ireneliu,a9rashid,ppoupart}@uwaterloo.ca
{ivan.kobyzev,mehdi.rezagholizadeh}@huawei.com

Abstract

Prompt-tuning has become an increasingly popular parameter-efficient method for adapting large pretrained language models to downstream tasks. However, both discrete prompting and continuous prompting assume fixed prompts for all data samples within a task, neglecting the fact that inputs vary greatly in some tasks such as open-domain dialogue generation. In this paper, we present a novel, instance-specific prompt-tuning algorithm for dialogue generation. Specifically, we generate prompts based on instance-level control code, rather than the conversation history, to explore their impact on controlled dialogue generation. Experiments on popular open-domain dialogue datasets, evaluated on both automated metrics and human evaluation, demonstrate that our method is superior to prompting baselines and comparable to fine-tuning with only 5%-6% of total parameters.

1 Introduction

Fine-tuning has been frequently used when deploying generative pretrained language models (PLMs) to downstream tasks since the advent of GPT (Radford et al.) and BERT (Devlin et al., 2019). However, this requires storing a full copy of parameter states for every downstream task, which is memory-consuming and expensive to serve when working with large-scale models with billions of parameters like GPT-3 (Brown et al., 2020).

In this work, we design a lightweight prompting module for adapting pretrained language models for attribute controlled dialogue generation. More precisely, for each attribute such as persona, intention, emotion etc. we only save an additional prompt module. Since the prompting module is a fraction of the size of the pretrained dialogue model, this allows many controlled dialogue systems to be stored on a device without too much overhead. We present results on both intent and persona controlled dialogue.

2 Related Work

GPT-3 (Brown et al., 2020) introduces *prompting*, a method to steer a frozen PLM by transforming inputs into cloze-style phrases with task description and some task examples. Though it is memory-efficient since one single copy of the PLM can be shared across different tasks, the model's performance is largely restricted by the maximum conditional input length, the model size and manual guesswork for prompts (Zhao et al., 2021; Schick and Schütze, 2021a,b; Jiang et al., 2020). Other works focus on automatically searching for better discrete prompts (Jiang et al., 2020; Shin et al., 2020; Gao et al., 2021; Ben-David et al., 2021).

Recently, there has been an increased interest in *continuous prompts / prompt-tuning*, which bridges the gap between prompting and fine-tuning, while remaining efficient during training (Lester et al., 2021; Li and Liang, 2021; Liu et al., 2021, 2022). Continuous prompts extend prompt selection to the entire space of embeddings, including vector embeddings that do not correspond to any human-interpretable natural language tokens. Hence, soft prompts are more expressive than discrete prompts.

However, both deep prompts and shallow prompts assume a *static prompt / task-level prompt* for all samples within a task, neglecting the fact that samples might vary greatly, especially in the field of conversation generation. There are recent papers exploring possible *instance-specific prompts*. For instance, Control-prefixes (Clive et al., 2021) generates attribute-level prompts for input labels, but its expressiveness is limited to four labels. IPL (Jin et al., 2022) includes a look-up module to reweight prompt tokens before passing the updated embedding-only prompt into the transformer, but IPL updates all model parameters, which loses the efficiency benefits of prompt-

[*] Work done during an internship at Huawei.

Findings of the Association for Computational Linguistics: ACL 2023, pages 2380–2389
July 9-14, 2023 ©2023 Association for Computational Linguistics

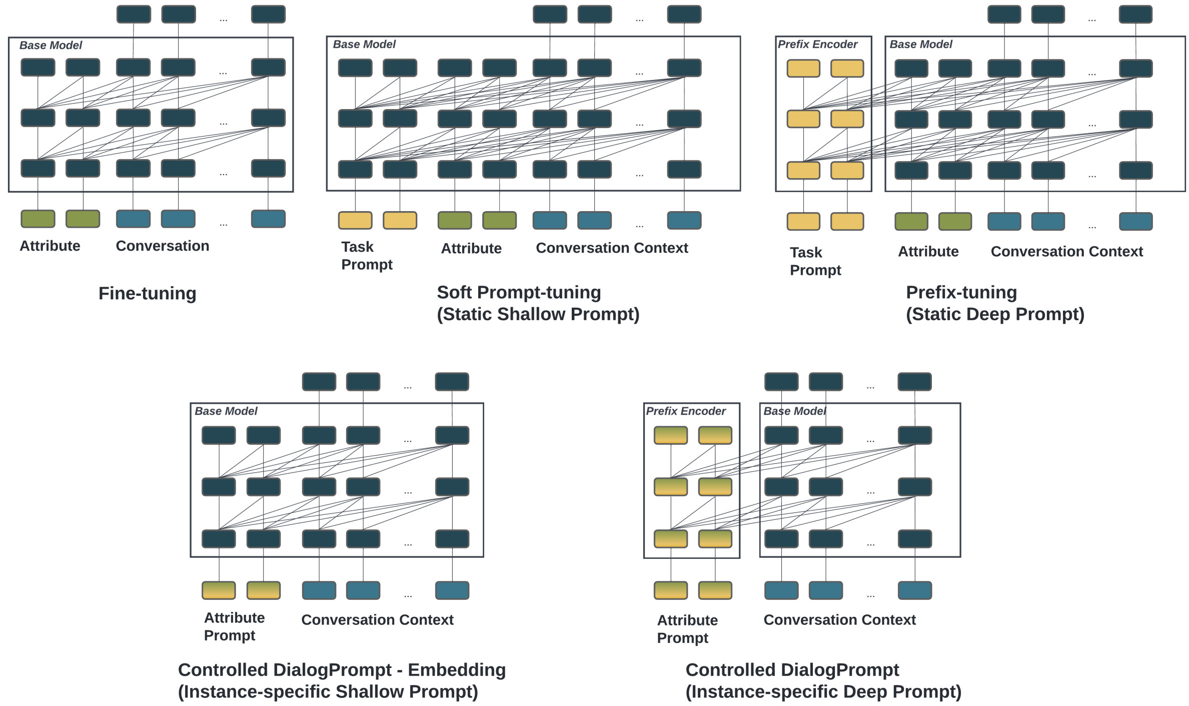

Figure 1: Diagrams illustrating attention mechanisms for different configurations of (task/attribute) prompt, attribute and conversation context.

ing. IDPG (Wu et al., 2022) consumes inputs in a two-layer perceptron module to generate instance-dependent prompts in classification tasks rather than generation tasks. In addition, (Gu et al., 2021) proposes DialogPrompt which performs instance-specific prompting for dialogue generation by conditioning the prompt on the entire dialogue history. However, their prompting module consists of GPT-2, which is a full-fledged language model, and the approach is as costly as storing an entire fine-tuned base model. Recent works Contrastive prefixes (Qian et al., 2022) and Tailor (Yang et al., 2022) both propose *attribute-based prompts*, instead of instance-specific, to include either single-attribute or multi-attribute prompts into controlled text generation tasks, which reveal the powerful potential of controllability of continuous prompts.

In contrast to previous work, we propose Controlled DialogPrompt for applying prompt-tuning in controlled dialogue generation, which optimizes prompts based on provided control codes rather than the previous conversation history and we further explore the controllability of prompts at the instance level. The size of the prompt encoder is strictly limited and we freeze the pretrained transformer during training in order to preserve memory efficiency. In addition, we would like to highlight that our work focuses more on open-ended text generation rather than natural language understanding, such as entailment, paraphrase detection, extractive QA, as seen in other parameter-efficient fine-tuning methods (He et al., 2022; Guo et al., 2021; Wu et al., 2022). We posit that generating high-quality text is a more challenging task that requires a more nuanced approach to prompt tuning.

3 Controlled DialogPrompt

In this section, we present Controlled DialogPrompt (Controlled DP) for dialogue generation, which is expected to provide attribute information such as the dialogue intention or the user's persona within the prompt and steer the pretrained model efficiently.

Soft Prompt-tuning (Lester et al., 2021; Liu et al., 2021) learns soft tokens for different tasks and then prepends them to the conversation context as well as control attributes. This approach yields a *static shallow* prompt since the soft tokens are static (i.e., fixed for a task) and shallow (only added as an input to the language model).

In contrast, Prefix-tuning proposes a more effective technique that adds soft tokens in the form of key-value pairs at every attention block of the transformer (Li and Liang, 2021; Liu et al., 2022).

This allows the soft tokens to influence each stage of the language model and therefore it is referred to as a *static deep* prompt.

Figure 1(bottom right) shows our proposed controlled dialogue prompt (Deep version). Instead of training static soft tokens for the dialogue task, we train a lightweight prompt module that takes as input a control attribute, either an intention label or persona sentences, and outputs key-value pairs that are prepended to each layer of the language model. Since the soft token embeddings change depending on the control attribute, this corresponds to an *instance-specific* prompt. For the shallow prompt (Figure 1 bottom left), we follow Soft Prompt-tuning which adds an additional trainable embedding layer to encode the attribute. For the deep prompt module, we consider two architectures: i) a simple multilayer perceptron (two fully connected layers of size 512 with tanh activation) applied to each token of the control attribute, and ii) a two-layer transformer decoder with embedding size of 256. The embedding size of each architecture was chosen to yield roughly the same number of parameters. This number of parameters is about 5%-6% of the number of parameters of the language model. For a given domain, training the prompt module is done as follows. An intention label or persona sentences are fed to the prompting module, which outputs key-value pairs added at each layer of the frozen pretrained dialogue system. Gradients to maximize the likelihood of response tokens are back-propagated through the dialogue system and prompting module, but only the weights of the prompting module are updated.

4 Experiments

4.1 Datasets and baseline models

We evaluate the proposed method on two publicly available datasets: Dailydialog (Li et al.) for label control and FoCus (Jang et al., 2021) for document control. Dailydialog (Li et al.) is a widely used daily conversation dataset that provides a dialogue act for every sentence that indicates the communication function of each utterance. There are 4 types of dialogue acts in total. FoCus(Jang et al., 2021) is a new persona-grounded dataset that aims to provide informative answers based on the user's persona about the geographical landmark. We provide the detailed dataset setups in Appendix A.1.

To demonstrate better performance of Controlled DialogPrompt, we compare our model with other competitive prompt-tuning techniques. The backbone model is DialoGPT-Large (Zhang et al., 2020). Details are provided in Appendix A.2.

4.2 Evaluation Methods

We use both automatic evaluation metrics and human evaluation to measure the performance.

Automated metrics For controllability, we follow (Du and Ji, 2021) to evaluate whether models can customize responses based on specified control attributes. Details about controllability measures are provided in Appendix B.1 Regarding response quality, we use n-gram based metrics such as BLEU (B-2, B-4) (Papineni et al., 2002), NIST (N-2, N-4) (Doddington, 2002), ROUGE-L (Lin, 2004), METEOR (Agarwal and Lavie, 2007) to evaluate fluency and adequacy and distinct n-gram distribution metrics such as Dist (D-1, D-2) (Li et al., 2016) and Entropy (E-4) (Zhang et al., 2018) to measure the diversity of the response.

Human Evaluation Human evaluation on the other hand is used to measure consistency between dialogue context and response and attribute controllability. We adopt single-turn pairwise evaluations to prevent annotator bias in numerical score evaluation. Details on question settings and annotators are provided in Appendix B.2

5 Result and Analysis

5.1 DialogAct / Intention

Table 1 summarizes the automatic evaluation results on the DialogAct label control task. Compared to static task prompts, instance-level controlled prompts achieve better performance consistently on both deep and shallow prompt levels. Since the controlled attribute is injected independently through the prompts, it does not affect the understanding and generation ability of the pretrained transformer. Both Controlled DP deep methods show higher controllability and response quality than Controlled DP embedding, in line with (Li and Liang, 2021; Liu et al., 2022; Qin and Eisner, 2021) indicating the expressiveness of

[1] Controlled DP (Embedding) involves training an embedding layer in a size of (prompt_vocab_size * base_model_n_embd). In DialogAct control, we use only 4 labels, resulting in a size of 4 * 1280. In User's Persona, since there are many words in the corpus, we adopt the base model vocab size as the prompt vocab size and the embedding layer is 50257 * 1280. Therefore, the proportion of tunable parameters is higher in User's Persona Control.

Method	ϕ%	**Controllability**	**BLEU ↑**		**NIST ↑**		**ROUGE-L ↑**	**METEOR ↑**	**Dist ↑**		**Entropy ↑**
		Accuracy	B-2	B-4	N-2	N-4			D-1	D-2	E-4
Pretrained (Zhang et al., 2020)	0%	58.30%	10.31%	1.73%	0.18	0.18	19.43%	7.30%	**7.61%**	**40.00%**	10.03
Fine-tuning	100%	**80.25%**	**21.03%**	**5.70%**	0.96	0.98	**34.38%**	**13.05%**	6.02%	34.51%	10.21
Soft Prompt-tuning (Lester et al., 2021)	0.008%	70.51%	18.15%	4.08%	0.56	0.57	31.58%	11.46%	5.33%	30.82%	10.02
Prefix-tuning (Li and Liang, 2021)	3.1%	75.02%	19.94%	5.12%	0.91	0.93	33.29%	12.54%	5.59%	32.46%	10.17
Controlled DialogPrompt (Embedding)	0.001%[1]	69.06%	**20.11%**	4.91%	0.71	0.73	32.80%	12.19%	5.18%	30.07%	10.03
Controlled DialogPrompt (MLP)	3.1%	78.36%	19.92%	**5.43%**	0.98	1.01	33.12%	12.61%	5.71%	32.42%	10.20
Controlled DialogPrompt (2-layer Transformer)	3.3%	**78.58%**	19.86%	5.26%	**1.01**	**1.04**	**33.35%**	**12.64%**	**5.82%**	**33.16%**	**10.23**

Table 1: **DialogAct label** control performance under Dailydialog multi-reference evaluation. $\phi\%$ denotes the $\%$ of tunable parameters to the frozen-LM parameters required at training time. Red number is the best value in every metric on all methods. Blue number is the best value in every metric among prompting methods.

Method	ϕ%	**Controllability**	**BLEU ↑**		**NIST ↑**		**ROUGE-L ↑**	**METEOR ↑**	**Dist ↑**		**Entropy ↑**
		Similarity	B-2	B-4	N-2	N-4			D-1	D-2	E-4
Pretrained (Zhang et al., 2020)	0%	51.40%	1.63%	0.42%	0.02	0.02	6.62%	3.67%	7.62%	34.44%	10.15
Fine-tuning	100%	**75.21%**	**37.38%**	**25.77%**	**5.80**	**6.30**	**27.71%**	**24.43%**	**7.93%**	**38.20%**	**11.28**
Soft Prompt-tuning (Lester et al., 2021)	0.008%	62.69%	18.01%	9.50%	2.72	2.87	16.53%	13.29%	6.77%	32.19%	10.96
Prefix-tuning (Li and Liang, 2021)	6.2%	**66.89%**	27.18%	16.73%	4.35	4.63	21.38%	18.56%	7.60%	36.88%	**11.25**
Controlled DialogPrompt (Embedding)	8.3%[1]	61.16%	13.01%	5.12%	1.89	1.96	14.84%	10.28%	5.21%	26.45%	10.82
Controlled DialogPrompt (MLP)	6.2%	64.96%	26.82%	17.09%	4.25	4.54	21.40%	18.47%	7.85%	37.58%	11.18
Controlled DialogPrompt (2-layer Transformer)	5.0%	66.34%	**31.85%**	**21.67%**	**5.00**	**5.40**	**24.20%**	**21.16%**	**7.85%**	**37.86%**	11.24

Table 2: **User's Persona** control performance under FoCus validation dataset. $\phi\%$ denotes the $\%$ of tunable parameters to the frozen-LM parameters required at training time. Red number is the best value in every metric on all methods. Blue number is the best value in every metric among prompting methods.

Methods	Attribute Relevancy	Consistency
Controlled DP (Deep)	30.7%	32.0%
Soft Prompt-tuning	20.0%	20.0%
Neutral	49.3%	48.0%
Controlled DP (Deep)	25.3%	37.3%
Prefix-tuning	16.0%	16.0%
Neutral	58.7%	46.7%
Controlled DP (Deep)	34.7%	38.7%
Controlled DP (Shallow)	9.3%	25.3%
Neutral	56.0%	36.0%

Table 3: Human evaluation on Dailydialog dataset. "Controlled DP (Deep)" represents Controlled DialogPrompt with 2-layer transformer decoder as the prompt module. "Controlled DP (Shallow)" represents Controlled DialogPrompt on the embedding layer. “Neutral” means that there is no preference between the two answers according to the annotators.

Methods	Persona Controllability	Consistency
Controlled DP (Deep)	41.3%	44.0%
Soft Prompt-tuning	5.3%	13.3%
Neutral	53.3%	42.7%
Controlled DP (Deep)	22.7%	28.0%
Prefix-tuning	26.7%	8.0%
Neutral	50.7%	64.0%
Controlled DP (Deep)	29.3%	41.3%
Controlled DP (Shallow)	21.3%	9.3%
Neutral	49.3%	49.3%

Table 4: Human evaluation on Focus dataset. "Controlled DP (Deep)" represents Controlled DialogPrompt with 2-layer transformer decoder as the prompt module. "Controlled DP (Shallow)" represents Controlled DialogPrompt on the embedding layer. “Neutral” means that there is no preference between the two answers according to the annotators.

deep prompts. Also, Controlled DP deep methods show performance close to fine-tuning and even outperform on some metrics such as NIST. This is because NIST is weighted-BLEU with higher weights on rarer words and fine-tuning tends to generate from a more limited vocabulary whereas Controlled DialogPrompt sometimes generates less frequent words and can attain a better NIST score. Human evaluation (Table 3) also shows that Controlled DP deep has a significantly higher winning rate than other prompting techniques on both control attribute relevancy and conversation consistency.

5.2 User's Persona

Table 2 shows that our model displays advantages over other prompting methods in terms of response quality, which shows a promising sign that controlled DP can be adapted to more challenging document control scenarios. Note that the difference in BLEU-2 is more pronounced for Focus compared to DailyDialog, as Focus is more complicated and uses sentences as the attribute rather than labels. Although controlled DP methods perform slightly lower than Prefix-tuning on the similarity scores with given user's persona and Entropy-4 values, we find it to be highly consistent with the previous conversation history upon human evaluation (Table 4). Similar results are observed with FoCus (Jang et al.,

2021) where models with high generation abilities do not always ensure high grounding abilities. In addition, the difference between static/instance-specific deep prompts and static/instance-specific shallow prompts emphasizes the direct impact of deep prompts in complex tasks. Fine-tuning performs the best, but with approximately $20X$ more tunable parameters.

6 Conclusion and Future Work

In summary, we presented a novel prompting technique, conditioned on a dialogue attribute (persona or intent), for controlled dialogue generation. The prompting module requires only 5%-6% of the total number of parameters, which allows the storage of several fined-tuned prompting modules for different dialogue generation tasks at a fraction of the cost of a full dialogue model.

However, Controlled DialogPrompt currently studies conditioning on simple control attribute sentences like the user's persona and the work can be extended to more extensive and complex sentences such as background knowledge documents to further evaluate the controlled prompt's encoding capabilities. Additionally, combining multiple Controlled DialogPrompts on several control attributes and automatically triggering various dialogue skills is an interesting and unexplored direction.

Limitations

In our current experiments, prompt-based methods are primarily storage-efficient or parameter-efficient solutions. Since these methods all require backpropagation to the bottom layer, the training time of prompt-based methods are closely resembles that of traditional fine-tuning approach.

Acknowledgements

This research was funded by Huawei Canada and the National Sciences and Engineering Research Council of Canada. Resources used in preparing this research at the University of Waterloo were provided by the province of Ontario and the government of Canada through CIFAR and companies sponsoring the Vector Institute.

References

Abhaya Agarwal and Alon Lavie. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of WMT-08*.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2021. Pada: A prompt-based autoregressive approach for adaptation to unseen domains. *arXiv preprint arXiv:2102.12206*.

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

Jordan Clive, Kris Cao, and Marek Rei. 2021. Control prefixes for text generation. *arXiv preprint arXiv:2110.08329*.

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.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Wanyu Du and Yangfeng Ji. 2021. Sidecontrol: Controlled open-domain dialogue generation via additive side networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2175–2194.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830.

Xiaodong Gu, Kang Min Yoo, and Sang-Woo Lee. 2021. Response generation with context-aware prompt learning. *arXiv preprint arXiv:2111.02643*.

Demi Guo, Alexander M Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. 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 4884–4896.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391.

Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, et al. 2022. Hyperprompt: Prompt-based task-conditioning of transformers. In *International Conference on Machine Learning*, pages 8678–8690. PMLR.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2021. Call for customized conversation: Customized conversation grounding persona and knowledge. *arXiv preprint arXiv:2112.08619*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Feihu Jin, Jinliang Lu, Jiajun Zhang, and Chengqing Zong. 2022. Instance-aware prompt learning for language understanding and generation. *arXiv preprint arXiv:2201.07126*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

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

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

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

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.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212.

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

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Rui Hou, Yuxiao Dong, VG Vydiswaran, and Hao Ma. 2022. Idpg: An instance-dependent prompt generation method. *arXiv preprint arXiv:2204.04497*.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2022. Tailor: A prompt-based approach to attribute-based controlled text generation. *arXiv preprint arXiv:2204.13362*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

A Experimental Setups

A.1 Datasets

A.1.1 Label control

Dailydialog (Li et al.) is a widely used daily conversation dataset that provides a dialogue act for every sentence. Dialogue acts indicate the communication function of each utterance and there are 4 types of dialogue acts: inform, questions, directives, and commissives. We follow the standard split of the original Dailydialog dataset, limit the conversation context to a maximum of four sentences, and remove any sentence that has more than 25 words to maintain computation efficiency. As a result, we obtain 61,669 training samples, 5769 validation samples, and 5453 testing samples.

We additionally use the Dailydialog multi-reference dataset from (Gupta et al., 2019) during metrics computation to mitigate the one-to-many possible response problem.

A.1.2 Document control

FoCus(Jang et al., 2021) is a persona-grounded dataset. Unlike DailyDialog, FoCus aims to build a dialogue agent that provides informative answers based on the user's persona about the geographical landmark; therefore, it is more content-rich and challenging. The selected knowledge candidate sentence is prepended to the conversation and regarded as part of the input.

The input to the base model has the template: "*Knowledge: [Selected knowledge sentence] Conversation: [Previous utterances]*". The persona sentences are given as the input to the prompt encoder. In fine-tuning (no prompt encoder) and static prompt methods (the prompt encoder does not take attribute information), the persona sentences are concatenated together with the knowledge and previous utterances and form the input to base model as "*Knowledge: [Selected knowledge sentence] Persona: [User's Personas] Conversation: [Previous utterances]*"

Since the grounded answer of the test set has not been released, we shuffle and split the original training set to construct our training samples and validation samples (70% training and 30% validation) and the original validation set as our testing samples. We further restrict conversation context to at most three sentences because the bot's utterances are much longer than human's utterances. In total, we have 49,198 samples for training, 21,134 samples for validation, and 5,639 samples for testing.

A.2 Baseline models

To demonstrate better performance of Controlled DialogPrompt, we compare our model with other competitive prompt-tuning techniques.

- **Pretrained DialoGPT** (Zhang et al., 2020): DialoGPT-large has shown its superiority for a wide range of open-domain dialogue generation tasks by pretraining on a massive corpus.
- **Fine-tuning**: Fine-tuning, though memory-consuming, is the most straightforward and prevalent adaptation technique to downstream tasks. Fine-tuning has been considered as the benchmark for all light-weight fine-tuning methods including prompt-tuning.
- **Soft Prompt-tuning (static shallow prompt)** (Lester et al., 2021): The method applies a static task prompt to the embedding of every input. We experiment with different lengths (length 10 and length 50) of the static shallow prompt and use the better length 50.
- **Prefix-tuning (static deep prompt)** (Li and Liang, 2021): Prefix prompts are added to every layer during computation. We experiment with different lengths (length 10 and length 50) and we report the better prompt result with length 10.
- **Controlled DP - Embedding (instance-specific shallow prompt)**: The shallow version of our method with controlled prompts added only in the embedding layer. It is used to demonstrate the expressiveness of the deep Controlled DialogPrompt.

- **Controlled DP - MLP / 2-layer Transformer (instance-specific deep prompt)**: We explore different prompt encoder structures, among which MLP prompt encoder shares the frozen pretrained transformer embedding layer to reduce tunable parameters.

During our experiments, we utilize DialoGPT-large as the frozen backbone model and train all models on two Nvidia V100 32G GPUs. We train models for 10 epochs with training batch size 2 per GPU and learning rate of 1e-4 except for fine-tuning, which is set to 5e-5 in the FoCus dataset and 1e-5 in the Dailydialog dataset. Models that achieve the lowest validation losses are saved during the training. We perform optimization with the AdamW optimizer with maximum gradient clipping set to 1. For decoding, we choose top-k sampling provided in Huggingface where k=10 and temperature T=0.9. The result is generated with random seed=42.

B Evaluation Methods

B.1 Automated metrics

For controllability, we follow (Du and Ji, 2021) to evaluate whether models can customize responses based on specified control attributes. (1) For label control, we fine tune an independent BERT classifier (Devlin et al., 2019) which can take a sentence and predict its dialogue intention. We train the classifier on the same training set and achieve 83.23% accuracy on the test set. (2) For document control, we also compute the cosine similarity between the Glove embedding of the generated responses and grounded persona documents. As FoCus dataset contains human-annotated labels for used persona sentences, only those that are actually used are evaluated. Detailed training information is provided in (Du and Ji, 2021).

Regarding response quality, we utilize different variants of n-gram based metrics such as BLEU (B-2, B-4) (Papineni et al., 2002), NIST (N-2, N-4) (Doddington, 2002), ROUGE-L (Lin, 2004), METEOR (Agarwal and Lavie, 2007) to evaluate fluency and adequacy and distinct n-gram distribution metrics such as Dist (D-1, D-2) (Li et al., 2016) and Entropy (E-4) (Zhang et al., 2018) to measure the diversity of the response. We follow the metrics setting in (Zhang et al., 2020).

B.2 Human Evaluation

Human evaluation on the other hand is used to measure consistency between dialogue context and response and attribute controllability. Similar to ACUTE-Eval in (Li et al., 2019; Roller et al., 2021), we adopt single-turn pairwise evaluations to prevent annotator bias in numerical score evaluation. We compare Controlled DialogPrompt with every other prompt-tuning methods, covering static shallow prompt, static deep prompt and instance-specific shallow prompt. In each comparison group, there are two questions designed separately to assess response's dialogact/personality controllability as well as consistency to the previous conversation context. For dialogact controllability, we have the question: *Which response do you think is more related to the given dialog act (intention)?*. For personality controllability, we set the question as *Which response do you think is more related to the personality?*. For the consistency to the previous conversation context, we set the question as *Which response do you think is more consistent to the above conversation context?* We sample 15 conversations from each comparison group and there are 5 conversations overlapped across different groups. Annotators are industrial NLP researchers and NLP graduate students. We collected 900 annotations in total.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

□ 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?
Section 1

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

B ☑ Did you use or create scientific artifacts?

Code from pretrained model and evaluation metrics; Pretrained models;

☑ B1. Did you cite the creators of artifacts you used?
Section 4 and Appendix

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4 and Appendix

☑ 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)?
We mentioned we use the datasets and models following the existing papers. Section 4 and Appendix.

□ 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.?
Appendix A

☑ 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.
Appendix A.1

C ☑ Did you run computational experiments?

Section 4 and Section 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?
Section 5 and table result

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?
Appendix A.2

☑ 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?
Appedix A.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.)?
Appendix B.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4.2

☑ 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.?
Appendix B.2

☑ 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)?
Appendix B.2

☑ 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?
Appendix B.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.