KEEP CHATTING! An Attractive Dataset for Continuous Conversation Agents

Yihe Wang[1], **Jin Liu**[1*], **Yao Wan**[2], **Yitong Li**[3], **Zifeng Liu**[1], **Weipeng Chen**[1]
[1]School of Computer Science, Wuhan University
[2]Huazhong University of Science and Technology [3]Huawei Technologies Ltd.
{yihewang, jinliu, liuzf22, weipengchen}@whu.edu.cn wanyao@hust.edu.cn liyitong3@huawei.com

Abstract

Ongoing conversation is crucial for conversational agents to build long-term connections with users. However, users tend to quickly lose interest if the conversational agent's responses are not engaging enough. In this paper, we introduce a novel task aimed at increasing users' willingness to continue interacting with the agent. We create a dataset named CONTINUOUSCHAT by: (i) collecting and revising personas, then expanding them into detailed personas through experiences, daily life, future plans, or interesting stories; (ii) transforming detailed personas into dialogues infused with emotions and feelings; (iii) rewriting the dialogues in specific styles using few-shot prompts, conditioned on handwritten style-specific examples. We benchmark Large Language Models (LLMs) on the CONTINUOUSCHAT dataset using both fine-tuning and in-context learning settings. Experiments with publicly available models show that while there is substantial room for improvement in generating style-specific dialogues, our CONTINUOUSCHAT dataset is valuable for guiding conversational agents to produce more engaging dialogues and increase users' willingness to continue conversations.[1]

1 Introduction

Open-domain dialogue is a longstanding challenge in Natural Language Processing (NLP) that has garnered widespread interest from researchers. Numerous approaches have been explored, and recently, generation models trained on large-scale datasets have gained significant attention (Adiwardana et al., 2020; Roller et al., 2021; Xu et al., 2022a; Bao et al., 2020; Zhang et al., 2020; Mi et al., 2022). Maintaining ongoing conversations is crucial for conversational agents to establish long-term connections with users. However, people tend to quickly lose interest if the conversational agent's responses are not sufficiently engaging.

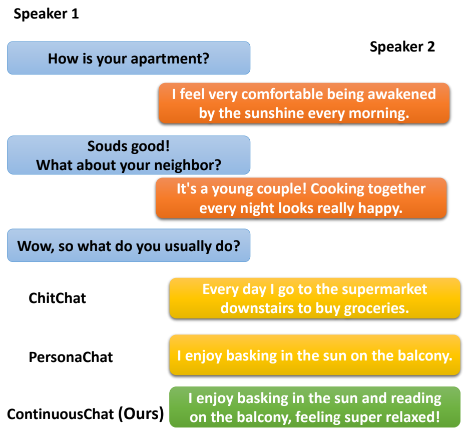

Figure 1: An example to illustrate our motivation.

One common solution is to endow chatbot with a configurable persona. Existing works about persona dialogue such as PersonaChat (Zhang et al., 2018; Dinan et al., 2020) have greatly facilitated the chatbot with configurable and persistent personalities. DuLeMon (Xu et al., 2022b) focuses not only on the consistency of the bot's own persona but also on the active construction and utilization of the user's persona in a long-term interaction.

We argue that most current open-domain dialogue datasets primarily feature "two strangers trying to get to know each other", and there is a lack of conversations designed to teach agents how to generate more engaging dialogues that increase users' willingness to continue interacting with the agent. Additionally, the language styles in these datasets are quite similar despite featuring various personas. Moreover, most personas lack detailed characteristics, making it challenging to create a lasting impression on users.

Motivated by the aforementioned limitations, we would like to investigate the following research question: *can we have better ways of increasing*

[*]Corresponding Author
[1]The dataset is available at https://drive.google.com/drive/folders/11GPd6N_gl15ihvwsMbO8nPF9TY6yHItl?usp=drive_link

Findings of the Association for Computational Linguistics: ACL 2024, pages 16408–16414
August 11-16, 2024 ©2024 Association for Computational Linguistics

users' willingness to continue talking to the agent? Figure 1 shows an example to illustrate the motivation of this paper.

In this paper, we introduce a dataset named CONTINUOUSCHAT, which consists of engaging conversations. We begin by collecting and refining personas, then expand these base personas into detailed personas through descriptions of experiences, daily life, future plans, and interesting stories. Subsequently, we develop these detailed personas into dialogues, infusing them with emotions and feelings. Finally, we rewrite the dialogues in specific styles using few-shot prompts, conditioned on handwritten style-specific examples.

To validate the usefulness of the created dataset, we benchmark Large Language Models (LLMs) on the CONTINUOUSCHAT dataset using both fine-tuning and in-context learning settings. Experiments with publicly available models demonstrate that our CONTINUOUSCHAT dataset effectively guides conversational agents in generating more engaging dialogues, thereby increasing users' willingness to continue conversations.

The key contributions of this paper are summarized as follows.

- We are the first to propose a novel task of increasing users' willingness to continue chatting with the agent.
- For this new task, we collect a dataset named CONTINUOUSCHAT, which is consists of attractive conversations generated by ChatGPT (OpenAI, 2021).
- Experiments over publicly available models demonstrate that our CONTINUOUSCHAT dataset is valuable in guiding conversational agents to generate more attractive dialogues and increase users' willingness to continue the conversations.

2 Related Work

Persona Dialogue. Kim et al. (2015) proposed a retrieval-based implicit role model to integrate personas and user interests into dialogue systems. Target response generation in implicit models is not easy to explain and control because roles are represented in the form of semantic role vectors. Qian et al. (2018) proposed an explicit persona model to generate consistent responses to given persona information. The role information of the machine includes name, gender, hobbies, etc. Furthermore, many persona chat datasets (Zheng et al., 2019; Mazare et al., 2018; Xu et al., 2022a) have been constructed for model developments. In particular, the introduction of the PersonaChat (Zhang et al., 2018; Dinan et al., 2020) dataset has broadly advanced the field, where the crowd-workers are simply asked to "*chat with the other person naturally and try to get to know each other*". DuLeMon (Xu et al., 2022b) ask the chatbot to actively remember and use the user's persona to improve conversational engagements and increase the intimacy between interlocutors in long-term interactions.

Item	Number
styles	13
Personas	1250
Detailed Personas	5000
Dialogues	25000
Utterances	302406

Table 1: Statistics of the CONTINUOUSCHAT dataset.

Style-Control Generation. Previous works have shown that focus on positive style result in engaging chats (Shuster et al., 2019). Niu and Bansal (2018) proposed a set of models for generating polite dialogues. The models are guided by a politeness classifier in generating responses. Su et al. (2021) first extracted a non-stylistic prototype from a generic dialogue system and then generate stylistic response via GPT-2. Zheng et al. (2021) attempted to using monolingual stylistic data to increase the style intensity of dialogue response. Dathathri et al. (2020) proposed a plug-and-play method (PPLM) to control text style, which is an iterative generation method using a classifier on top of a pre-trained generation model. Another work that is achieving fine-grained control using a very large architecture is the CTRL model (Keskar et al., 2019). The style conditioning relies on control codes obtained from the training data (meta-data). Lample et al. proposed a style transfer architecture using noisy encoders and decoders and style conditioning through an additional token.

3 The CONTINUOUSCHAT Dataset

The aim of this work is to enhance users' willingness to continue chatting with models. We collect a Chinese dialogue dataset named CONTINUOUSCHAT, which consists of engaging conversations initially generated by ChatGPT and subsequently revised by hand. Detailed information about the dataset is shown in Table 1. The data collection process comprises three stages.

Focus on 'I have a pug named Tea', talk about your past experience in 6 sentences, each sentence should not exceed 20 words, pay attention to control the number of words.

(a)

Character A's personality is...(*detailed persona*), please let character A and character B start a dialogue around A's personality, no less than 10 sentences, each sentence no more than 30 words, pay attention to control the number of words. Please use parentheses to mark the emotion of each sentence

(b)

Examples of cross-talk style are as follows:...(*Handwritten style Examples*) Please rewrite the following dialogue in cross-talk style, please do not change the format of the dialogue: ...(*Base Conversations*)

(c)

Figure 2: Examples of Prompts during (a) Personas Expansion (b) Base Continuous-Chat Generation (c) Style Continuous-Chat Generation

Categories	Examples
Hometown/Residence	I grew up in shanghai
Family/Partner	My dad and I have a bad relationship
Study/Work	I'm in college
Wish/Plans	I want to be a marine biologist
Opinions/Ideas	I think everyone is equal before the law
Feelings/Emotions	I'm very lonely
Hobbies	I like reading
Health	I am colorblind
Pets	I have a pug named Tea
Food	I am vegetarian
Idols	My favorite singer is Taylor Swift
Good at/Not good at	I'm not good at communication
Like/Dislike	I like snow
Habits	I always put on my left sock first
Dress	I put on green nail polish
Travel	I always travel alone
Social	I hang out with my friends on weekends
Others	I participated in a math competition

Table 2: Categories and examples of revised personas. Note that these personas are in Chinese, we translated them into English to make them easy to read.

Personas Collection and Expansion. We collect and revise personas, then expand these base personas into detailed personas by incorporating experiences, daily life, future plans, and interesting stories.

Base Continuous-Chat Generation. We then transform the detailed personas into dialogues, ensuring that emotions and feelings are expressed within these conversations.

Style Continuous-Chat Generation. Next, we use ChatGPT to rewrite the dialogues in specific styles by employing few-shot prompts, conditioned on handwritten style-specific examples.

3.1 Personas Collection and Expansion

There are numerous persona chat datasets that have been developed to train conversational models. We initially collect personas from the DuLeMon dataset (Xu et al., 2022b), as it includes personas in Chinese. We then refine these personas by:

(i) Rewriting personas with unclear meanings. For instance, we change *"She taught me to cook"* to *"Grandma taught me to cook"* to clarify the ambiguity of the former statement.

(ii) Rewriting personas that are not idiomatic. For example, we revise *"Both my parents are dead"* to *"Both my parents passed away"* to make it sound more natural.

(iii) Retaining and adding "imaginative" personas to make the dialogues more engaging, such as *"I want to be a dog when I grow up"* or *"I think if dogs are well trained, they can learn to read"*.

(iv) Keeping the one with the most vivid expression when multiple personas have the same meaning. For example, we kept *"I have a pug named Tea"* and removed *"I have 2 dogs"*.

Table 2 shows the categories and examples of revised personas. However, the revised personas are still not attractive enough due to the lack of details. In other words, an attractive persona should be more than a brief sentence; it should be a detailed paragraph. To address this issue, we expand the base personas into detailed personas by incorporating elements such as (i) experiences, (ii) daily life, (iii) future plans, and (iv) random interesting stories.

Due to the high cost and slow speed of manual writing, we use ChatGPT [2] to complete this. We manually write prompts for above 4 topics. The requirements for the number and length of sentences are also written in the prompts. An example of prompts is showed in Figure 2 (a).

3.2 Base Continuous-Chat Generation

ChatGPT is used to expand detailed-personas into conversations, conditioning on handwritten prompts. In particular, we ask the model to express emotions and feelings in dialogues, and mark the emotion after each sentence. An example of prompts is showed in Figure 2 (b).

[2] `https://platform.openai.com/docs/guides/gpt`

Style	Example
Martial-arts novel style	This good meeting is full of great excitement. When we meet again, let us drink together.
Chinese-Internet style	What bad thoughts can a kitten have?
Cross-talk style	Hey, guess what!
Elegant	The dew in the morning is just right for making tea. Let's go after drinking this cup of tea.
Literary	I look up at the moon, listen to the wind with my ears, there is wine in the pot, and my heart is at peace.
Lively	Good morning! How about going for a morning jog together?
Sweet	Such a fine weather, um suitable for a brighter style of music!
Imaginative	Take a deep breath, blow up a big balloon, put your sleepiness into the balloon, exhale-let the balloon fly away!
Childlike	This box... seems to be able to sleep in it.
Humorous	Listen, does the cry of the frog seem to be accompanied by the sound of rain? Quack quack.
Dramatic	There are such treasures hidden in this barren world!
Sentimental	I like maple leaves very much, but it's a pity that when the maple leaves are red, there are always many farewells.
Heroic	This kind of trivial matter, let me settle it!

Table 3: Examples of handwritten styles. Note that these examples are in Chinese, we translated them into English to make them easy to read.

Chinese-Alpaca-Pro-7B	Fluency↑	Engagingness↑	Consistency↑	Attraction↑	Stylization↑
BASE	4.40	4.20	3.85	3.80	3.35
PROMPT W/O FINE-TUNING	4.45	4.25	3.90	3.90	3.50
FINE-TUNING+PROMPT	**4.50**	**4.35**	**4.05**	**4.05**	**3.70**
ChatGLM2-6B	Fluency↑	Engagingness↑	Consistency↑	Attraction↑	Stylization↑
BASE	4.35	4.25	4.05	3.80	3.45
PROMPT W/O FINE-TUNING	4.40	4.30	4.00	3.90	3.55
FINE-TUNING+PROMPT	**4.50**	**4.35**	**4.10**	**4.15**	**3.80**

Table 4: Human evaluation of our baselines over CONTINUOUSCHAT datasets. PROMPT denotes "Few-shot Prompt", The best results are in **bold**.

3.3 Style Continuous-Chat Generation

Previous works have shown that focus on positive style result in engaging chats (Shuster et al., 2019). Therefore, we rewrite the conversations generated in the previous step into conversations with a positive style, to make them more attractive. We first picked out 10 common positive styles in Chinese from 217 styles (Shuster et al., 2019). In particular, we add 3 special Chinese language styles: Chinese-Internet style, cross-talk style and martial-arts novel style. We manually write examples for each style. ChatGPT was used to rewrite the dialogues in specific styles through few-shot prompt, conditioning on handwritten style-specific examples. Table 3 shows examples of handwritten style. Due to computational resource constraints, we cannot rewrite the conversations in all styles. For each conversation above, we randomly choose 4 out of 10 styles and 1 out of 3 special Chinese language styles as target styles. An example of prompts is showed in Figure 2 (c).

4 Experiments

We build multiple strong baselines in our experiments. Specifically, we implement different generative models in two ways: fine-tuning and in-context learning. we use Chinese-Alpaca-Pro-7B and ChatGLM2-6B for our dataset.

4.1 Implementation and Setup

We conduct experiments using our proposed CONTINUOUSCHAT dataset, which comprises 23,000 samples in the training set, 1,480 samples in the validation set, and 520 samples in the test set.

4.2 Evaluation Metrics

As automated metrics are notoriously poor for evaluating dialogue, we perform human evaluation over the generated response. Following Zhang et al. (2018), we consider three conventional criteria: **Fluency, Engagingness and Consistency**. In particular, we use **Attraction** to evaluate the ability to engage users in long-term chats, and use **Stylization** to assess the ability to generate dialogues of specific styles. We sample 520 instances(40 samples for each of 13 styles) for each model's evaluation and each sample was evaluated by three people. Each criterion is rated on five scales, where 1, 3, and 5 indicate unacceptable, moderate, and perfect performance, respectively. We report the average Fleiss's kappa score, 0.56, indicating annotators have reached moderate agreement.

We also introduce a special test that allowed volunteers to engage in multiple rounds of continuous conversations with the conversational agent until

(i) Base
User: Did you receive chocolates for Valentine's Day yesterday?
Agent: I got it, but I hope to receive your sincerity instead of the sweet chocolate.
(ii) Prompt w/o Fine-tuning
User: Did you receive chocolates for Valentine's Day yesterday?
Agent: Oops, I got it! However, I'm not sure if it's real chocolate because they turned into weird shapes.
User: Oh? That looks interesting! Who gave it to you?
Agent: It's from my dog! I think it must be trying to express its love for me by biting the chocolate into various strange shapes.
User: Haha, so cute! It seems your dog is also a creative gift deliverer.
Agent: Yes, although these "chocolates" look a little strange, I still feel warm inside because this is my dog's way of expression.
(iii) Fine-tuning + Prompt
User: Did you receive chocolates for Valentine's Day yesterday?
Agent: Oh, chocolate? received! But unfortunately, I bought it for myself.
User: Haha, it's a good idea to buy chocolates for yourself during the narcissistic holiday!
Agent: Yes, I think if I don't buy it for myself, I might not even receive the chocolate. What about you, did you receive any gifts?
User: Oh, I, besides receiving a box of chocolates yesterday, I also received a mysterious gift!
Agent: A mysterious gift? really interesting! Who sent it?
User: I don't know either. There is only a note on it that says: "To the cutest XX". However, I really don't know who sent it.
Agent: Wow, what a mystery! However, since you received the chocolate, at least you didn't spend the holidays empty-handed.
User: Haha, yes! Even if you don't know who sent it, at least you have a sweet surprise. Next Valentine's Day, let's play a mystery gift-giving game together!
Agent: OK, that sounds interesting! However, I have to start saving money to buy myself a gift next time.

Figure 3: Examples of a user having sustained multi-round conversations with conversational agents in different settings, with rounds that the user finds uninteresting marked in red.

Chinese-Alpaca-Pro-7B	Average Rounds↑
BASE	5.92
PROMPT W/O FINE-TUNING	7.06
FINE-TUNING+PROMPT	**8.14**
ChatGLM2-6B	**Average Rounds↑**
BASE	6.08
PROMPT W/O FINE-TUNING	7.24
FINE-TUNING+PROMPT	**8.40**

Table 5: The average number of dialogue duration rounds for each conversational agent. The best results are in **bold**.

the volunteers subjectively felt that the conversation became boring. We count **the average rounds** of dialogue for each conversational agent to verify whether our dataset can help agents generate more engaging conversations and increase users' willingness to continue talking to the agent.

4.3 Results and Analysis

Table 4 reports the overall experimental results. Our FUNE-TUNING + PROMPT setting achieve better performance consistently across all human evaluation metrics, especially **Attraction** and **Stylization**, demonstrating that our approach could utilize more of the CONTINUOUSCHAT data, help guiding conversation agents to generate more attractive dialogues and increase users' willingness to continue the conversations. Table 5 shows that our dataset could help conversation agents to have more rounds of conversation with the user.

4.4 Case Study

Figure 3 shows examples of a user having a sustained multi-round conversation with agents in three different settings. We can see that the agent in the Base setting (which does not use our dataset) is more likely to cause users to lose interest in the conversation. Either fine-tuning or in-context learning using our proposed dataset can help the agent generate more engaging dialogues. This shows that fine-tuning the model using our proposed dataset can help conversational agents generate more engaging conversations and increase users' willingness to continue talking to the agent.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants No. 61972290.

Limitations

In this paper, we propose a novel task aimed at increasing users' willingness to continue engaging with the agent, and we introduce a dataset named CONTINUOUSCHAT. A primary limitation of our work is that the CONTINUOUSCHAT dataset is in Chinese, reflecting Chinese culture and expression habits. Consequently, direct translation into other languages may not be feasible. Adapting our dataset to other languages will require further investigation. Additionally, while the emotions of each utterance were annotated by ChatGPT, our experiments did not leverage this information. We hope that future research will fully utilize the emotion labels in the dataset.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Yonghee Kim, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2015. Acquisition and use of long-term memory for personalized dialog systems. In *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction: Second International Workshop, MA3HMI 2014, Held in Conjunction with INTERSPEECH 2014, Singapore, Singapore, September 14, 2014, Revised Selected Papers 2*, pages 78–87. Springer.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Pierre-Emmanuel Mazare, 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.

Fei Mi, Yitong Li, Yulong Zeng, Jingyan Zhou, Yasheng Wang, Chuanfei Xu, Lifeng Shang, Xin Jiang, Shiqi Zhao, and Qun Liu. 2022. Pangubot: Efficient generative dialogue pre-training from pre-trained language model. *arXiv preprint arXiv:2203.17090*.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

OpenAI. 2021. Chatgpt. Computer software.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4279–4285.

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.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526.

Yixuan Su, Yan Wang, Deng Cai, Simon Baker, Anna Korhonen, and Nigel Collier. 2021. Prototype-to-style: Dialogue generation with style-aware editing on retrieval memory. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2152–2161.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

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

System Demonstrations, pages 270–278, Online. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021. Stylized dialogue response generation using stylized unpaired texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14558–14567.