Persona-based Dialogue Response Generation Using Personal Facts and Personality Traits

Weiwen Su, Naoki Yoshinaga, Yuma Tsuta and Masashi Toyoda

Abstract Persona-based chatbots, which assume a specific human-like persona for chatbots, have been studied to generate consistent and engaging responses. The common approach to this problem is to provide concrete profiles in text. Although text profiles can describe not only personal facts (*e.g.*, "I have a dog") but also personality traits ("I'm likely to follow other's opinions."), the existing persona-based dialogue datasets such as Muti-Session Chat (MSC) contain mainly personal facts. In this study, we augment the MSC datasets with profiles on predicted personality traits to train and evaluate a persona-based chatbot based on both personal facts and personality traits. We explore methods to verbalize and incorporate the personality traits in persona-based chatbots, and then propose a reranking method for response candidates to increase personality consistency. Experimental results on the augmented MSC dataset confirm that the personality traits help the chatbot generate more consistent responses.

1 Introduction

Open-domain dialogue systems or chatbots, such as Siri, Microsoft XiaoIce [1], and ChatGPT, have become more common in our daily lives. As a partner of daily conversation, we expect chatbots to converse like humans, namely, to generate consistent responses based on a consistent persona. However, since the conversation data used to train open-domain chatbots usually compile conversations done by various persons, the resulting chatbots are likely to generate inconsistent responses.

Weiwen Su and Yuma Tsuta
The University of Tokyo
e-mail: {su-w,tsuta}@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga and Masashi Toyoda
Institute of Industrial Science, The University of Tokyo
e-mail: ynaga@iis.u-tokyo.ac.jp, toyoda@tkl.iis.u-tokyo.ac.jp

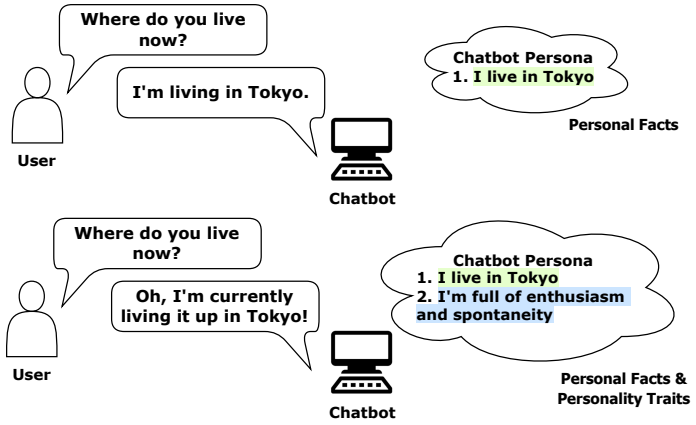

Fig. 1: Persona-based chatbots with only personal facts (above) and with both personal facts and personality traits (below).

To generate consistent responses by chatbots, researchers take the identity of speakers into consideration [2, 3, 4, 5, 6, 7, 8]. Li et al. [2] have initiated to model the speaker identity implicitly by using a speaker embedding to distinguish individual speakers in the training data. To explicitly specify the chatbot's persona, Zhang et al. [4] built PersonaChat dataset, which provides speaker profiles as text descriptions (*e.g.*, "I have a dog."); most of the following studies on persona-based chatbots leverage this dataset or its extension, Multi-Session Chat (MSC) [9]. Although the existing datasets for persona-based chatbots contain various profiles to characterize a chatbot, those profiles are mainly personal facts such as personal tastes, relatives, social status, and experiences, and are barely personality traits such as agreeableness and extraversion. Meanwhile, although Saha et al. [10] predicted Big-Five personality traits for speakers in several dialogue datasets to control the style of generated utterances, these datasets do not contain personal facts.

In this study, aiming to investigate the impact of using both personal facts and personality traits in training persona-based chatbots, we automatically annotate an existing persona-based dataset, MSC, with Big-Five personality traits using a detector trained on the Pandora dataset [11]; we then train and evaluate a persona-based chatbot using this augmented MSC dataset. An issue here is how to represent predicted personality traits (category with intensity) as profiles. We adopt the same short text descriptions as the original profiles in the MSC dataset on personal facts, to maintain the interpretability and flexibility of profiles, combine those profiles, and prepend them to the input utterance to feed a persona-based chatbot. To enhance the personality consistency, we incorporate a response ranking model inspired by [8] that computes the consistency between the personality profiles and the generated utterance by choosing the response candidate with the highest consistency among multiple response candidates generated by using top-k sampling.

We use our personality-augmented MSC dataset to evaluate the impact of using both profiles on the original personal facts and predicted personality traits in a persona-based chatbot. The automatic and human evaluation confirmed the effectiveness of personality profiles when we use the proposed reranking model. We will release our augmented MSC dataset to promote the reproducibility of our results.[1]

The contributions of this paper are as follows:

- We augmented the MSC dataset with Big-Five personality traits using a detector, to train and evaluate a persona-based chatbot that takes both personal facts and personality traits into account.
- We explored effective methods for leveraging personality traits in a persona-based chatbot. Specifically, we design a method of verbalizing personality traits and reranking response candidates to improve personality consistency
- We confirmed the merits of using both personal facts and personality traits in persona-based chatbots through automatic and human evaluation of generated responses on the augmented MSC dataset.

2 Related Work

In this section, we first review existing studies on persona-based response generation. We next introduce personality-controlled dialogue generation and then mention approaches to detect speakers' personality traits from the text.

2.1 Persona-aware Response Generation

Li et al. [2] have first pointed out the problem of inconsistent responses generated by data-driven chatbots. To address this problem, they induced speaker embeddings from the speakers' dialogue histories to model individual speakers in a conversation dataset and generate more consistent responses. Ma et al. [6] induced speakers' implicit profiles from dialogue histories, and Zhong et al. [12] further refined the speaker's dialogue histories to obtain more precise and abundant persona information. Although these approaches allow us to model speakers in dialogue datasets, the data-driven persona representations will capture only the explicitly mentioned persona of the speakers, since open-domain dialogue datasets are often sourced from microblogs such as X (formerly, Twitter); the speakers may hesitate to expose their detailed persona information due to privacy concerns.

Zhang et al. [4] thereby created PersonaChat, the most commonly used dataset for persona-based chatbots; it compiles conversations between a pair of speakers that role-play given profiles, a series of text descriptions. Majumder et al. [13] expanded the original profiles in this dataset using commonsense knowledge. Liu et

[1] https://github.com/NioHww/Extended-MSC

al. [14] incorporated the profiles of system users to encourage mutual persona perception via reinforcement learning. In the context of long-term conversation, the profiles of system users may change over time. Xu et al. [9] thus released Multi-Session Chat, which extended PersonaChat with future dialogue sessions.

The profiles in the PersonaChat and MSC contain mainly personal facts such as personal tastes, relatives, and social status, and do not include personality traits that will contribute to responses. We thus propose to add personality traits as text descriptions to the MSC dataset to train and evaluate persona-based chatbots that take both personal facts and personality traits into consideration.

2.2 Personality-controlled Text Generation

In dialogue modeling, personality traits such as agreeableness and extraversion are considered to affect speaking styles. Mairesse and Walker [15] first proposed a personalty-aware text generation model, focusing on the extraversion personality. Recently, Wang et al. [16] proposed a seq2seq model for Big-Five personality-conditioned response generation. Saha et al. [10] adopted Big-Five personality traits and discourse intents as stylistic control codes to generate stylistic responses. Xu et al. [17] targeted three of the Big-Five personality dimensions according to the results of principal component analysis for narrative dialogue generation. The current studies on personality-controlled response generation mainly depend on predicted personality traits using a personality detector, due to the absence of personality-labeled dialogue datasets.[2] Meanwhile, although personal facts and personality traits are two important factors that characterize a speaker, no study attempts to model both persona information into account in response generation, due to the lack of datasets.

2.3 Personality Detection From Text

Self-reported personalities suffer from biases (*e.g.*, answers may be influenced by social trends) [19] and assessment from psychologists is hard to acquire. Therefore, automatic personality detection from text is an important task in performing personality-aware language modeling. Early studies [20] mainly utilized linguistic features for personality detection. Recently, neural network-based methods have been widely applied to this task. Ren et al. [21] leveraged BERT [22] and emotional information for the detection. For more specialized models, Tao et al. [23] and Zhu et al. [24] elaborated graph neural networks to better detect personality from the text. In this study, we follow these studies to obtain personality traits for the annotation.

[2] Very recently, Yamashita et al. [18] have built the RealPersonaChat dataset in Japanese, including massive conversations with both personal facts and personality traits provided by the speakers themselves. However, the paper and the dataset have not been made to the public at the moment.

3 Approach

In this section, we describe our method to generate a response using not only personal facts but also personality traits in open-domain conversation. The task is defined as follows. Given the dialogue context C and chatbot's personal facts P_{fact} and personality traits P_{trait}, the model optimizes the generation of response R to maximize a conditional probability, $P(R \mid C, P_{\text{fact}}, P_{\text{trait}})$, on the training data.

Because existing datasets with personal facts such as PersonaChat [4] and MSC [9] do not contain the personality traits of the speakers, we augment the existing dataset (MSC, in this study) with personality traits of the speakers to construct the desired dialogue dataset. For this augmentation, we train a personality detector using Pandora dataset [11] (§ 3.1) and attach the estimated personality traits to the MSC dialogues (§ 3.2). In the development of persona-based chatbots, we explore the effective method to incorporate two types of persona information and introduce a response reranking method to improve personality consistency (§ 3.3).

3.1 Building Personality Detector

To augment existing profiles attached to the dialogue dataset with personality traits, we train a personality detector using Pandora [11] dataset, which consists of Big-Five personality intensities for more than 1500 Reddit users. We develop a RoBERTa [25]-based regression model to predict the target user's intensities of all Big-Five personality dimensions at once from each of the target user's utterance. We explore the following two variations of the personality detector:

Personality detection using accumulated utterances (AU) Following previous work [10], this detector predicts a target user's Big-Five personality intensities from accumulated target user's utterances. Since available responses vary across users, we randomly pick 2 to 8 non-overlap utterances from one user's utterances and concatenate them as input to the detector.

Personality detector using contextualized utterances (CU) Since the context of the target utterance will help detect the personality intensities, we feed an utterance to which the target utterance replies, if any, with the target utterance as an input of the detector. Since the Pandora dataset provides only the utterances of each user, we retrieve Reddit posts that have parent post IDs for the post IDs of the target utterances.

For given utterances of the target user, we prepare multiple inputs to the above personality detectors and accumulate (average) the resulting intensities to obtain the reliable personality intensities of the target user. To see how this accumulation contributes to the correlation with the gold personality intensities, we evaluate the detector outputs for single inputs and averages of multi inputs (here, three inputs). Following the previous study [11], we use Pearson correlation between the detected

Detector	1-sample	3-sample (average)
AU Personality Detector	0.272	0.391
CU Personality Detector	0.357	0.465

Table 1: Results of Person correlation for the personality intensities detected by AU and CU personality detectors. Each sample for the AU personality detector contains accumulated 2-8 utterances. Pearson correlation are $p < 0.05$.

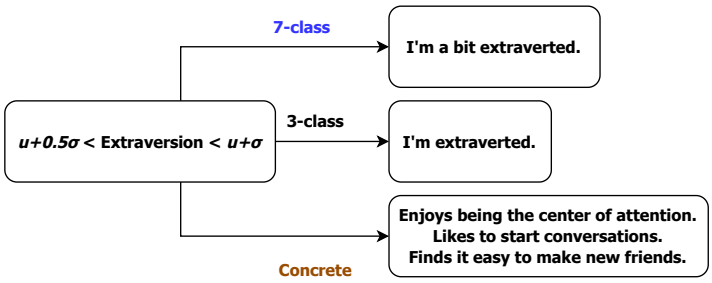

Fig. 2: A schema of the three verbalization methods to get personality-related profiles. A case of extraversion intensity between $u + 0.5\sigma$ and $u + \sigma$ is shown.

Big-Five personality intensity and the provided gold intensity in each personality dimension of the target user, and average over personality dimensions.

Table 1 shows the evaluation results of the AU and CU personality detectors. This result confirms that we can obtain more reliable personality intensities by averaging the results of multiple samples. Since we employ these detectors on dialogues in the MSC dataset, which are different from the Pandora dataset, we compare both detectors in the following experiments.

3.2 Annotating Multi-Session Chat with Personality Traits

We have chosen the MSC dataset [9] as the target persona-based dialogue dataset to augment personality traits since it is larger than PersonaChat [4]. We estimate the personality traits (Big-Five personality intensities) of the speakers for individual sessions rather than all sessions. This is because different speakers may role-play the same profiles across sessions and the same speakers may role-play different profiles across sessions, which will affect the discharge of personality.

Following the results in § 3.1, we average personality intensities estimated for the utterances of the target speaker in each session to acquire a more reliable personality. Because the MSC dataset is created by role-playing given profiles, this role-playing may affect the speakers' personality traits. We thus calculated the standard deviation for each dimension of personality traits estimated for each turn in a

session and ignored the speakers' personality dimensions with extremely high standard deviations, since it means an unstable personality trait.

There is an issue in using the estimated intensities of personality traits for our purpose; since personal facts are given as text descriptions, we want to represent the estimated intensities of personal traits as text descriptions. Therefore, we explore the following three verbalization approaches to obtain text descriptions of personality traits as shown in Figure 2:

3-class verbalization Split each personality intensity into three classes with 0.5 standard deviation σ from the mean u of the estimated intensity for each personality trait. We then verbalize the personality traits using the corresponding adjectives and their antonyms of each Big-Five personality trait (*e.g.* "extraverted" and "introverted"). We add nothing for the neutral class.

7-class verbalization Rather than the three classes, using 0.5, 1, and 2 σ from u as the threshold, we further divide each personality trait into seven classes by adding adverbs ("a bit", no adverb, and "quite") of the degree to create more refined descriptions. Specifically, this results in six descriptions in addition to the empty description for the neutral class.

Concrete verbalization We take more concrete descriptions from the psychology websites[3] for those who own such personality traits as the concrete descriptions of personality traits. Specifically, we randomly sampled the concrete descriptions according to the predicted intensity; 33%, 66%, and 100% descriptions among all the descriptions are used and then combined.

3.3 Response generation using personal facts and personality traits

In this section, we first discuss how to incorporate verbalized personal traits into a persona-based chatbot and then propose a reranking method for response candidates to improve personality consistency.

3.3.1 Incorporating Verbalized Personality Traits

An issue in training a persona-based chatbot with text profiles of personal facts and personality traits is how to feed them to the model. The influence of the profiles on utterances depends on individual profiles. Personal facts affect topics in a conversation, whereas personality traits affect (re)actions and speaking styles. We compare the following two ways to feed the profiles. The model here used is GPT-2 [26] based model.

Mixed The profiles of personal facts and personality traits are mixed, randomly ordered, concatenated without any special tokens, and then prepended to the dialogue contexts for response generation.

[3] https://www.verywellmind.com/the-big-five-personality-dimensions-2795422

Separated The profiles of personality traits and personality facts are interdependently concatenated without any special tokens; these two sequences of profiles are then concatenated and prepended to the dialogue contexts. Distinct special tokens are inserted before each sequence of profiles to indicate their types.

In both methods, distinct special tokens are inserted between the profiles and dialogue contexts and between utterances in the contexts.

3.3.2 Personality-aware Reranking

Naively incorporating personality traits may not facilitate the model to fully utilize the given personality traits. Thus, inspired by [8] that performs reranking of verbosely-generated response candidates to improve consistency between given personal facts and system responses, we propose to rerank verbosely-generated response candidates to improve consistency between given personality traits and system responses.

To train a scorer that measures the personality consistency for response reranking, we reuse the training split of the augmented MSC dataset. Specifically, we train a RoBERTa [25]-based response scoring model[4] to compute the consistency between the given utterance and profiles of personality traits. To train this model, we collect triplets of the target utterance, the target personality traits, and their consistency in terms of personality as the training data. To form these triplets, we combine each utterance and its speaker's personality traits in the augmented MSC dataset; their consistency score is 1. For each of these triplets, we then pick an utterance from different speakers, while keeping the previous personality traits, and compute the cosine similarity of the personality intensities of the two speakers as the consistency. We create two triplets for each triplet, and and combine all for training.

Moreover, to ensure the generalization ability of the reranking model to deal with profiles of personality traits created using three verbalization methods, we pick 33% non-overlap examples from the three versions of the augmented MSC dataset. In the inference stage, given profiles of personality traits and one candidate response, the model could predict the personality consistency between them.

In evaluation, we generate five response candidates for given dialogue contexts and profiles to rerank the candidates in terms of the estimated personality consistency. We then pick the one with the highest personality consistency as the response.

4 Experiments

In this section, we train persona-based chatbots on our augmented MSC dataset to confirm the effectiveness of using both personal facts and personality traits.

[4] We did not reuse personality intensity estimator in § 3.1 for this reranking, since the chatbot's personality traits will be provided by a system user in the text format.

4.1 Datasets

We built the six personality-augmented MSC dataset, by combining the two personality detectors and the three verbalization methods for personality traits. Since individual dialogue sessions, a series of consecutive utterances, in the MSC datasets can be generated by different pairs of speakers even if they maintain the same personal facts, we ignore multi-session settings in the dataset and handle individual sessions as independent dialogues to guarantee the consistency of speakers' personality traits; namely, we estimate speakers' personality traits for individual sessions. We used the original split of training, validation, and test data for evaluation.

4.2 Models

We adopt the DialoGPT model [27] as an base model of persona-based chatbots, and fine-tune this model on the six personality-augmented MSC datasets. The DialoGPT is based on GPT-2 [26] and has been trained on Reddit comment chains data. The models to be compared in the experiment are as follows:

Baseline We fine-tuned the DialoGPT model on one augmented MSC dataset (ignoring personality traits) with and without the personal facts as baselines. We hereafter referred to them as **Baseline** and **+ personal facts**, respectively.

Proposed We fine-tuned three DialoGPT models on the personality-augmented MSC datasets by combining three personality verbalization (3-class, 7-class, and Concrete) and the mixed or separated incorporation method for personality traits. We referred them to as **3-class Mixed/Separated**, **7-class Mixed/Separated**, **Concrete Mixed/Separated** respectively.

We perform the reranking of response candidates for the proposed models. We evaluate these models trained with mixed personality traits incorporation method as the main results and later compare two personality incorporation methods.

We adopt the Huggingface's implementations, `roberta-base`[5] for personality detectors and personality-aware reranking model, and use `DialoGPT-small`[6] for response generation. We use AdamW optimizer [28] to fine-tune these models. The hyperparameters for fine-tuning RoBERTa and DialoGPT are as follows: batch size of 32 and 4, learning rate of 2e-5 and 5e-5, respectively. At the decoding stage for the response generation models, we first sample five response candidates using top-k sampling and then rerank them using our proposed method; here, k is set as 50. We compute evaluation metric C_{fact}, which we will explain later, using `roberta-large` with a learning rate of 1e-6 and batch size of 8.[7] All the experiments were conducted via PyTorch on one NVIDIA Quadro P6000 GPU.

[5] https://huggingface.co/roberta-base

[6] https://huggingface.co/microsoft/DialoGPT-small

[7] https://huggingface.co/roberta-large

4.3 Metrics

To evaluate the persona-based chatbots, we consider two aspects in automatic and human evaluation; one is the quality of system response for the given dialogue contexts and the other is the consistency between the system response and given profiles, namely, personal facts and personality traits. In what follows, we explain automatic and human evaluation, respectively.

To measure the quality of responses against the given dialogue contexts, we use perplexity and BLEU-1/2 [29] as basic metrics. We also evaluate the Distinct-1/2 (DIST-1/2) [30] to show whether the system responses exhibit a certain degree of diversity.

To measure the consistency between responses and given profiles, we used two metrics, each of which measures the consistency against personal facts and personality traits, respectively. Following the previous studies, we evaluate the consistency on personal facts by using consistency score (C_{fact}) [31] which is a textual entailment score computed using a RoBERTa [25] model trained on the DNLI dataset [32]. For each response, we compute C_{fact} as whether the response entails each profile on personal facts and sum the result (-1 for contradiction, 0 for neural, and 1 for entailment) up. Apart from C_{fact}, it is also necessary to evaluate the personality consistency, like the utterance level Pearson correlation used by [10]. Since the detection based on a single response is unstable, we compute the Pearson correlation between the average of personality intensities computed for system responses in each session and the given personality as personality correlation (C_{trait}).

In human evaluation, as for the response quality, following the setting of [8], we adopt Fluency, Coherence, and Informativeness, and ask three human subjects (graduate students) to annotate five-point Likert scales for the three dimensions, where 1-point means bad quality, 3-point means moderate and 5-point means perfect. As for the consistency of personal facts, a point of -1, 0, and 1 is assigned to each response, which means contradicted, neutral, or entailed to all given personal facts. As for personality consistency, we asked the subjects to assign a point of 1 meaning reflecting the given personality, and 0 meaning not reflecting that.

4.4 Main Results

Table 2 and Table 3 list the evaluation results of the generated responses for the MSC datasets augmented by AU and CU personality detectors, respectively. As for the results of AU personality traits, we can observe that the proposed method improves BLEU-1/2 and C_{trait} and is comparable to the baselines in terms of the other metrics. As for the results of CU personality traits, we can observe that the proposed method improves C_{trait}, and is comparable to the baselines in terms of the other metrics than C_{fact}. The 7-class verbalization of personal traits shows better scores than the other two verbalization methods.

Models	perplexity	BLEU-1	BLEU-2	DIST-1	DIST-2	C_{fact}	C_{trait}
Baseline	18.81	10.86	1.99	2.87	28.70	0.332	0.386
+ personal facts	18.52	10.86	2.06	2.83	28.73	0.443	0.390
Proposed (Baseline + personal facts and personality traits)							
3-class traits, Mixed	18.61	11.20	2.02	2.77	28.54	0.434	0.646
7-class traits, Mixed	18.56	11.45	2.13	2.80	28.58	0.460	0.651
Concrete traits, Mixed	18.56	11.21	2.10	2.77	28.42	0.442	0.640

Table 2: Evaluation results of persona-based chatbots using AU personality traits. Pearson correlations are $p < 0.05$.

Models	perplexity	BLEU-1	BLEU-2	DIST-1	DIST-2	C_{fact}	C_{trait}
Baseline	18.81	10.86	1.99	2.87	28.70	0.332	0.639
+ personal facts	18.52	10.86	2.06	2.83	28.73	0.443	0.649
Proposed (Baseline + personal facts and personality traits)							
3-class traits, Mixed	18.66	10.79	2.03	2.81	28.44	0.430	0.697
7-class traits, Mixed	18.50	10.93	2.06	2.88	28.89	0.421	0.706
Concrete traits, Mixed	18.58	10.83	1.91	2.75	28.33	0.437	0.688

Table 3: Evaluation results of persona-based chatbots using CU personality traits. Pearson correlations are $p < 0.05$.

Models	Fluency	Coherence	Informativeness	$Consistency_{fact}$
Baseline + personal facts	3.85	2.77	2.65	0.17
Proposed (7-class Mixed)	3.99	2.69	2.54	0.19
Human	4.88	4.78	4.53	0.66

Table 4: Human evaluation for response quality using AU personality traits.

Intensity	Agreeableness	Openness	Conscientiousness	Extraversion	Neuroticism
High	0.71	0.35	0.65	0.61	0.30
Low	0.31	0.21	0.61	0.35	0.75

Table 5: Human evaluation for pair-wise personality consistency using AU personality traits; High and Low rows show results on target personality traits with high and low intensities (*e.g.*, extraverted and introverted for extraversion), respectively.

Table 4 shows the human evaluation of for 100 system responses generated by the best-performing model with the 7-class verbalization of personality traits using the AU personality detector, for metrics other than personality consistency. The results show that the 7-class setting slightly outperformed **+ personal facts** in terms of Fluency and Consistency, and the model trained with only personal facts (**+ per-**

Models	perplexity	BLEU-1	BLEU-2	DIST-1	DIST-2	C_{fact}	C_{trait}
Proposed (using AU personality traits)							
3-class traits, Separated	18.51	11.39	2.17	2.83	28.72	0.450	0.646
7-class traits, Separated	18.46	11.11	2.04	2.82	28.75	0.429	0.664
Concrete traits, Separated	18.56	11.14	2.11	2.73	28.34	0.443	0.638
Proposed (using CU personality traits)							
3-class traits, Separated	18.50	10.95	2.03	2.86	28.77	0.435	0.698
7-class traits, Separated	18.52	10.90	2.01	2.84	28.71	0.424	0.707
Concrete traits, Separated	18.56	10.89	2.01	2.73	28.10	0.448	0.694

Table 6: Automatic results with proposed setting using the separated input method. Pearson correlations are $p < 0.05$.

sonal facts) achieved better results in Coherence and Informativeness. The use of personality traits did not show a negative impact on the response quality.

We also conduct a human evaluation to examine personality consistency. Because we found that evaluating personality consistency only with a given personality trait and a single system response is difficult for the human subjects, we obtained 100 pairs of generated responses using the 7-class Mixed model with original profiles of personality traits and with the same profiles with the intensity of one (target) dimension inverted. The human subjects are then asked for each pair of responses to judge if both responses were consistent in terms of the target personality traits, by a point of 1 (consistent) or 0 (inconsistent). The results are shown in Table 5. We can observe that responses based on "conscientiousness" profiles attain high consistency regardless of the intensity, while responses on "openness" profiles attain low consistency regardless of the intensity. These results highlight the remaining challenges in generating personality-consistent responses.

4.5 Ablation Studies

In this section, we evaluate alternative methods to those evaluated in the main results. In what follows, we first evaluate the separated incorporation method of personality traits and compare the results with the main results using the mixed incorporation method.

Table 6 shows the evaluation results with the AU personality detector. From the results, we can observe that the separated incorporation method leads to a similar performance to the mixed input method. We conclude that the way of incorporating personality traits did not affect much on the system response.

Table 7 shows an ablation study to examine the influence of the reranking of response candidates in terms of personality consistency. The results without reranking (w/o Personality-aware Reranking) confirm the huge impact of the reranking on personality consistency. The model trained with personality traits could not im-

Models	perplexity	BLEU-1	BLEU-2	DIST-1	DIST-2	C_{fact}	C_{trait}
Baseline + personal facts	18.52	10.86	2.06	2.83	28.73	0.443	0.390
Proposed w/o Personality-aware Reranking							
3-class traits, Mixed	18.61	10.92	2.03	2.75	28.27	0.427	0.394
7-class traits, Mixed	18.56	11.07	20.4	2.83	28.55	0.442	0.386
Concrete traits, Mixed	18.56	10.96	2.00	2.75	28.14	0.415	0.378
Proposed w/o Personal Facts & Personality-aware Reranking							
3-class traits, mixed	18.79	11.13	2.07	2.82	28.78	0.352	0.415
7-class traits, mixed	18.74	11.04	2.05	2.82	28.70	0.339	0.410
Concrete traits, mixed	18.81	11.00	2.05	2.90	28.83	0.335	0.416

Table 7: Ablation studies using AU personality traits. Pearson correlations are $p < 0.05$.

Models	C_{trait}
7-class traits, Mixed	0.651
+ 1 Dimension Reverse	0.455
+ All Dimensions Reverse	−0.085

Table 8: Result of personality control using AU personality traits. Pearson correlations are $p < 0.05$.

prove C_{trait}. This is probably because the model may have a strong emphasis on personal facts due to the process of creating the MSC dataset, and naively incorporating personality traits did not contribute to the personality consistency. Further ablation test on the proposed model based only on personality traits (w/o Personal Facts & Personality-aware Reranking) confirmed that the personal facts caused a slightly negative impact on C_{trait}. In conclusion, it would be better to design a more effective method to respectively utilize the two kinds of profiles. At the moment, personality-aware reranking is vital to improve personality consistency.

4.6 Personality Control

To check the ability of the profiles of personality traits to control the personality, we respectively reverse the intensity (and the 7-class verbalization) of one or all dimensions of personality traits and evaluate the Pearson correlation between the detected personality traits of generated response and gold personality traits. The results are shown in Table 8. From the results, we can observe that the descriptions of personality traits successfully control the personality traits of the generated responses, which also serves as an indirect reflection of personality consistency. Also, we can notice that the C_{trait} does not attain an exact negative value of the original one when revers-

Models	Attention to personal facts	Attention to personality traits
w/o personality-aware reranking		
AU personality traits	0.897	0.856
CU personality traits	0.893	0.790
with personality-aware reranking		
AU personality traits	0.897	0.851
CU personality traits	0.900	0.794

Table 9: Attention analysis for response generation using 7-class separated setting. The score is scaled by multiplying 1000.

ing all of the descriptions of personality traits due to a large portion of the dialogues not being endowed with descriptions of personality traits in all dimensions.

4.7 Attention Analysis of Response Generation

We conduct an attention analysis of response generation to see how the models distribute their attention on the personality traits and personal facts, which will provide a view of the difference in their performance. We analyze the attention of models using the 7-class verbalization of AU and CU personality traits with and without the personality-aware reranking method, which result in four models in total. The results with the reranking method target at the finally-chosen responses.

We calculate the attention score by averaging the attention from tokens in responses to tokens in personality traits and personal facts respectively. More concretely, for each token in personality traits, we calculate the average (12 layers, 12 heads, length of the response) attention from all tokens in the response to it. Then the attention score to personality traits is the average of such averaged attention of every token in it. We show the attention scores of the models in Table 9. From the results, we can observe that the models using AU personality traits paid more attention to personality traits than models using CU personality traits while their attention to personal facts almost the same. The improvement of attention to personality traits led to better performance. Meanwhile, the reranking method did not lead to additional attention to the personality traits.

4.8 Example of Reranking Response Candidates

Table 10 shows an example of reranking response candidates by the chatbot with the 7-class traits with mixed incorporation method using AU personality traits. The response candidates are shown with the corresponding score from the personality-aware scoring model. We can observe that the second candidate with a higher con-

Profiles of Chatbot	
Personal Facts	I am starting a new job. I'm going to work in computer programming. I have no wife or kids. I like to go on cruises during my vacations. I live in San Francisco, I just moved here and am trying to make more friends. I lived in Kansas City.
Personality Traits	I am disagreeable. I am a bit closed-minded.
Dialogue Context	
User	I might go to San Francisco for my next trip.
Chatbot's Response Candidates (Score for Reranking Method)	
- You will be so excited! I really feel like it would be a fun trip! (-0.068)	
- How does San Francisco compare to New York? I know California gets a little hot around here. (0.697)	
- How much are tickets? I hope you will have a lot of good time to go out and visit! (0.175)	
- That does sound pretty cool! Are you planning on going anywhere at all? (0.152)	
- Yeah, I saw a couple of pictures that look amazing! What was your favorite part? (0.035)	

Table 10: An example of personality-aware reranking.

sistency score shows more disagreeableness (trying to give other choices), which is consistent with the given *"I am disagreeable.,"* while candidates with lower consistency scores show more agreeableness (advocate user's decision).

5 Conclusions

In this paper, we study a persona-based chatbot based on not only personal facts but also personality traits in open-domain conversation. We first constructed a dataset to train and test a persona-based chabot by extending the MSC dataset [9], which contains personal facts, with personality traits. We train two types of personality detectors using the Pandora dataset [11] and annotate the personality traits of the speakers in the MSC dataset. To train a persona-based chatbot using the augmented MSC dataset, we explore methods for verbalizing personality traits and combining them with the profiles on personal facts. Then, we propose a reranking method to pick response candidates with better personality consistency. Experimental results on the personality-augmented MSC dataset show an improvement in consistency in terms of personality traits. As for future work, we plan to design a more effective model structure to fully utilize personal facts and personality traits at the same time.

Acknowledgements This work was partially supported by the special fund of Institute of Industrial Science, The University of Tokyo, by JSPS KAKENHI Grant Number JP21H03494, JP21H03445, and by JST, CREST Grant Number JPMJCR19A, Japan.

References

1. Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.
2. Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.
3. Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Modeling situations in neural chat bots. In Allyson Ettinger, Spandana Gella, Matthieu Labeau, Cecilia Ovesdotter Alm, Marine Carpuat, and Mark Dredze, editors, *Proceedings of ACL 2017, Student Research Workshop*, pages 120–127, Vancouver, Canada, July 2017. Association for Computational Linguistics.
4. Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.
5. Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization, 7 2018.
6. Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 555–564, New York, NY, USA, 2021. Association for Computing Machinery.
7. Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics.
8. Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. SimOAP: Improve coherence and consistency in persona-based dialogue generation via over-sampling and post-evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9945–9959, Toronto, Canada, July 2023. Association for Computational Linguistics.
9. Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics.
10. Sougata Saha, Souvik Das, and Rohini Srihari. Stylistic response generation by controlling personality traits and intent. In Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi, editors, *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 197–211, Dublin, Ireland, May 2022. Association for Computational Linguistics.
11. Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. PANDORA talks: Personality and demographics on Reddit. In Lun-Wei Ku and Cheng-Te Li, editors, *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online, June 2021. Association for Computational Linguistics.

12. Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. Less is more: Learning to refine dialogue history for personalized dialogue generation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States, July 2022. Association for Computational Linguistics.
13. Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online, November 2020. Association for Computational Linguistics.
14. Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. You impress me: Dialogue generation via mutual persona perception. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online, July 2020. Association for Computational Linguistics.
15. François Mairesse and Marilyn Walker. PERSONAGE: Personality generation for dialogue. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
16. Wanqi Wu and Tetsuya Sakai. Response generation based on the big five personality traits. 2020.
17. Weilai Xu, Fred Charles, and Charlie Hargood. Generating stylistic and personalized dialogues for virtual agents in narratives. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 737–746, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.
18. Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. Realpersonachat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, 2023.
19. Sanja Stajner and Seren Yenikent. A survey of automatic personality detection from texts. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6284–6295, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.
20. François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Int. Res.*, 30(1):457–500, nov 2007.
21. Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xu. A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, 58(3):102532, 2021.
22. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics.
23. Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. Orders are unwanted: Dynamic deep graph convolutional network for personality detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13896–13904, Jun. 2023.
24. Yangfu Zhu, Linmei Hu, Xinkai Ge, Wanrong Peng, and Bin Wu. Contrastive graph transformer network for personality detection. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4559–4565. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

25. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
26. Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
27. Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*, 2020.
28. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.
29. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
30. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *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, June 2016. Association for Computational Linguistics.
31. Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy, July 2019. Association for Computational Linguistics.
32. Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics.