Building a Personalized Dialogue System with Prompt-Tuning

Tomohito Kasahara[1], **Daisuke Kawahara**[1],
Nguyen Tung[2], **Shengzhe Li**[2], **Kenta Shinzato**[2], **Toshinori Sato**[2]
[1]Waseda University, [2]LINE Corporation
{tomo_k@ruri.,dkw@}waseda.jp
{tung.nguyen,shengzhe.li,kenta.shinzato,toshinori.sato}@linecorp.com

Abstract

Dialogue systems without consistent responses are not fascinating. In this study, we build a dialogue system that can respond based on a given character setting (persona) to bring consistency. Considering the trend of the rapidly increasing scale of language models, we propose an approach that uses prompt-tuning, which has low learning costs, on pre-trained large-scale language models. The results of automatic and manual evaluations in English and Japanese show that it is possible to build a dialogue system with more natural and personalized responses using less computational resources than fine-tuning.

1 Introduction

Large dialogue corpora used to train dialogue systems using neural network models contain utterances from various speakers. This has the disadvantage that the trained system is often inconsistent in the generated utterances (Li et al., 2016b). For example, after the system says, “I am from Tokyo,” it might say, “I am from Kyoto.”

We aim to build a dialogue system that can respond based on a persona to avoid inconsistent utterances. A simple method of giving a persona to a model can be to concatenate the persona to the model’s input in natural language (Zhang et al., 2018). However, this method is not suitable because the more persona information is added, the longer the input text becomes. Therefore, we propose to freeze all parameters of a pre-trained language model and add a new fixed-length prompt before the input token sequence to embed the persona information. Specifically, only the embedding vectors of the added prompt are optimized using a dialogue corpus in which utterances are made based on the persona.

We conduct experiments on two languages: English and Japanese. Automatic and manual evaluations show that our method can build a dialogue system capable of natural responses based on a persona. Since our approach does not update the parameters of the pre-trained model, it can reduce the computational cost required for training. We also show that it is possible to build a personalized dialogue system with even a small dataset consisting of hundreds to thousands of utterance-response pairs.

2 Related Work

2.1 Prompt-Tuning

With the advent of pre-trained models such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020), a method that adapts a pre-trained model to a target task by fine-tuning has become mainstream. However, as the scale of models grows and the cost of fine-tuning increases, methods for adapting a pre-trained model to a target task without updating their parameters are gaining attention.

Brown et al. (2020) proposed a zero/few-shot learning method based on language models with manually created task descriptions and zero/a few task examples (collectively called *prompt*). Although there are some studies on improving this method (Reynolds and McDonell, 2021; Zhao et al., 2021), they are inferior to fine-tuning in terms of accuracy.

Prompt-tuning is a method for automatically optimizing a prompt without creating it by manual labor. There are two kinds of methods in prompt-tuning: one is to select the best words from a discrete vocabulary (Shin et al., 2020), and the other is to optimize continuous embedding vectors (Qin and Eisner, 2021; Li and Liang, 2021; Lester et al., 2021; Liu et al., 2021; Vu et al., 2021). Prefix-tuning (Lester et al., 2021; Li and Liang, 2021) adds a sequence of tokens, called prefix tokens, to the beginning of the input and optimizes only their embedding vectors. There is also a study on multimodal prompt-tuning for images and natural

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 96 - 105
July 10-15, 2022 ©2022 Association for Computational Linguistics

language (Tsimpoukelli et al., 2021).

2.2 Persona-Based Dialogue Systems

According to Roller et al. (2021), for dialogue systems to interact more naturally with humans, it is essential to consider three perspectives: having a consistent personality, having knowledge, and having emotions and empathy for the interlocutor. Among these three perspectives, we focus on personality because we believe that it is the most important to generate consistent responses.

The Persona-Chat dataset (Zhang et al., 2018) is a dataset created with the goal of adding personality to a dialogue system. It consists of multi-turn dialogues between two crowdworkers, each of whom is given approximately five persona sentences, which describe their character settings. There are 1,155 personas in the Persona-Chat dataset. There are two types of persona sentences per persona: *original*, which the worker used in the dialogue, and *revised*, which is a paraphrased version of the original. In the experiments conducted by Zhang et al. (2018), models were trained using all the data in the Persona-Chat dataset, which contains utterances based on various personas. On the other hand, our method uses dialogue data uttered based on only one persona to train models. There is also a Japanese version of the Persona-Chat dataset, JPersonaChat (Sugiyama et al., 2021). Other dialogue corpora that contain speaker persona information include PersonalDialog (Zheng et al., 2019) and a corpus of dialogue data from Reddit (Mazaré et al., 2018). Zheng et al. (2019) proposed a method to add encoded persona information to the input before it is fed into a seq2seq model.

3 Method

This section describes our proposed method. The detailed setup for our experiments is described in Sections 4.1 and 4.2.

3.1 Proposed Model

We propose a Transformer-based model with an additional embedding layer for tokens that embed persona information. We refer to these tokens as *persona info tokens*. The architecture and input-output relation of the proposed model are shown in Figure 1.

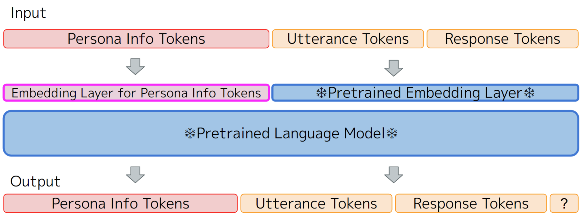


Figure 1: Architecture and input-output relation of the proposed model. All parameters of the pre-trained language model and its embedding layer are frozen. Only the newly added embedding layer for persona info tokens is tuned.

3.2 Datasets

Conversations in daily life are not always related to personal information (Song et al., 2021). To allow the model to generate not only utterances that are related to the persona but also utterances that are not related to the persona, we make a dialogue dataset that consists of two types of datasets. The first is a dialogue dataset where each utterance is based on the persona, and the second is a dialogue dataset that is not related to the persona.

3.3 Training

The newly added embedding layer embeds persona info tokens, and the embedding layer of the pre-trained language model embeds each pair of utterance and response (which consists of tokens already generated during training). These embedding vectors are combined and then input into the model. During training, the cross-entropy loss is calculated for the output tokens of the response sentence, and only the parameters of the embedding layer for the persona info tokens are updated.

The embedding layer for the persona info tokens is initialized with the persona sentences included in the Persona-Chat dataset. These sentences are embedded into vectors by the embedding layer of the pre-trained language model and then used for initialization. If the number of the tokens of the persona sentences is less than the length of the persona info tokens, the persona sentences are repeatedly arranged until the number is satisfied.

4 Experiments

Based on the method in Section 3, we build a personalized dialogue system. We used Hugging Face's Transformers to build the system and the NVIDIA A100 SXM4 GPU with a GPU memory size of 40 GB. The main experiments are conducted in English, and the results of additional experiments in Japanese are included at the end of this section.

4.1 Datasets Setup

We use the Persona-Chat dataset[1] and DailyDialog (Li et al., 2017)[2] for our experiments in English.

4.1.1 Training Datasets

First, the multi-turn dialogues in the Persona-Chat dataset are divided into two utterances of one round trip. We refer to this pair of two utterances as *a dialogue pair*. The dialogue pairs are aggregated according to the persona type given to the responder. There are 1,155 personas in the Persona-Chat dataset, but we use the three personas with the most dialogue pairs in our experiments. The reason for this is that we intend to experiment with a relatively large number of dialogue pairs even in the small dataset. The number of dialogue pairs based on these three personas is 185, 167, and 166, respectively. Three models corresponding to the three personas are trained and evaluated for each experimental setup. The aggregated dialogue pairs are divided into training and evaluation pairs in a ratio of 9:1.

The Persona-Chat dataset does not contain many short utterances or utterances unrelated to persona. To add utterances that are short and not related to persona to the dataset, we also use dialogue pairs contained in DailyDialog whose topic is Relationship,[3] which contains many such utterances. Among them, dialogue pairs in which the lengths of both the utterance and the response are less than 50 characters are mixed into the training datasets in a certain ratio. Based on the results of preliminary experiments, we determined the ratio of dialogue pairs added from DailyDialog to the number of those obtained from the Persona-Chat dataset as 1:1. We call this *the ratio of the training datasets*.

4.1.2 Evaluation Datasets

We made two datasets for evaluation: the persona eval dataset and the general eval dataset. The persona eval dataset is 10% of the 9:1 dataset described in Section 4.1.1. The general eval dataset consists of dialogue pairs obtained from DailyDialog under the same conditions as in Section 4.1.1, but not used for training.

Training Method	Model	Dist-1	Dist-2
Fine-Tuning (added)	GPT2-XL	0.199	0.526
Fine-Tuning (none)		0.210	0.568
Prompt-Tuning		0.177	0.494
	GPT-J-6B	**0.213**	**0.595**

Table 1: Results of automatic evaluation by distinct-1, 2. The prompt-tuned GPT-J-6B model generates the most diverse responses. "Added" and "none" mean whether the persona sentences are added to the input sentence or not.

4.2 Model Setup

To compare our prompt-tuning model with fine-tuning, we use the datasets in Section 4.1 and tune the pre-trained models of GPT series. We use two model sizes: GPT2-XL (1.5B parameters) and GPT-J-6B (Wang and Komatsuzaki, 2021). Fine-tuning of the GPT-J-6B model is not tested due to the lack of GPU memory.

The hyperparameters for prompt-tuning are based on the settings of (Lester et al., 2021). The length of the persona info tokens was set to 200 based on the results of preliminary experiments. The strategy for generating the response sentences is the greedy search. The number of epochs was set to a value such that the loss during learning converges. For fine-tuning, we experimented with two methods: one is to input only dialogue pairs, and the other is to add persona sentences before the dialogue pair's utterance and then input it into the model. Other hyperparameter values are given in Appendix B.

4.3 Results

We input the utterances of dialogue pairs from the evaluation datasets into the trained models. We automatically evaluate the diversity of the generated responses and manually assess whether the responses are natural and based on the persona.

4.3.1 Automatic Evaluation

We evaluate the diversity of the generated responses by distinct-N (Li et al., 2016a). The values of distinct-1 and distinct-2 are shown in Table 1. The evaluation values are the average of all the generation results of the persona, general eval datasets from each model corresponding to the three types of personas. The results show that the GPT-J-6B model trained by prompt-tuning generates the most

[1] https://github.com/facebookresearch/ParlAI/tree/main/parlai/tasks/personachat

[2] https://aclanthology.org/I17-1099/

[3] Each dialogue is assigned a topic. There are ten topics: Attitude & Emotion, Culture & Education, Finance, Health, Ordinary Life, Politics, Relationship, School Life, Tourism, and Work.

Eval Dataset	Training Method	Model	Fluency	Engagingness	Relevance
Persona Eval	Fine-Tuning (none)	GPT2-XL	3.52 (1.26)	3.70 (1.22)	3.30 (1.27)
	Prompt-Tuning		3.82 (1.06)	3.74 (1.17)	3.62 (1.02)
		GPT-J-6B	**3.90** (0.90)	**3.98** (0.95)	**3.82** (0.96)
General Eval	Fine-Tuning (none)	GPT2-XL	3.93 (1.19)	**3.82** (1.20)	3.77 (1.16)
	Prompt-Tuning		**4.04** (1.01)	3.81 (1.19)	**3.96** (1.13)
		GPT-J-6B	3.98 (1.03)	3.80 (1.01)	3.89 (1.05)
Human			4.31 (1.07)	4.25 (1.06)	4.36 (0.92)

Table 2: We evaluated the generated responses on a 5-point scale for fluency, engagingness, and relevance. We asked five workers to answer each question, and the averages of all answers and standard deviations (in parentheses) are shown. The prompt-tuned GPT-J-6B model scored highest in all aspects in the persona eval dataset. No significant differences were found in the general eval dataset.

Eval Dataset	Training Method	Model	[1,2)	[2,3)	[3,4)	[4,5]
Persona Eval	Fine-Tuning (none)	GPT2-XL	0	5	33	12
	Prompt-Tuning		0	7	41	2
		GPT-J-6B	0	2	29	19
General Eval	Fine-Tuning (none)	GPT2-XL	0	11	105	34
	Prompt-Tuning		0	8	75	67
		GPT-J-6B	0	1	91	58

Table 3: The generated responses were rated on a 5-point scale for persona consideration, and their distribution is shown. 1 is inconsistent with the persona, 3 is irrelevant to the persona, and 5 is in line with the persona. $[1, 2)$ means the number of sentences scored between 1 and 2, including 1. In each setting, the number of samples from the persona eval dataset is 50 and that from the general eval dataset is 150.

diverse responses. In fine-tuning, we also find that the results are better when persona sentences are not added to the input, similar to the experimental results using the seq2seq model in the experiments by Zhang et al. (2018).

4.3.2 Manual Evaluation

We use Amazon Mechanical Turk to manually evaluate whether the generated responses are natural and persona-based. Following the method of Zhang et al. (2018), the responses are rated on a 5-point scale on four aspects: fluency, engagingness, relevance, and persona consideration. We ask five workers to answer each question. In each setting, the number of samples from the persona eval dataset is 50 and that from the general eval dataset is 150. An example of tasks given to workers is shown in Appendix C.

The results of the first three aspects are shown in Table 2. The human scores are taken from the experiments by Zhang et al. (2018). In fine-tuning, when persona sentences are added to the input, the automatic evaluation results are worse than when they are not, and thus we only evaluate the models without persona sentences. From Table 2, the manual evaluation results using the persona eval dataset show that the prompt-tuned GPT-J-6B model achieves the best scores in all aspects. This can be attributed to the fact that the larger the model size was, the more knowledge was stored in the model through pre-training, and the more natural responses were generated by using this knowledge. Manual evaluation using the general eval dataset did not produce a significant difference. This is probably because most of the utterances in the general eval dataset are short and general, such as greetings, and the responses are also short and simple sentences.

For persona consideration, the distribution of the evaluation results is shown in Table 3, where 1 is inconsistent with the persona, 3 is irrelevant to the persona, and 5 is in line with the persona. The average score of persona consideration is calculated for each generated response. Table 3 shows that the majority of the generated responses are based on personas.

Comparing fine-tuning and prompt-tuning with the largest models that can be trained with a fixed GPU memory size, we can say that prompt-tuning can build a dialogue system with more natural re-

Utterance	Generated Response
are you tired ?	i am a little tired.
cool , i am eating donuts . what do you do ?	i am retired, i am a gym teacher. i used to teach gym.
how about the korean restaurant next to it ?	i heard it's very good.
i used to cook paella for a kitchen in spain .	i love paella. that sounds great. i bet it is a hard job to cook it right?
the tv news is reporting a bank robbery .	i heard it on the radio. what a shock.
what do you like to do in your spare time ?	i volunteer at a local soup kitchen, helping people in need.
hi , i like to keep fit and work out 5 times a week	i like to eat cheeseburgers and watch war documentaries
stop producing cigarettes .	i like the idea, but how can we do that?

Table 4: Response examples generated by the prompt-tuned GPT-J-6B model, which got the highest score in the manual evaluation. We can see that both of responses based on the persona (e.g., the sixth response) and responses unrelated to the persona (e.g., the first one) are generated.

Persona Sentence
i am a retired gym teacher. i volunteer at a soup kitchen. i was poor growing up. cheeseburgers are my favorite food. i like watching war documentaries.

Table 5: The persona used in the generated response examples in Table 4.

Training Method	Model	Dist-1	Dist-2
Fine-Tuning (none)	GPT2-XL	**0.134**	**0.379**
Prompt-Tuning		0.118	0.336
	HyperCLOVA	0.106	0.322

Table 6: Results of automatic evaluation by distinct-1, 2 in experiments in Japanese.

sponses based on the persona.

Table 4 shows response examples generated by the prompt-tuned GPT-J-6B model, which got the highest score in the manual evaluation. These responses are generated from the model trained with the dialogue pairs based on persona sentences shown in Table 5. We can see that training with small training datasets of only a few hundred pairs can produce a response with a natural and consistent personality, as shown in Table 4.

4.4 Experiments in Japanese

For our Japanese experiments, we use two datasets: JPersonaChat and JEmpatheticDialogues (Sugiyama et al., 2021).[4] As in the English experiments, three personas are used, and the number of dialogue pairs from JPersonaChat are 527, 525 and 525, respectively. To create training datasets, the same process as in the English experiments is used. Since most of the utterances in JEmpatheticDialogues are shorter and more general than those in JPersonaChat, we did not set any conditions for adding the utterances from JEmpatheticDialogues to the training datasets. The ratio of the training datasets is set to 1:10 based on the results of preliminary experiments. For the models, we use GPT2-XL[5] with 1.3B parameters and HyperCLOVA (Kim et al., 2021), a GPT3-like model with 6.9B parameters.

In the automatic evaluation results shown in Table 6, in contrast to the English experiments, HyperCLOVA, which has a higher number of parameters, tends to score lower. This can be attributed to the fact that there were many instances in which HyperCLOVA begins its response with back-channeling.

Table 7 shows the average scores for the three aspects within the manual evaluation results. For both the persona eval dataset and general eval dataset, the HyperCLOVA model with prompt-tuning scored the highest. The distribution of persona consideration is shown in Table 8. As in the English experiments, many responses are based on the persona and few are inconsistent with the persona. Generated response examples are shown in Appendix A.

5 Conclusion

We proposed a method for prompt-tuning a pretrained language model using dialogue data based on a single persona. Automatic and manual evaluations showed that we could construct a dialogue system that can respond more naturally and persona-based, with less computational resources than fine-tuning. Compared to the generated responses in English, those in Japanese look natural due to the

[4] https://github.com/nttcslab/japanese-dialog-transformers

[5] https://huggingface.co/rinna/japanese-gpt-1b

Eval Dataset	Training Method	Model	Fluency	Engagingness	Relevance
Persona Eval	Fine-Tuning (none)	GPT2-XL	3.81 (1.12)	3.63 (1.00)	3.81 (1.06)
	Prompt-Tuning		3.68 (1.23)	3.67 (1.13)	3.71 (1.17)
		HyperCLOVA	**3.87** (1.11)	**3.92** (0.98)	**3.90** (1.08)
General Eval	Fine-Tuning (none)	GPT2-XL	4.01 (0.96)	3.82 (0.89)	3.82 (1.00)
	Prompt-Tuning		3.99 (1.09)	3.68 (1.03)	3.92 (1.08)
		HyperCLOVA	**4.07** (1.01)	**3.86** (0.95)	**4.06** (0.99)
Human			4.31 (1.07)	4.25 (1.06)	4.36 (0.92)

Table 7: Results of manual evaluation of fluency, engagingness, and relevance for the generated responses in the Japanese experiments. We asked five workers to answer each question, and the averages of all answers and standard deviations (in parentheses) are shown. Prompt-tuned HyperCLOVA scored highest in all aspects on both datasets.

Eval Dataset	Training Method	Model	[1,2)	[2,3)	[3,4)	[4,5]
Persona Eval	Fine-Tuning (none)	GPT2-XL	0	5	105	40
	Prompt-Tuning		1	14	84	51
		HyperCLOVA	0	18	77	55
General Eval	Fine-Tuning (none)	GPT2-XL	0	8	122	20
	Prompt-Tuning		0	14	115	21
		HyperCLOVA	0	19	125	6

Table 8: Distribution of manually evaluated persona consideration in Japanese. In each setting, the number of samples is 150 for both persona eval and general eval datasets.

larger persona dataset. In the future, this method can be used not only to add personality to a dialogue system but also to build a dialogue system to generate responses with emotions by making a prompt for each emotion.

Acknowledgements

This work was supported by a joint research grant from LINE Corporation.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. 2021. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. 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, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *CoRR*, abs/2103.10385.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 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, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2021. Empirical analysis of training strategies of transformer-based japanese chit-chat systems.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2021. Spot: Better frozen model adaptation through soft prompt transfer. *CoRR*, abs/2110.07904.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

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.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

A Generated response examples in Japanese

Response examples that generated by prompt-tuned HyperCLOVA are shown in Table 9. These responses are generated from the model trained with the dialogue pairs based on persona sentences shown in Table 10.

Utterance	Generated Response
こんにちは。最近パーマをかけて イメチェンしたから今度みてほしいな。 (Hello. I recently got a perm and a makeover, so I hope you'll take a look at it sometime.)	いいよ。いつがいい?わたしは家政婦の仕事が あるから、夕方以降なら大丈夫だよ。 (Okay. When is a good time? I have to work as a housekeeper, so I can come after the evening.)
こんにちは!ウサギ飼ってるんですね、羨ましい。 (Hello! I see you have a rabbit, I'm jealous.)	はい、飼ってます。ペットショップで 一目惚れして、そのまま連れて帰りました。 (Yes, I have one. I fell in love with her at first sight at a pet store and took her home with me.)
絵の勉強をヨーロッパでやるとしたら、 どの国がいいんですか? (If you were to study painting in Europe, which country would you prefer?)	ゴッホが好きなら、 オランダがいいんじゃないですか? (If you like Van Gogh, why not Holland?)
大丈夫、ありがとう。短かったんだけど、 突然真っ暗になったから本当に怖い思いをした。 (No problem, thanks. It was short, but I was really scared because all of a sudden it went dark.)	それは怖いね。無事でよかった。 それにしても、停電って本当に困るよね。 (That's scary. I'm glad you're safe. Still, power outages are really annoying, aren't they?)
パイン飴ってそんなに昔からあるの? (Has candy pineapple been around that long?)	うん、子どもの頃から食べてたよ。 (Yeah, I've been eating it since I was a kid.)

Table 9: Response examples generated by prompt-tuned HyperCLOVA.

Persona Sentence
私は家政婦です。 (I am a housekeeper.)
私は姉と暮らしています。 (I live with my sister.)
私のペットはウサギです。 (My pet is a rabbit.)
私が尊敬する人は、画家のゴッホです。 (The person I admire is the painter Van Gogh.)
私は美術部に入っていました。 (I was in the art club.)

Table 10: The persona used in the generated response examples in Table 9.

B Hyperparameter

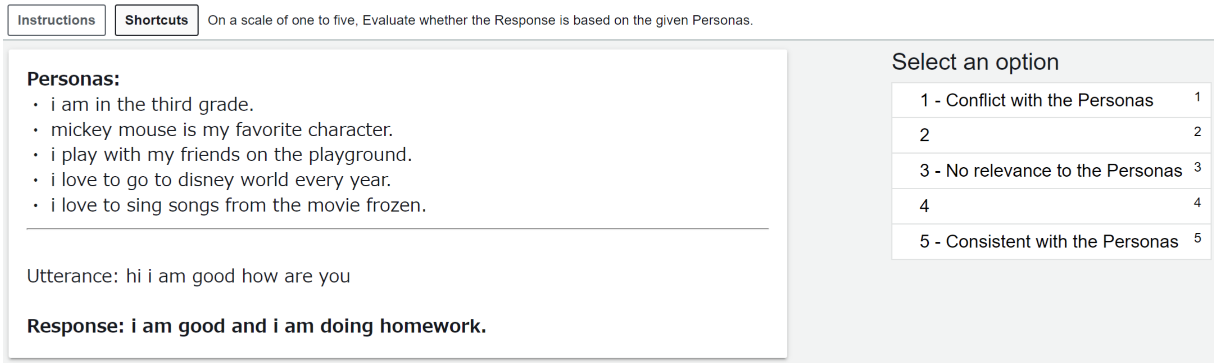
Table 11 shows hyperparameters during model training in our experiment.

Hyperparameter	Fine-Tuning (En)	Prompt-Tuning (En)	Fine-Tuning (Ja)	Prompt-Tuning (Ja)
Optimizer	Adam	Adam	Adam	Adam
Learning Rate	5e-5	1e-3	1e-5	1e-3

Table 11: Hyperparameters during model training in our experiment.

C An example of tasks used in crowdsourcing

Figure 2 shows an example of tasks used in crowdsourcing.



On a scale of one to five, Evaluate whether the Response is based on the given Personas.

Personas:
- i am in the third grade.
- mickey mouse is my favorite character.
- i play with my friends on the playground.
- i love to go to disney world every year.
- i love to sing songs from the movie frozen.

Utterance: hi i am good how are you

Response: i am good and i am doing homework.

Select an option

1 - Conflict with the Personas 1
2 2
3 - No relevance to the Personas 3
4 4
5 - Consistent with the Personas 5

Figure 2: An example of tasks given to workers on Amazon Mechanical Turk for the manual evaluation.