Recent Trends in Personalized Dialogue Generation: A Review of Datasets, Methodologies, and Evaluations

Yi-Pei Chen*†, Noriki Nishida†, Hideki Nakayama*, Yuji Matsumoto†
* University of Tokyo, † RIKEN AIP
ypc@g.ecc.u-tokyo.ac.jp, nakayama@ci.i.u-tokyo.ac.jp
{noriki.nishida, yuji.matsumoto}@riken.jp

Abstract

Enhancing user engagement through personalization in conversational agents has gained significance, especially with the advent of large language models that generate fluent responses. Personalized dialogue generation, however, is multifaceted and varies in its definition – ranging from instilling a persona in the agent to capturing users' explicit and implicit cues. This paper seeks to systemically survey the recent landscape of personalized dialogue generation, including the datasets employed, methodologies developed, and evaluation metrics applied. Covering 22 datasets, we highlight benchmark datasets and newer ones enriched with additional features. We further analyze 17 seminal works from top conferences between 2021-2023 and identify five distinct types of problems. We also shed light on recent progress by LLMs in personalized dialogue generation. Our evaluation section offers a comprehensive summary of assessment facets and metrics utilized in these works. In conclusion, we discuss prevailing challenges and envision prospect directions for future research in personalized dialogue generation.

Keywords: Personalized dialogue systems, personalized response generation, persona-based conversation

1. Introduction

Personalization can enhance a user's engagement with conversational agents (Zhang et al., 2018a; Kwon et al., 2023). The ability of large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023) to generate fluent and coherent responses to human queries underscores the importance of building personalized systems that cater to each individual's background and preferences more than ever before.

However, *personalization* remains an open question, with varying definitions among different individuals. Figure 1 illustrates the three scenarios currently being explored in research on personalized dialogue generation. This may involve endowing the agent with a persona, modeling the other party's persona, or both. In this context, a persona refers to characteristics such as personal background, interests, or behaviors that shape the identity or personality of the user or the agent. Personalized response generation can be seen as a conditional text generation task where a response is generated based on the given context and conditioned on either the speakers' explicitly provided persona or implicit attributes embedded in the dialogue history.

We aim to find out what is being personalized (**Dataset**), how the dialogue systems implement personalization (**Methodology**), and how previous research evaluates the personalization (**Evaluation**) in this systematic survey. In the Dataset Section (Sec. 2), we introduce 22 datasets, including benchmark datasets that were frequently used in previous personalized dialogue research and recently published datasets that propose more features to add on the existing ones. In the Methodology Section (Sec. 3), we center on 17 works published in recent three years from 2021 to Oct. 2023 at top conferences including ACL, NAACL, EMNLP, AAAI, etc., based on the keyword searching and the related works in each paper. In the Evaluation Section (Sec. 4), we summarize the evaluated

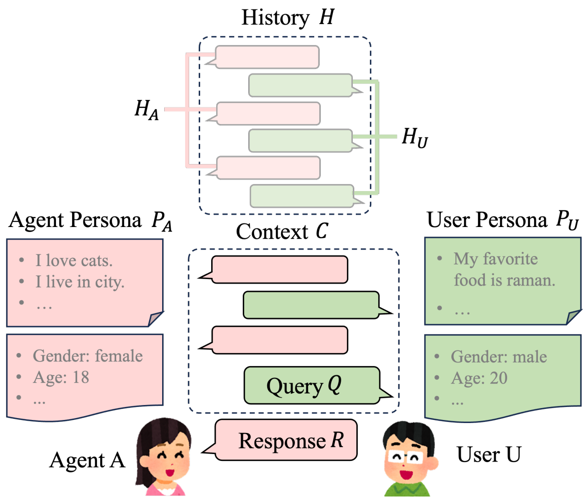

Figure 1: An overview of personalized dialogue generation. Assumed that the conversation is performed by two speakers, i.e., an agent A and a user U, the goal is to generate the response R given the dialogue context C or the last utterance Q, plus the persona of the agent or user (P_A or P_U) (explicit), or utterance histories of them (H_A or H_U) (implicit).

LREC-COLING 2024, pages 13650–13665
20-25 May, 2024.

Dataset		Lang	Persona	Source	D	U/D	Role	Pub.	P Ground.	Multi Session	Feature
PersonaChat	(2018a)	EN	Description	Crowd	10.9K	**14.9**	H-H	✔	✘	✘	
ConvAI2	(2020)	EN	Description	Crowd	11.0K	**15.0**	H-H	✔	✘	✘	
BlendedSkillTalk	(2020)	EN	Description	Crowd	6.8K	**11.2**	H-H	✔	✘	✘	KG, EMP
MSC	(2022a)	EN	Description	Crowd	24.0K	**12.5**	H-H	✔	✘	✔	
FoCus	(2022)	EN	Description	Crowd	13.5K	**11.3**	H-A	✔	✔	✘	KG, ConvQA
PersonaMinEdit	(2021a)	EN	Description	Crowd	121.8K	2.0	H-H	✔	✘	✘	OOD
IT-ConvAI2	(2022)	EN	Description	ConvAI2	1.6K	2.0	H-H	✔	✘	✘	OOD
Reddit	(2018)	EN	Description	Reddit	**700M**	-	H-H	✘	✘	✘	
PEC	(2020)	EN	Description	Reddit (H, O)	355.0K	2.4	H-H	✔	✘	✘	EMP
MPChat	(2023)	EN	Description	Reddit (648*)	15.0K	2.8	H-H	✔	✔	✘	Multi-modal
PER-CHAT	(2021b)	EN	KV + History	Reddit (A)	**1.5M**	2.0	H-H	✔	✘	✘	QA
Reddit (DialoGPT)	(2020)	EN	UserID	Reddit	**147M**	-	H-H	✔	✘	✘	
Persona Reddit	(2021)	EN	UserID	Reddit	**3.1M**	-	H-H	△	✘	✘	
Baidu PersonaChat	N/A	ZH	Description	Crowd	24.5K	**16.3**	H-H	✔	✘	✘	
DuLeMon	(2022b)	ZH	Description	Crowd	27.5K	**16.3**	H-A	✔	✔	✔	
LiveChat	(2023)	ZH	Description + KV	Douyin	1.3K	2.0	H-H	✔	✘	✘	LiveStream
WD-PB	(2018)	ZH	KV	Weibo	76.9K	2.0	H-H	△	✘	✘	
Personal Dialog	(2019)	ZH	KV	Weibo	**20.8M**	2.7	H-H	✔	✘	✘	
PchatbotW	(2021)	ZH	UserID	Weibo	**139.4M**	2.0	H-H	△	✘	✘	
PchatbotL	(2021)	ZH	UserID	Forums	**59.4M**	2.0	H-H	△	✘	✘	Judiciary
MSPD	(2023)	KR	Description	Crowd	53.9K	**11.2**	H-A	✘	✔	✔	
XPersona	(2021)	Multi	Description	ConvAI2	3.3K	15.6	H-H	✔	✘	✘	

Table 1: Dataset summary. Persona can be represented by descriptive sentences (Description) or key-value dictionary (KV). Data source is mainly from crowdsourcing (Crowd) and Reddit (subreddit) (H: happy, O: offmychest, A: AskReddit, 648*: from 648 subreddits, see the paper for the list). D: number of dialogues. U/D: average utterance per dialogue. H-H: human-human conversation. H-A: human-agent conversation. △: Available upon request. Persona Grounding (P Ground) means each utterance has the label of which persona sentence the utterance is grounding on. KG: knowledge. EMP: empathy. OOD: out-of-distribution persona. We added background colors to emphasize the less-frequent options.

aspects and evaluation metrics from each work.

We address the challenges and potential future trajectories in terms of Datasets, Methodologies, and Evaluation in the Discussion Section (Sec. 5). We believe that the primary issues with current datasets revolve around their size, quality, and diversity. In terms of methodology, we stress the need to critically examine the assumptions underpinning approaches to personalization. Finally, we advocate for a standardized evaluation benchmark equipped with advanced metrics to provide a fair assessment of contributions in this domain.

2. Datasets

In this section, we first review datasets that have been used in personalized dialogue generation literature. We then discuss the characteristics of the datasets, with a focus on persona representations and domain and language biases.

2.1. Datasets Review

Table 1 summarizes the datasets. One of the first and most widely used dataset in personalized dialogue research is the **PersonaChat** dataset (Zhang et al., 2018a), which is also used in The Second Conversational Intelligence Challenge (ConvAI2) (Dinan et al., 2020). Nine out of eighteen works introduced in Sec. 3 evaluated their systems on PersonaChat/ConvAI2. PersonaChat consists of 10.9K English dialogues in total, which were collected through crowdsourcing. Each dialogue in PersonaChat consists of two human speakers trying to know each other, and about five persona descriptive sentences are provided to each speaker.

The **FoCus** dataset (Jang et al., 2022) introduced *persona grounding*, indicating the specific persona sentence to which each utterance is anchored. With persona grounding label, model can learn to extract the most relevant personal information from persona. Datasets such as **MPChat** (Ahn et al., 2023), **DuLeMon** (Xu et al., 2022b), and **MSPD** (Kwon et al., 2023) also contain persona grounding.

Some datasets introduce multi-session dialogues. The **MSC** dataset (Xu et al., 2022a) is similar to PersonaChat, except that MSC has four to five sessions for the same pair of speakers. **DuLeMon** (Xu et al., 2022b) and **MSPD** (Kwon et al., 2023) also introduced multi-session into their datasets.

Some datasets also augment persona with other features such as knowledge, empathy, and vision. The **BlendedSkillTalk** (BST) dataset (Smith et al., 2020) differs from PersonaChat in that BST endows agents persona, knowledge (KG), and empathy (EMP) by combining PersonaChat with Wizard of Wikipedia (Dinan et al., 2019) and Empathetic Dialogues (Rashkin et al., 2019). Each agent is given two persona sentences and each utterance is labeled with the type of bias they were grounded on (persona, KG, or EMP). The **Persona-based Empathetic Conversation** (PEC) dataset (Zhong et al., 2020) also investigated the impact of persona

Dataset	Attribute Keys
WD-PB (Qian et al., 2018)	gender, location, age, name, weight, constellation
Personal Dialog (Zheng et al., 2019)	gender, location, age, self-description, interest tags
PER-CHAT (Wu et al., 2021b)	gender, location, self-description, ID, pets, family, favorites, partner, possessions
LiveChat (Gao et al., 2023)	gender, location, age, ID, character, fans number, live time, reply barrage, audiences, skill

Table 2: The types of personal attributes collected in each key-value dataset.

on empathetic responding. PEC differs from BST for extracting persona from Reddit. The **MPChat** dataset (Ahn et al., 2023) introduced the first multi-modal persona, where the persona is not only text descriptions but also paired with an image.

Due to the limited scale of crowdsourced datasets, the persona distribution in real-world data often exceeds that of the datasets. **PersonaMinEdit** (Wu et al., 2021a) and **Inadequate-Tiny-ConvAI2** (IT-ConvAI2) (Liu et al., 2022) are specifically designed to test the generation grounding on unseen personas. Specifically, these datasets ensure that their test set contains only unseen personas the response should conditioned on.

2.2. Facets

2.2.1. Persona Representation

It is not obvious how persona information should be represented. Through the literature survey, we found that different datasets employ different ways of representing persona information, and that the persona representation can be classified into three categories: (1) persona description, (2) key-value attributes, (3) user ID and comment histories.

Most datasets employ **descriptive sentences** as the persona representation (Dinan et al., 2020; Mazaré et al., 2018; Smith et al., 2020; Zhong et al., 2020; Wu et al., 2021a; Xu et al., 2022a,b; Liu et al., 2022; Jang et al., 2022; Ahn et al., 2023; Kwon et al., 2023). For example, PersonaChat (Zhang et al., 2018a) contains 5 descriptive sentences for each speaker. These datasets primarily recruited annotators to chat based on given persona descriptions, thus avoiding privacy concerns. Mazaré et al. (2018) extracted persona descriptions from Reddit using heuristic rules to gather large datasets, which was followed by Zhong et al. (2020) and Ahn et al. (2023).

Some datasets represent personal information using **sparse key-value attributes** (Qian et al., 2018; Zheng et al., 2019; Wu et al., 2021b; Gao et al., 2023). The examples of key-value attributes are shown in Table 2. For example, WD-PB (Qian et al., 2018) defines 6 attribute keys like gender, location, and age, and the values corresponding to these keys are recorded as persona information for each user. In addition to key-value attributes, LiveChat (Gao et al., 2023) provides dense persona description extracted by a rule-based method similar to Mazaré et al. (2018), and PER-CHAT (Wu et al., 2021b) provides query-related comment histories extracted by a pretrained IR system as additional personal information. They first define what attribute types they want to extract, then collect the results from posts/responses using regular expression (Qian et al., 2018; Wu et al., 2021b), from the profile provided by the users (Zheng et al., 2019; Gao et al., 2023), or recruit annotators to label from the given context (Gao et al., 2023).

Some datasets collected from social platform only provide **speaker/user IDs** as persona information (Zhang et al., 2020; Qian et al., 2021; Zeng and Nie, 2021). They assume that the speaker IDs are used to retrieve comment histories of the corresponding users in the social platform. The speaker IDs have also been used in other tasks such as speaker identification. They consider the same user's utterances contain implicit personal information and the personalization can be measured by the similarity between the generated response and the ground-truth response.

2.2.2. Domain and Language Biases

Personalized dialogue generation is an open-domain task: Human speakers are allowed to talk whatever topics they like. However, through the literature survey, we found that there are domain differences between datasets. PchatbotL (Qian et al., 2021) was collected from Chinese judicial forums. PER-CHAT (Wu et al., 2021b) crawled from Subreddit AskMeQuestion is more similar to a question answering (QA) dataset. FoCus (Jang et al., 2022) asked annotators to discuss about landmarks from Google Landmarks Dataset v2 (Weyand et al., 2020). It is more like a conversational question answering dataset where a user asks questions about a landmark and the agent answers, rather than a natural conversation between humans. LiveChat (Gao et al., 2023) gathered data from live streaming on Douyin (Chinese TikTok). There might be multiple people (multi-party) response to the streamer, while the streamer only response to one, making it a 1-1 dialogue.

Also, persona dialogue generation is not limited to specific languages; however, the languages of the datasets are highly biased. XPersona (Multilingual Persona-Chat) (Lin et al., 2021) translates a portion of PersonaChat into Chinese, French, Indonesian, Italian, Korean, and Japanese. How-

Model		Whose Persona	Approaches	Dataset	Model Input			NLI
					D	P	Other	
UA-CVAE	(2022)	Self	Coherence	CAI, ED	C	P_A		
BoB	(2021)	Self	Consistency	CAI, PD	Q	P_A		✔
PCF	(2023a)	Self	Consistency, Coherence	CAI, PD	Q	P_A		✔
LMEDR	(2023b)	Self	Consistency, Coherence	PC, AVSD	Q	P_A		✔
SimOAP	(2023a)	Self	Consistency, Coherence	PC	C	P_A		✔
PAA	(2023)	Self	Balance	CAI	C	P_A		
D3	(2022)	Self	Data Scarcity	PC	-	-		✔
GME	(2021a)	Self	Long-tail	PME, BST	Q	(P_A)	P_{OOD}	
PS-Transformer	(2022)	Self	Select , OOD, Long-tail	CAI, IT-CAI	Q	P_A	P_{OOD}, P Pool	✔
DHAP	(2021)	Self	Unknown	PCW, Reddit-D	Q	H_A		
FoCus	(2022)	Other	Select	FoCus	C	P_U	Knowledge	
INFO	(2022)	Other	Select	FoCus	C	P_U	Knowledge	
WWH	(2023)	Other	Select, Balance	MSPD	C	P_U		
IUPD	(2022)	Other	Unknown	CAI	C	(P_U)		
CLV	(2023)	Other	Unknown	CAI, Baidu PC	Q	(P_U)		
MSP	(2022)	Other	Unknown	PCW, Reddit-D	Q	H_U	H_{sim}	
DuLeMon	(2022b)	Both	Select, Unknown	DuLeMon	C	H_A, H_U		

Table 3: An overview of Personalized Dialogue Systems reviewed in Sec 3. We show (1) **whose persona** each model aims to represent (Self: agent; Other: user), (2) **approaches** involved in each work (Sec. 3.2), (3) **datasets** they are trained on (PC: PersonaChat; CAI: ConvAI2; PD: PeresonalDialog; PCW: PChatbotW; Reddit-D: Reddit (DialogGPT); ED: EmpatheticDialog; PME: PersonaMinEdit; BST: BlendedSkillTalk), (4) **model inputs**, i.e., dialogue types D (C: context; Q: query), persona types P (P_A: agent persona; P_U: user persona; H_A: agent history; H_U: user history), other modalities (P_{OOD}: out-of-distribution persona; P Pool: additional personas pool), (5) and whether additional **NLI** data/model is used. The parentheses like "(P_A)" indicate that the persona information is only provided during training.

ever, the number of dialogues in each language is very limited. For example, there are only 280 dialogues for Italian. There are other translations of PersonaChat, such as Japanese (Sugiyama et al., 2021) and Korean[1]. Other than translation, Baidu constructed and released the Chinese PersonaChat dataset [2] which is similar to PersonaChat. DuLeMon (Xu et al., 2022b) and Multi-Session Personalized Dialogue (MSPD) (Kwon et al., 2023) are the Chinese and Korean version of Multi-Session Chat (MSC) (Xu et al., 2022a).

3. Methodology

In this section, we first introduce the task definition of personalized dialogue generation, and then discuss recent methodology advances in personalized dialogue generation published at top conferences from 2021 to 2023. Finally, we review recent studies on large language models in personalised dialogue generation.

3.1. Problem Statement

In this survey, we focus on the bilateral conversation, i.e., conversation between two parties. As shown in Fig. 1, we define two speakers as an agent A and a user U with the current dialogue context C, where the last utterance of C is defined as the query Q uttered by the user U. There might also be past dialogue history H, where the history utterances by the agent A and the user U are denoted as H_A and H_U, respectively.

The goal of personalized dialogue generation is to generate a response R as an agent conditioned on the input dialogue D and the persona P. The input dialogue D can be a single query Q (i.e., the last utterance by the user) or the dialogue context C. The persona P can be in various representation formats, such as descriptive sentences and sparse key-value attributes, as seen in Section 2. Persona P for agent A is denoted as P_A, while that for user U is represented as P_U.

There are two main streams of the personalized dialogue generation. One direction endows the agent its own persona P_A, and focuses on generating R that is consistent with persona P_A and coherent with the dialogue context C. In most studies, P_A is explicitly provided as an input; one exception is Ma et al. (2021), in which P_A is not provided and the agent's history responses H_A are considered as the persona information.

The other direction aims to model the user's persona P_U to generate responses that better tailored to the user's needs. The emphasis is often on selecting segments from the provided P_U that are most relevant to Q, or on establishing P_U when it

[1] https://aihub.or.kr/

[2] https://www.luge.ai/#/luge/dataDetail?id=38

is not directly given. In cases where P_U is absent, one might derive explicit or implicit P_U from the user's dialogue history H_U, or infer implict P_U via conditional variational inference.

3.2. Approaches

Through the survey, we found that recent approaches to personalized dialogue generation can be classified into 5 groups based on their motivations and target issues: Consistency and Coherence, Persona-Context Balancing, Relevant Persona Selection, Unknown Persona Modeling, and Data Scarcity. Note that a single paper may cover more than one of these issues.

As a side note, although a closely related line of research focuses on extracting personas from dialogues, such as Zhu et al. (2023), it falls beyond the scope of this paper, which is concentrated on personalized dialogue generation.

3.2.1. Consistency and Coherence

As mentioned in the previous section, most research on endowing models with personas focuses on generating responses that are simultaneously consistent with the given persona and coherent with the context.

Uncertainty Aware CVAE (UA-CVAE) (Lee et al., 2022) tackled the coherence problem using conditional variational autoencoder (VAE) training. They proposed to generate response R conditioned on the context C, the agent persona P_A, and the latent variable z. The latent variable z is sampled form latent Gaussian distribution $p(z|C, P_A)$, where the variance of z acts as an approximation to the uncertainty in the input (C, P_A).

Natural Language Inference (NLI) is commonly used to solve the consistency problem, which predicts whether a premise and a hypothesis are entailed, neutral, or contradictory. BoB (Song et al., 2021) trained a decoder with NLI data to ensure the consistency and minimize contradiction between the response R and the agent persona P_A.

Based on BoB, PCF (Wang et al., 2023a) further added another NLI module between R and the query Q to maintain the coherence of the dialogue. LMEDR (Chen et al., 2023b) fine-tuned a pretrained NLI model with two additional matrix parameters which act as additional "memory" for language modeling, one for consistency and the other for coherence. SimOAP (Zhou et al., 2023a) demonstrated that responses with high probabilities aren't always superior to those with lower probabilities. Consequently, their approach involves generating an extensive list of response candidates and post-filtering. The candidates are first filtered for coherence with C using the TF-IDF method (Salton and Buckley, 1988), then selected for consistent with P_A through a pretrained NLI model.

3.2.2. Persona-Context Balancing

Since not all responses need personalization, deciding when to condition more on the context and when to weave in more personal information into the response is important. Persona-Adaptive Attention (PAA) (Huang et al., 2023) separately encode P_A and C, then design attention mechanism to combine them dynamically. WWH (Kwon et al., 2023) integrates non-personalized datasets with personalized dataset, adjusting training data sampling based on each dataset's size to yield more natural responses.

3.2.3. Relevant Persona Selection

While not all information in the given persona $P \in \{P_A, P_U\}$ is related to the dialogue $D \in \{Q, C\}$, selecting the most relevant persona sentence becomes crucial to generate a natural and engaging response. PS-Transformer (Liu et al., 2022) and FoCUS (Jang et al., 2022) trained a binary classifier for each persona sentence to evaluate their likelihood of being utilized. In contrast, INFO (Lim et al., 2022) obtained the weights over all P_A candidates via a multi-class classifier. DuLeMon (Xu et al., 2022b) and WWH (Kwon et al., 2023) trained LLM to discriminate the negative persona sentences from the positive one based on C. Note that a dataset with persona-grounding label - each utterance is associated with a specific persona attribute - is usually required for learning relevant persona selection.

3.2.4. Unknown Persona Modeling

In the case when persona P is not explicitly given, the personal information could be extracted from dialogue histories of the speaker (H_A or H_U) (Ma et al., 2021; Zhong et al., 2022; Xu et al., 2022b), or implicitly modeled by latent variables (Cho et al., 2022; Tang et al., 2023).

DHAP (Ma et al., 2021), MSP (Zhong et al., 2022), and DuLeMon (Xu et al., 2022b) extract explicit (DuLeMon, MSP) or implicit (DHAP) personal information from dialogue histories. DuLeMon was the first to conduct personal information management for **both** U and A, training a classifier to determine whether a clause in an utterance contains personal information. Following a similarity check, clauses containing persona information were then added or updated in either P_A or P_U. MSP derived the user persona P_U from the dialogue history of the user H_U and similar users H_{sim}. Histories unrelated to the query Q are filtered out. From the remaining histories, the top k

tokens are selected based on the attention weight between Q and the histories. DHAP encoded H_A as implicit agent persona P_A and constructed a personalized vocabulary consisting of words from H_A. During decoding, DHAP switched between this personalized vocabulary and a general one.

Both IUPD (Cho et al., 2022) and CLV (Tang et al., 2023) model implicit user persona from D via conditional variational inference (CVAE) training. CVAE has been used in dialogue generation to address the challenge of producing diverse responses for a single query. It also facilitates response generation under various conditions (Sohn et al., 2015; Zhao et al., 2017; Song et al., 2019; Chen et al., 2022)[3].

IUPD proposed both a persona latent variable Z_P and a fader latent variable Z_α, along with special tokens tok_P and tok_α for the input. The Z_P variable captures the latent distribution of P_U, thereby connecting the context C to the response R. Meanwhile, Z_α measures the extent to which Z_P's persona information impacts R under C. The response generation is expressed as: $p(R|Z_P, Z_\alpha, C) = p(R|Z_P, Z_\alpha, C)p(Z_P|C)p(Z_\alpha|Z_P, C)$, where $p(R|Z_P, Z_\alpha, C)$ is the generator and $p(Z_P|C)$, $p(Z_\alpha|Z_P, C)$ are prior networks. For Z_P, inputs to it's prior and recognition networks combine the variable token with the condition, namely $[tok_P, C]$ and $[tok_P, P]$. For Z_α, these are $[tok_\alpha, Z_P, C]$ and $[Z_\alpha, P, R]$.

CLV also proposed a persona latent variable Z_P and additionally, a response latent variable Z_R. They assume that Z_P and Z_R are independent. The response generation is formulated as: $p(R|Z_P, Z_R, C) = p(R|Z_P, Z_R, C)p(Z_P|C)p(Z_R|C)$. In contrast to IUPD, CLV's prior and recognition networks for Z_P utilize Q and $[Q, P]$. For Z_R, the inputs are Q and $[Q, R]$.

Note that although UA-CVAE (Lee et al., 2022) also using CVAE, they have the persona P_A as input during inference time and is described in Sec. 3.2.1.

3.2.5. Data Scarcity

As shown in the Dataset section (Sec. 2), crowd-sourced persona-based datasets usually have more dialogue context but are notable for their limited size. Data augmentation is a naive answer to solve the data scarcity problem. D^3 (Cao et al., 2022) is a model-agnostic method that purely manipulates the data. They kept only persona-related (Q, R) dialogues, removed irrelevant P_A which are not entailed by R, and enlarged the number of persona-related dialogues to 1.8 times and personas to 3 times by BERT, GPT2, and back translation technique (Sennrich et al., 2016). Although not trained with NLI data, pretrained NLI models have been extensively used to judge the consistency between the augmented $\tilde{P}_A$ and $\tilde{R}$, as well as to evaluate the coherence between the augmented $\tilde{Q}$ and $\tilde{R}$.

Data scarcity also evokes the problem of out-of-distribution (OOD) personas. That is, the limited data provides the agent with a restricted persona P_A and as a result, personas related to certain real-world queries might not be present in P_A (Liu et al., 2022) . To solve this problem, Liu et al. (2022) proposed to retrieve unseen persona from an external persona pool based on a NLI model.

Nevertheless, even the unseen persona is provided, it is difficult for the model to ground on the out-of-distribution persona, i.e. the long-tail problem (Liu et al., 2022). GME (Wu et al., 2021a) enforced the grounding on unseen persona P_{OOD} during inference by masking persona-spans in the original response and re-generate the response conditioned on P_{OOD}, Q, and the masked response. Liu et al. (2022), on the other hand, solve this problem by training a persona selection module over P_A and P_{OOD} (which learned to select P_{OOD}), and generating the response weighted on P.

3.3. Large Language Models and In-Context Learning

Given the rising prominence of ChatGPT, we are motivated to examine the impact of large language models (LLMs) and in-context learning on personalized dialogue generation. As demonstrated in Salewski et al. (2023); Jiang et al. (2023), LLMs can reflect personas or personality traits provided in prompts, evident in corresponding personality tests or tasks like writing or reasoning. Recent works have prompted LLMs as various characters for multi-agent simulation or collaboration (Chen et al., 2023c; Qian et al., 2023; Park et al., 2023; Wang et al., 2023b). Tu et al. (2023) prompted ChatGPT with MBTI personalities to create a dialogue dataset between characters. Chen et al. (2023a) pretrained their own LLM with persona information augmented to the original dialogue context and demonstrated that prompting such a model improves the agent's persona consistency.

While there's enthusiasm around using prompts with LLMs to imbue them with personas, evaluations typically focus on the success rate of assigned tasks, overlooking the quality of LLM-generated conversations. Furthermore, to the best of our knowledge, existing works have centered on conferring agent personas P_A, with no studies exploring the use of in-context learning to extract or model unknown user personas P_U.

[3]Not included in this work because they are not specifically designed for personalized dialogue generation.

Model		Fluency	Diversity	Coherence	Personalization
UA-CVAE	(2022)	PPL, ROUGE, METEOR	Dist-1, Dist-2, Dist-3	UE-Score	✘
BoB	(2021)	PPL	Dist-1, Dist-2, Dist-Avg	✘	Delta Perplexity
PCF	(2023a)	PPL, BLEU	Dist-1, Dist-2, Dist-Avg	UE-Score	C-Score
LMEDR	(2023b)	PPL, F1	Dist-1, Dist-2	✘	C-Score
SimOAP	(2023a)	PPL	Dist-1, Dist-2, Rep	TF-IDF	C-Score
PAA	(2023)	PPL, BLEU, F1	Dist-1, Dist-2	✘	✘
D3	(2022)	PPL,BLEU,NIST-4, BERTScore	Dist-1, Dist-2, Dist-3, Entropy-n	✘	C-Score
GME	(2021a)	BLEU	✘	✘	C-Score
PS-Transformer	(2022)	BLEU, ROUGE, CIDEr	✘	✘	✘
DHAP	(2021)	BLEU, ROUGE, EMB	Dist-1, Dist-2	✘	Persona-F1, Persona-Coverage
FoCus	(2022)	PPL, BLEU, ROUGE, chrF++	✘	✘	P-Grounding-Acc
INFO	(2022)	BLEU, ROUGE, BERTScore, chrF++	✘	✘	P-Grounding-Acc, P-Grounding-F1
WWH	(2023)	PPL	✘	✘	Persona-F1, Persona-Coverage
IUPD	(2022)	PPL	Dist-1, Dist-2	✘	Persona-Distance
CLV	(2023)	BLEU, ROUGE	Dist-1, Dist-2	Coh-Con-Score	C-Score*
MSP	(2022)	BLEU, ROUGE, EMB	Dist-1, Dist-2	✘	Persona-F1, Persona-Coverage
DuLeMon	(2022b)	PPL, BLEU, F1	Dist-1, Dist-2	✘	✘

Table 4: Evaluation metrics used in each model introduced in Sec. 3. EMB: embedding metric. C-Score*: a different implementation of C-Score.

4. Evaluation

Personalized dialogue generation literature typically assess the quality of the generated responses across various dimensions. The most commonly examined dimensions include fluency, diversity, coherence, and personalization.

4.1. Fluency

The fluency evaluation usually refers to the common generation metrics. *Perplexity (PPL)* is often reported as an indication of fluency. Besides, most works measure the similarity between the generated response to the reference response. The most popular similarity metrics include the **lexical overlap metrics**, e.g., *F1*, *BLEU* (Papineni et al., 2002), *ROUGE* (Lin, 2004), *NIST* (Doddington, 2002), *METEOR* (Banerjee and Lavie, 2005), *chrF++* (Popović, 2017), and **representation-based metrics**, e.g., *bag-of-word embedding score* (Chan et al., 2019) and *BERTScore* (Zhang* et al., 2020). Other metric like *CIDEr* (Vedantam et al., 2015) has also been used.

4.2. Diversity

Distinct-1 (Dist-1) and *Distinct-2 (Dist-2)* (Li et al., 2016) are the most widely adopted metrics for diversity evaluation. 12 out of 17 papers in Sec. 3 includes Dist-1 and Dist-2 in their evaluation. Entropy-n (Zhang et al., 2018b) is the entropy derived from a sentence's n-gram distribution, measuring the uniformity of the empirical n-gram distribution. *Repetition Rate (Rep)* (Zhou et al., 2023a) is proposed for evaluating diversity at the sentence level, which counts the number of identical responses in the candidates that differ from the ground truth. Comparing the repetition of candidates, this metric is specifically designed for their post-filtering method over the candidate responses.

4.3. Coherence

Coherence refers to the logical and meaningful continuity of a conversation. A Coherent response ensures that the conversation makes sense and flows naturally from the previous turn. Although coherence evaluation metrics have been proposed (Ghazarian et al., 2022; Ye et al., 2021), they were used in none of the papers in Sec 3. Many studies omit the coherence evaluation and consider that the Fluency metrics described in the previous section can also measure the coherence of dialogues.

Zhou et al. (2023a) calculate the cosine similarity between the TF-IDF vectors (Salton and Buckley, 1988) of the context and the response as a measure of coherence evaluation. *Coherence-Consistency Score (Coh-Con.Score)* (Tang et al., 2023) measures both dialogue coherence and persona consistency simultaneously. It takes the query Q, response R, and persona P as input and assigns 2, 1, 0 for the following scenarios respectively: both (P, R) and (Q, R) are entailed, only (P, R) is entailed, and otherwise.

Utterance Entailment score (UE-Score) (Lee et al., 2022) compute the NLI score between utterance and response as the coherence score.

Although both Coh-Con.Score and UE-Score utilize NLI model, the pretrained data and backbone are different. While Coh-Con.Score using RoBERTa (Liu et al., 2019) trained on Dialogue NLI dataset (DNLI) (Welleck et al., 2019) and finetuned on ConvAI2 and Baidu PersonaChat as the NLI model for English and Chinese, UE-Score finetuned BERT on SNLI dataset (Bowman et al., 2015).

4.4. Personalization

Personalization can be evaluated from two aspects: consistency and coverage.

Consistency reflects whether the generated response R is consistent with the given personal information P. *C-Score* (Madotto et al., 2019) finetuned BERT on DNLI and assigned 1, 0, and -1 for entailment, neutral, and contradiction, respectively. The final C-Score for an utterance is the sum over all personas: $\text{C-Score}(R) = \sum_i \text{NLI}(R, P_i)$. *P-Score* proposed in (Wu et al., 2021a) is actually the same as C-Score. *Consistency Score* (Tang et al., 2023) reduced the three classes in C-Score to binary classes, i.e., assigning 1 for entailment and neutral labels and 0 for the contradiction label.

Li et al. (2020) first showed that the perplexity of entailed dialogues would be lower than that of contradicted dialogues. Building on this, BoB (Song et al., 2021) reported the PPL of entailed and contradicted dialogues, as well as their subtraction, denoted as Delta Perplexity, to highlight model's capability of distinguishing between entailment and contradiction.

Coverage shows how much given personal information is reflected in the generated response. *Persona-F1* (Lian et al., 2019) is determined by the overlap of the set of non-stopword unigrams between R and P, and *Persona Coverage* (Song et al., 2019) quantifies the IDF-weighted word overlap between R and P. *Persona Distance* (Cho et al., 2022) is defined as the average word2vec cosine similarity between the keywords of R and P, where the keywords are decided by the word frequency after removing stopwords.

5. Discussion

5.1. Dataset

The persona-based datasets are expensive to collect and notable for its limited size. They are also considered artificial as the annotators are playing the given role and not act like themselves. The limited size also causes the out-of-domain persona problem discussed in Sec. 3.2.5, limiting the adaptation to the real-world scenario.

Crawling from social platforms addresses the issue of data size and provides original utterances from humans. However, the quality of these datasets is questionable. They originate from posts and comments on social media rather than from natural conversations, and thus the average turns per "dialogue" is relatively low. In addition, the extracted personas may not always align with the actual topic of discussion. Users may also post contradictory statements, making it challenging to deduce a consistent persona. And some situational or fleeting statements might be inadvertently extracted (Mazaré et al., 2018).

In addition, the diversity of the model are constraint, especially in domain and language variances, as discussed in Sec. 2.2.2. Multilingual dialogue data is pivotal in capturing the nuances of diverse cultural backgrounds. For example, a model trained solely on English or translated dialogue datasets may not effectively cater to Japanese contexts, given the stark contrast between English's low-context communication style and Japanese's high-context nature.

5.2. Methodology

Many works are based on the assumption that the query Q, response R, and persona P are mutually dependent, that is, assuming the three components are interconnected and that changes or variations in one might influence the other two. However, some of the works overlooked that not all R are personalized and not all personal information in the given P should be reflected.

Furthermore, research on personalized dialogue generation divides into two main streams, and each only models one single party of the dialogue speaker, as described in Sec 3.1. Future research might explore modeling both P_A, P_U for a more comprehensive personalized dialogue generation.

5.3. Evaluation

The primary evaluation metrics employed in recent papers may be insufficient. Predominantly, these are generation metrics, specifically similarity-based metrics, used for machine translation (MT) or summarization tasks. However, they have proven ineffective for assessing complex, open-ended tasks such as dialogue generation (Gehrmann et al., 2023; Yeh et al., 2021; Deriu et al., 2021; Liu et al., 2016). While advanced metrics that align more closely with human judgments have been proposed for dialogue evaluation, none have been applied in the examined studies.

Moreover, each study employs its own data preprocessing and might have different implementation on evaluation methods, complicating direct comparisons between different models. We advocate for a standardized approach to data preprocessing and evaluation. Such consistency will enable more precise comparisons among models and ensure that progress in the field is gauged against a uniform benchmark.

6. Conclusion

This review delves into personalized dialogue generation, covering Datasets, Methodologies, and Evaluation techniques. The cornerstone dataset in this field is PersonaChat. Modern datasets have expanded in size, added persona grounding, and now cover a broader range of domains, sessions, modalities, and languages. Persona is represented in descriptions, key-value attributes, or basic user IDs to trace past dialogues. The primary methodologies are: (1) imparting a persona to the agent and ensuring its consistency, and (2) pinpointing user personas or choosing the right one for context. The central challenges addressed include maintaining consistency and coherence, appropriately balancing persona with context, choosing pertinent personal details, modeling personas when not directly provided, and navigating limited data. For evaluations, factors like Fluency, Diversity, Coherence, and Personalization are paramount, with perplexity, BLEU, and Distinct-N being commonly used metrics. Personalization assessment primarily gauges persona consistency and coverage.

We conclude by discussing the limitations in the three dimensions. The primary issues with datasets include their limited size, concerns about quality, and insufficient diversity in both domains and languages. In terms of methodology, incorrect assumptions about the interdependencies among persona, query, and response can pose challenges. And we advocate for consistent assessment standards across various models and benchmarks for effective dialogue evaluation.

We compile information on the publication venues and repositories of related works, including those analyzed in this paper, in Table 5 in the Appendix.

7. Acknowledgements

This work was supported by the Institute of AI and Beyond of the University of Tokyo and MEXT KAKENHI Grant Number JP22H05015. Chen is also supported by JST SPRING, Grant Number JPMJSP2108.

8. Bibliographical References

Harsh Agrawal, Aditya Mishra, Manish Gupta, and Mausam. 2023. Multimodal persona based generation of comic dialogs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14150–14164, Toronto, Canada. Association for Computational Linguistics.

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards multimodal persona-grounded conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized quest and dialogue generation in role-playing games: A knowledge graph-and language model-based approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002, Dublin, Ireland. Association for Computational Linguistics.

Zhangming Chan, Juntao Li, Xiaopeng Yang, Xiuying Chen, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Modeling personalization in continuous space for response generation via augmented Wasserstein autoencoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1931–1940, Hong Kong, China. Association for Computational Linguistics.

Liang Chen, Hongru Wang, Yang Deng, Wai Chung Kwan, Zezhong Wang, and Kam-Fai Wong. 2023a. Towards robust personalized dialogue generation via order-insensitive representation regularization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7337–7345, Toronto, Canada. Association for Computational Linguistics.

Ruijun Chen, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2023b. Learning to memorize entailment and discourse relations for persona-consistent dialogues. *arXiv preprint arXiv:2301.04871*.

Wei Chen, Yeyun Gong, Song Wang, Bolun Yao, Weizhen Qi, Zhongyu Wei, Xiaowu Hu, Bartuer Zhou, Yi Mao, Weizhu Chen, Biao Cheng, and Nan Duan. 2022. DialogVED: A pre-trained latent variable encoder-decoder model for dialog response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4852–4864, Dublin, Ireland. Association for Computational Linguistics.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023c. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.

Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. 2022. A personalized dialogue generator with implicit user persona detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 367–377, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

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

Emily Dinan, 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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Xu Han, Bin Guo, Yoon Jung, Benjamin Yao, Yu Zhang, Xiaohu Liu, and Chenlei Guo. 2023. PersonaPKT: Building personalized dialogue agents via parameter-efficient knowledge transfer. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 264–273, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and H Tang. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12916–12923.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10803–10812.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

Minju Kim, Beong-woo Kwak, Youngwook Kim, Hong-in Lee, Seung-won Hwang, and Jinyoung Yeo. 2022. Dual task framework for improving persona-grounded dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10912–10920.

Deuksin Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim, and Eric Davis. 2023. What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 707–719, Toronto, Canada. Association for Computational Linguistics.

Jing Yang Lee, Kong Aik Lee, and Woon Seng Gan. 2022. Improving contextual coherence in variational personalized and empathetic dialogue agents. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7052–7056. IEEE.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Yangqiu Song, and Lixin Fan. 2022. You don't know my favorite color: Preventing dialogue representations from revealing speakers' private personas. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5858–5870, Seattle, United States. Association for Computational Linguistics.

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

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.

Jungwoo Lim, Myunghoon Kang, Yuna Hur, Seung Won Jeong, Jinsung Kim, Yoonna Jang, Dongyub Lee, Hyesung Ji, DongHoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. You truly understand what I need : Intellectual and friendly dialog agents grounding persona and knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1053–1066, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021. Xpersona: Evaluating multilingual personalized chatbot. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023a. RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. Improving personality consistency in conversation by persona extending. In *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*, CIKM '22, page 1350–1359, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. 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*, pages 555–564.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, and Harsh Jhamtani. 2021. Unsupervised enrichment of persona-grounded dialog with background stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 585–592, Online. Association for Computational Linguistics.

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.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. 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 1699–1713, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the SIGIR 2021*. ACM.

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 Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *arXiv preprint arXiv:2305.14930*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5190–5196. International Joint Conferences on Artificial Intelligence Organization.

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. *CoRR*, abs/2109.05217.

Bin Sun, Shaoxiong Feng, Yiwei Li, Jiamou Liu, and Kan Li. 2021. Generating relevant and coherent dialogue responses using self-separated conditional variational AutoEncoders. 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 5624–5637, Online. Association for Computational Linguistics.

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5456–5468, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Fucheng Wang, Yunfei Yin, Faliang Huang, and Kaigui Wu. 2023a. Please don't answer out of context: Personalized dialogue generation fusing persona and context. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023b. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

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

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Chen Henry Wu, Yinhe Zheng, Xiaoxi Mao, and Minlie Huang. 2021a. Transferable persona-grounded dialogues via grounded minimal edits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2368–2382, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021b. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics.

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.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023. Towards zero-shot persona dialogue generation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398, Toronto, Canada. Association for Computational Linguistics.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, Brussels, Belgium. Association for Computational Linguistics.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. 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 2718–2729, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Yan Zeng and Jian-Yun Nie. 2021. A simple and efficient multi-task learning approach for conditioned dialogue generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4927–4939, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. 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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and 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.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. 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*.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *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. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023a. SimOAP: Improve coherence and consistency in persona-based dialogue generation via over-sampling and post-evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9945–9959, Toronto, Canada. Association for Computational Linguistics.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2023b. Learning to predict persona information for dialogue personalization without explicit persona description. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2979–2991, Toronto, Canada. Association for Computational Linguistics.

Luyao Zhu, Wei Li, Rui Mao, Vlad Pandelea, and Erik Cambria. 2023. PAED: Zero-shot persona attribute extraction in dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9771–9787, Toronto, Canada. Association for Computational Linguistics.

9. Appendix

Table 5 presents a list of works related to personalized dialogue generation with their publication venues and repositories.

Dataset	Model	Title		Venue	Repo
	BoB	BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data	Song et al. (2021)	ACL	link
	PABST	Unsupervised Enrichment of Persona-grounded Dialog with Background Stories	Majumder et al. (2021)	ACL	link
PersonaMinEdit	GME	Transferable Persona-Grounded Dialogues via Grounded Minimal Edits	Wu et al. (2021a)	EMNLP	link
PER-CHAT		Personalized Response Generation via Generative Split Memory Network	Wu et al. (2021b)	NAACL	link
Persona Reddit		A Simple and Efficient Multi-Task Learning Approach for Conditioned Dialogue Generation	Zeng and Nie (2021)	NAACL	link
	DHAP	One Chatbot Per Person: Creating Personalized Chatbots based on Implicit User Profiles	Ma et al. (2021)	SIGIR	link
PchatbotW/L		Pchatbot: A Large-Scale Dataset for Personalized Chatbot	Qian et al. (2021)	SIGIR	link
XPersona		XPersona: Evaluating Multilingual Personalized Chatbot	Lin et al. (2021)	NLP4ConvAI	link
FoCus	FoCus	Call for Customized Conversation: Customized Conversation Grounding Persona and Knowledge	Jang et al. (2022)	AAAI	link
		Dual Task Framework for Improving Persona-grounded Dialogue Dataset	Kim et al. (2022)	AAAI	x
	D3	A Model-Agnostic Data Manipulation Method for Persona-based Dialogue Generation	Cao et al. (2022)	ACL	link
		DialogVED: A Pre-trained Latent Variable Encoder-Decoder Model for Dialog Response Generation	Chen et al. (2022)	ACL	link
MSC	MSC	Beyond Goldfish Memory : Long-Term Open-Domain Conversation	Xu et al. (2022a)	ACL	link
DuLeMon	DuLeMon	Long Time No See! Open-Domain Conversation with Long-Term Persona Memory	Xu et al. (2022b)	ACL-Findings	link
IT-ConvAI2	PS-Transformer	Improving Personality Consistency in Conversation by Persona Extending	Liu et al. (2022)	CIKM	link
	IUPD	A Personalized Dialogue Generator with Implicit User Persona Detection	Cho et al. (2022)	COLING	x
	INFO	You Truly Understand What I Need: Intellectual and Friendly Dialogue Agents grounding Knowledge and Persona	Lim et al. (2022)	EMNLP-Findings	x
		Keep Me Updated! Memory Management in Long-term Conversations	Bae et al. (2022)	EMNLP-Findings	link
	UA-CVAE	Improving Contextual Coherence in Variational Personalized and Empathetic Dialogue Agents	Lee et al. (2022)	ICASSP	x
	MSP	Less is More: Learning to Refine Dialogue History for Personalized Dialogue Generation	Zhong et al. (2022)	NAACL	link
		You Don't Know My Favorite Color: Preventing Dialogue Representations from Revealing Speakers' Private Personas	Li et al. (2022)	NAACL	link
	LMEDR	Learning to Memorize Entailment and Discourse Relations for Persona-Consistent Dialogues	Chen et al. (2023b)	AAAI	link
	PAA	Personalized Dialogue Generation with Persona-Adaptive Attention	Huang et al. (2023)	AAAI	link
	CLV	Enhancing Personalized Dialogue Generation with Contrastive Latent Variables: Combining Sparse and Dense Persona	Tang et al. (2023)	ACL	link
	SimOAP	SimOAP: Improve Coherence and Consistency in Persona-based Dialogue Generation via Over-sampling and Post-evaluation	Zhou et al. (2023a)	ACL	x
MPChat		MPCHAT- Towards Multimodal Persona-Grounded Conversation	Ahn et al. (2023)	ACL	link
LiveChat		LiveChat: A Large-Scale Personalized Dialogue Dataset Automatically Constructed from Live Streaming	Gao et al. (2023)	ACL	link
		RECAP: Retrieval-Enhanced Context-Aware Prefix Encoder for Personalized Dialogue Response Generation	Liu et al. (2023a)	ACL	link
		PAED: Zero-Shot Persona Attribute Extraction in Dialogues	Zhu et al. (2023)	ACL	link
		Multimodal Persona Based Generation of Comic Dialogs	Agrawal et al. (2023)	ACL	link
		Towards Zero-Shot Persona Dialogue Generation with In-Context Learning	Xu et al. (2023)	ACL-Findings	x
		PAL: Persona-Augmented Emotional Support Conversation Generation	Cheng et al. (2023)	ACL-Findings	link
		Learning to Predict Persona Information for Dialogue Personalization without Explicit Persona Description	Zhou et al. (2023b)	ACL-Findings	x
		Towards Robust Personalized Dialogue Generation via Order-Insensitive Representation Regularization	Chen et al. (2023a)	ACL-Findings	link
MSPD	WWH	WHAT, WHEN, and HOW to Ground: Designing User Persona-Aware Conversational Agents for Engaging Dialogue	Kwon et al. (2023)	ACL-Industry	x
		Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-based Approach	Ashby et al. (2023)	CHI	link
	PCF	Please don't answer out of context: Personalized Dialogue Generation Fusing Persona and Context	Wang et al. (2023a)	IJCNN	x

Table 5: List of works related to personalized dialogue generation. The first two columns are the names used in Sec 2 and 3.