Enhancing Multi-turn Dialogue Consistency with Localized-Generalized Persona Expansion

Xiaohua Wu[1,2], Yanbing Chen[1], Xiaohui Tao[3], Peipei Wang[4], and Lin Li[1(✉)]

[1] Wuhan University of Technology, Wuhan 430070, China
[2] Queensland University of Technology, Brisbane, QLD 4000, Australia
[3] University of Southern Queensland, Springfield, QLD 4300, Australia
[4] Qilu University of Technology, Jinan 250353, China
{xhwu,chenyb,cathylilin}@whut.edu.cn, xiaohui.tao@usq.edu.au, ppwang@sdas.org

Abstract. Open-domain dialogue systems, while exhibiting exceptional potential for wide-ranging application scenarios, suffer from hallucinations and weaknesses in maintaining long-standing personality consistency in multi-turn conversations. To mitigate this problem, the **Lo**calized and **G**eneralized **P**ersona Expansion (**LoG-P**) is proposed to enhance personalized dialogue system for consistent response generation. The LoG-P incorporates interaction learning and metamorphic relation construction as key components in localized and generalized persona expansions, aiming at enhancing persona understanding and consistency in response selection. Those make it a well-rounded solution for accurate personalized response selection in multi-turn dialogue generation. Since the retrieval-augmented generation (RAG) can enhance generative dialogue systems by retrieving relevant responses from available data stores, leading to higher retrieval accuracy, we first conduct the experiments on two assigned persona versions focusing on consistent personality reveal that the LoG-P significantly improves retrieval accuracy over the state-of-the-art methods.

Keywords: Personality Consistency · Persona Expansion · Metamorphic Relation · Retrieval-Augmented Generation

1 Introduction

Multi-turn dialogue systems have been applied in the field of human-computer interaction and intelligent services [12]. They not only advance dialogue modeling technology but also significantly change the nature of human-computer communication, bringing new application scenarios and possibilities to various domains. Open-domain dialogue systems exhibit exceptional potential for wide-ranging application scenarios, which have been a considerable topic. Given their response generation mechanism, this system is categorized into three categories: retrieval-based, generative, and ensemble system [34, 37, 41]. The early works mainly focused on the retrieval methods and aimed to select the appropriate sentence for responses [24, 39, 45]. However, the dialogue generation based on

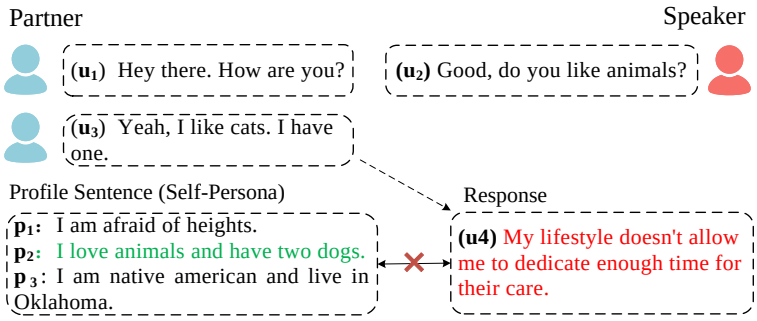

Fig. 1. The illustration of our motivation: In multi-turn dialogue, it is necessary to generate an appropriate response for the questions. However, the response generated by LLM contains some inconsistency with the persona of a speaker.

retrieval methods cannot deal with the situations where a new ground truth persona appears and methods of extracting attributes by non-sentence formats would miss semantic information implied within the sentence. In response, a few follow-up generation-based methodologies [29] including such as large language models (LLM) have been put forth, all capable of generating coherent responses, which exclusively rely on utterances from the dialogue history.

Maintaining persona consistency across multiple rounds is a key issue for multi-turn dialogue systems [40]. It is defined as ensuring historical context consistency, preserving the speaking style, and preserving personal information consistency. However, maintaining personality consistency is particularly challenging due to the diverse nature of training data comprising different speakers and the absence of explicit long-term memory for each speaker. Existing dialogue models exhibit weaknesses in maintaining long-standing personality consistency in multi-turn conversations. To address this, previous works typically introduce prior knowledge, such as persona [10, 13, 26, 28, 40] and background information [9, 17], into the dialogue system. For example, the self and partner persona sustain self-consistency [17] during a conversation. As shown in Fig. 1, the persona is described as a set of at least 5 profile sentences representing the personality of the speaker and partner [26]. This incorporation enhances accuracy and facilitates personalized dialogue systems [5, 11, 19].

Despite the notable successes of dialogue systems, they still face hurdles such as handling long-tail data, managing high training costs, and lacking personalized retrieval consistency. Specifically, dialogue systems may suffer from hallucination problems and sometimes tend to generate dialogue responses that appear plausible but are not consistent with speakers' persona [33] regardless of retrieval- or generation-based methods.

Taking the dialogue response generation topic as an example in Fig. 1, a model needs to generate the corresponding response for the current context of a multi-turn dialogue and the speaker's persona. However, the response may

not match the profile sentence and persona even though it is contextual coherence. Instead of the direct response generation, the novel retrieval-augmented generation (RAG) [1] has been proposed recently to mitigate this issue, which first retrieves relevant information from a large corpus of documents when the multi-turn dialogue system needs to generate text or answer questions. This retrieved information is then utilized to guide the text-generation process, thereby enhancing the quality and accuracy of the predictions.

In response, the **Lo**calized and **G**eneralized **P**ersona Expansion (**LoG-P**) framework is proposed to enhance the personalized dialogue generation using metamorphic relationship and interaction learning. First, the dialogue text is reconstructed by a designed template for generative models to model the inter-response (i.e., among pairwise responses) and intra-response (i.e., within one response) relationships. Then, localized persona expansion is introduced to softly expand the latest responses by considering the speaker's persona and conversation history based on an interaction learning mechanism. Moreover, the generalized persona expansion operates on the speaker's persona, generating expanded persona data using a type of metamorphic relation without additional manual costs. These localized-generalized persona expansions enhance persona understanding and consistency in dialogue systems. Finally, inspired by the RAG, we conducted experiments on a retrieval problem as an example based on two large-scale persona versions datasets that focus on maintaining responses with a consistent personality.

The main contribution of this work can be summarized as follows:

1) The **Lo**calized and **G**eneralized **P**ersona Expansion (**LoG-P**) framework is designed for facilitating the persona consistency in multi-turn dialogue systems. It can be easily transferred to dialogue retrieval and dialogue generation problems.
2) localized and generalized persona expansion by interaction learning and metamorphic relation is utilized to implicitly identify the semantic association between dialogue and persona, as well as capture the dialogue properties, which is beneficial to LoG-P for personalized response consistency.
3) Experimental results on dialogue retrieval problems based on two persona versions datasets as an example reveal that the proposed LoG-P achieves significant improvements in persona consistency, with a decrease in violation rates of up to 6.38%. Additionally, LoG-P improves retrieval accuracy by up to 13% in Hits@1 and 8.9% in MRR compared to state-of-the-art methods.

2 Related Work

With the development of natural language processing, dialogue retrieval, and generation are two mainstream methods in multi-turn dialogue systems.

2.1 Multi-turn Dialogue Retrieval

Dialogue retrieval models mainly encompass three paradigms: non-pretraining, pretraining and fine-tuning, and prompt learning.

Non-Pretraining Earlier multi-turn dialogue retrieval works focused on word or sentence matching, by deep learning methods, such as LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Networks) for response selection [24], and Deep Attention Network [45] for dialog information matching. However, these works cannot extract the interaction information in pairwise conversation thereby they have a limited performance in long-term dialogue.

Pretraining and Fine-tuning Previous works [9, 10, 39] show significant performance by pre-training and fine-tuning. Inspired them, FT-PC [26], TransferTransfo [38], and BERT-CRA [11] are proposed for better dialogue retrieval improvements. Recently, various fusion strategies, such as the BERT-Fusion strategy, have been effective for multi-turn dialogue retrieval [5]. However, the previous works do not consider the persona consistency in the multi-turn dialogue even though they reflect the characteristics.

Prompt Learning Since most pre-trained language models can be fine-tuned for downstream tasks with different goals, there is a gap between pre-training and fine-tuning. Therefore, recent work has employed prompt learning [23], a lightweight alternative to fine-tuning. By using prompts to fine-tune the pre-trained language model, the rich knowledge distributed in the pre-trained language model can be further stimulated to better serve downstream tasks, such as memes category detection [4] and P5 [16].

2.2 Dialogue Generation

Early dialogue response generation systems suffer from problems like dull and generic responses, and a lack of encoding context [18]. With the development of generation-based methods, recent works utilized MLM, such as DialogLM-MLM derived from the dialogue model DialogLM [43], BERT-MLM [7], and NSP problem such as BERT-NSP [32] and DialogLM-NSP, both converted into NSP template. These works indicated that the generation-based method is a novel and advanced method for dialogue generation. The latest work involves a chain of thoughts (CoT) that is employed for personalized response generation with reasonable steps [21, 31]. However, even with the powerful performance of generation-based methods for developing dialogue systems, it lacks a consistent persona in multi-turn conversations.

2.3 Persona Enhanced Dialogue System

The concept of persona is first introduced to ensure consistent characterization and clear history recall for generating rational responses [40, 44]. Inspired by these works, it is introduced as a means to maintain speaker information and enhance dialog coherence, e.g., [19]. While [40] built a personalized dialogue system and constructed a baseline model containing both the speaker and partner personas. The incorporation of conversational partner personas into response selection is explored by [11], and the proposed BERT-CRA also achieves high model performance.

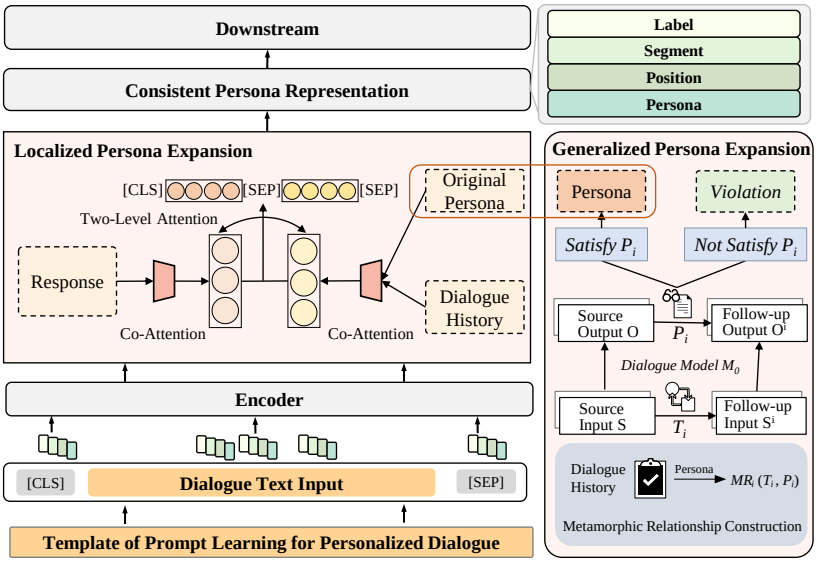

Fig. 2. The framework of the proposed framework LoG-P.

In summary, many works with retrieval-based [5, 7, 11, 20, 38, 43] and generation-based methods are introduced to mitigate the hallucination [33] and plausible dialogue response problems that are not consistent with the speaker's persona [9, 10, 16, 40, 44]. The current persona-enhanced methods still have challenges of handling long-tail data, managing high training costs, and lacking personalized generation consistency. Therefore, we utilize the Localized and Generalized Persona Expansion framework (LoG-P) with interaction and persona expansion components to enhance personalized dialogue systems.

3 LoG-P Framework

3.1 Problem Definition

It is given a dataset D consisting of n conversation tuples in the format of (c, p, r, y). $c = \{u_1, u_2, \ldots, u_{n_c}\}$ represents the dialogue history with n_c sentences, $p = \{p_1, p_2, \ldots, p_{n_p}\}$ is the n_p personas of the speaker, and r is the response candidate for c. The label $y \in \{0, 1\}$ indicates whether r is the appropriate response for (c, p), where $y = 1$ means it is appropriate, while $y = 0$ means it is not. The objective is to learn a matching function g from D such that, given any tuple (c, p, r), $g(c, p, r) = 0$ denotes the answer generated matches the question, and $g(c, p, r) = 0$ is not.

3.2 Methodology

To improve the persona consistency of dialogue systems, a **Lo**calized and **G**eneralized **P**ersona Expansion (**LoG-P**) framework is proposed, which includes two modules as shown in Fig. 2. Inspired by the significant performance of retrieval-augmented generative models [16], the personalized dialogue retrieval problem is first reconstructed into the ensemble of two pre-training tasks to align the advanced generative learning methods. As described in Fig. 3, dialogue retrieval is restructured using MLM and NSP by determining whether the answer matches the question. There are complex intra-sentence and inter-sentence relationships in multi-turn conversations. Interaction learning is then employed to expand the localized persona, while MT-based persona expansions are used to broaden the generalized persona. The extracted relations are represented as $h_{dialogue}$ and $h_{[CLS]}$ for response selection through the degree of matching.

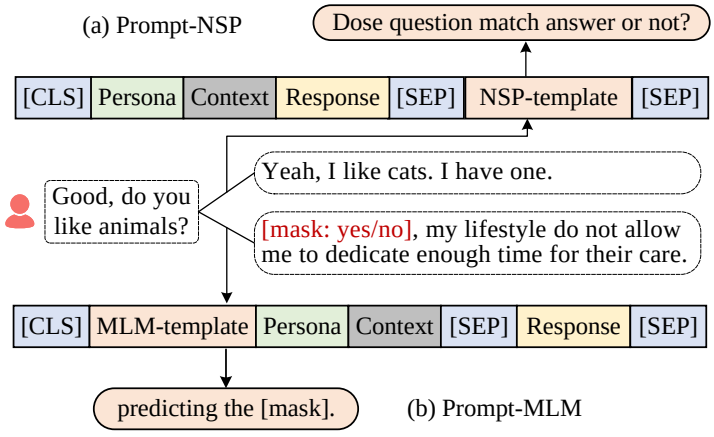

Fig. 3. The prompt templates of NSP and MLM.

Reconstructing Retrieval into Generative Learning The generative learning training paradigm in previous works shows a significantly effective performance in retrieval problems. Therefore, the LoG-P reconstructs the personalized dialogue retrieval problem into the ensemble of two pre-training paradigms, i.e., the masked language modeling (MLM) and the next sentence prediction (NSP), and employs prompt templates to model complex intra- and inter-sentence relationships in a multi-turn conversation from the task reconstruction perspective. The MLM works by masking a portion of the input tokens in a sentence at random and then asking the model to predict the masked tokens, e.g., the mask is yes or no. The NSP is a binary classification problem to predict whether the alternative answer originally matched the question. The prompt templates are as Fig. 3 and also shown in the left part of Fig. 2.

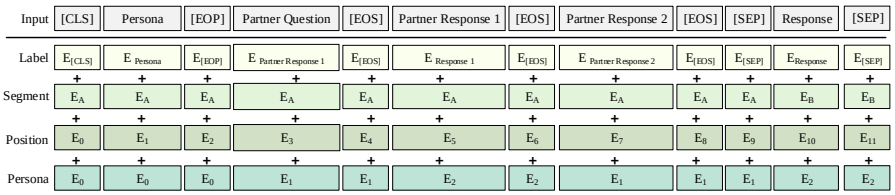

Fig. 4. The input fused representation.

Localized Persona Expansion To model the response interaction, interaction learning is introduced to implicitly identify the semantic association between dialog context and personas, as shown in the middle part of Fig. 2. Localized persona is defined as the deeper interaction between dialog history, speaker persona, and response in the current session of multi-turn dialogue.

1) Fused Speaker's Persona Representation Dialog context is further subdivided into segments from different speakers and their persona. Inspired by [8], the fused speaker response and its persona representation are used. Specifically, a speaker persona presentation E_0, speaker presentation E_1, and partner response presentation E_2 are introduced based on the conversation order, as shown in Fig. 4. The [EOP] and [EOS] are segmentation labels for marking the end of the persona and sequence, respectively. Therefore, the speaker's dialogue context and persona representations can be defined as Equation 1.

$$E = \sum_{1}^{n_p+2} E_0 + \sum_{1}^{n_{o_1}+1} E_1 + \sum_{1}^{n_{s_1}+1} E_2 + \cdots + \sum_{1}^{n_{o_i}+2} E_1 + \sum_{1}^{n_r+1} E_2, \tag{1}$$

where n_p is the number of words in all sentences of the persona; n_{o_i} is the number of words in one dialogue response of the partner, and i is the number of responses; n_{s_i} is the number of words in one response; n_r is the number of words in the set of all historical responses.

2) Dialog History Interaction Modeling Since the response to the dialog history of this round may appear in later rounds during the conversation, we propose a multi-level attention module to better capture the dialog properties. Specifically, this multi-level attention includes a co-attention layer and a two-level attention layer.

Co-Attention The speaker persona and dialogue history representation are fused as dialogue context representation vector $E_1^c \in \mathcal{R}^{n\times h}$. Then, the dialogue context representation can be defined as $E_1^r \in \mathcal{R}^{m\times h}$. Last, the final response fused with the last response is obtained by Equation 2 to 5.

$$C = (A^C)^T \left[E_1^r; \left((E_1^C)^T A^r\right)^T\right] \in \mathcal{R}^{n\times 2h}, \tag{2}$$

$$r = (A^r)^T \left[E_1^C; \left((E_1^r)^T A^C \right)^T \right] \in \mathcal{R}^{m \times 2h}, \tag{3}$$

$$C^r = layernorm(E_1^C + CW^C) \in \mathcal{R}^{n \times h}, \tag{4}$$

$$r^C = layernorm(E_1^r + rW^C) \in \mathcal{R}^{m \times h}, \tag{5}$$

where the $C \in \mathcal{R}^{n \times 2h}$ is the dialogue context representation, and r is the response representation, W^C and W^r are hyper-parameters; A^C is the weight of the word pair in the dialogue context, and A^r is the weights of the word pairs in the response.

Two-level Attention To model the multi-turn response relations, we also design a sliding window in the two-level attention. The sliding window w_1 of the first level is set to the length of two dialogue rounds. The second level of attention focuses on two dialogue information responses. Therefore, the adjacent set can be defined as $N(i, w_1) = \{i - w_1, \ldots, i, \ldots, i + w_1\}$. Then, the prompt, persona, and last sentence representation in dialogue history are added to a generalized set $G = \{q_1, q_2, \ldots, q_l\}$, so that all the words in the sliding window can simultaneously pay attention to the generalized set and themselves in their window. The word-sensing region not in the generalized set is $N(i, w_1) \cup G$. Therefore, the $N_G(i, w)$ is defined as Equation 6.

$$N_G(i, w) = \begin{cases} N(i, w) \cup G, & i \notin G \\ [1, \ldots, n], & others \end{cases}, \tag{6}$$

where the n is the sequence length. Then the first level of attention of i_{th} word can be defined as Equation 7.

$$y_i^T = softmax(\alpha q_i^T K_{N_g(i,w_1)}) V_{N_g(i,w_1)}^T, \tag{7}$$

where the q_i, $K_{N_g(i,w_1)}$, and $V_{N_g(i,w_1)}$ denote the query, key, and value of i_{th} word, respectively. Then, the second level of attention of i_{th} word is defined as Equation 8, which focuses between two sentences.

$$z_i^T = softmax(\alpha \hat{q}_i^T \hat{K}_{N_g(i,w_2)}) \hat{V}_{N_g(i,w_2)}^T \tag{8}$$

The $\hat{q}_i$, $\hat{K}_{N_g(i,w_2)}$, and $\hat{V}_{N_g(i,w_2)}$ represent the new query, key, and value generated from the output of the first level of attention $Y = \{y_1, y_2, \ldots, y_n\}$ within a narrower window. Therefore, the final attention of two levels is $h_{dialogue} = y_i^T + z_i^T$ by residual connection.

Generalized Persona Expansion The localized persona expansion by interaction learning focuses on the semantic attention between dialog context and personas. Moreover, the profile and persona can be described by many linguistic sentences that have a high semantic similarity. In addition to traditional persona expansion methods, such as similarity computation, a novel method called metamorphic relation demonstrates significant expansion performance without additional manual costs in various NLP problems [27]. However, the key lies

in discovering and designing appropriate metamorphic relations (MRs). In response, we introduce three metamorphic relations of self-persona consistency, which can be used for generalized personas expansion. Taking the synonyms as an example, the metamorphic relation and metamorphic tests (MTs) are designed below.

MR: Self-Persona Consistency.

Previous research [2, 25, 30] has employed the technique of replacing keywords or adjectives in a sentence with their respective synonyms to assess the model's comprehension within the sentence. This approach aims to evaluate the model's understanding rather than solely focusing on improving accuracy by selecting words that are strictly similar in literal terms. Considering the speaker's persona in dialogues, the overall semantics of persona is more important than individual words. Leveraging the principle that synonymous sentences preserve the original meaning, this work establishes a transformation relationship ensuring that modified inputs do not alter the prediction results of the original dialogue model, in the right part of Fig. 2.

MT: Synonymous Sentences Test. MR is applied where synonymous sentences replaced the *original* persona. The aim was to assess the model's ability to recognize synonymous sentences and adhere to persona-centric metamorphic relations, specifically in-variance.

Example 1:
Persona of source input: "I am afraid of heights."
Persona of follow-up input: "I have acrophobia."

Example 1 is an illustration of this metamorphic relation. Since the two personas have identical meanings if the model's output for these inputs is not the same, which would be considered an error, those metamorphic test results can be used as training data directly. The applied self-persona consistency shows its effectiveness in improving persona consistency and accuracy in multi-turn dialogue retrieval.

4 Experiments

4.1 Experimental Setup

Dataset Various types of datasets exist in the domain of personalized multi-turn dialogue systems, such as Persona-Chat dataset [40], DailyDialog [22], Personal-Dialog [42], personalized empathy datasets PEC [44], and FoCus incorporating personalized information with background knowledge [15]. For persona consistency problems, the two persona versions described from different perspectives with persona consistency are usually utilized [40]. It includes the original and revised versions, which consider the attribution of maintaining consistent personality in the responses. The two persona versions consist of 65,719 context-response

pairs for the training set, 7,801 pairs for the validation set, and 7,512 pairs for the test set. They include correct responses from real humans, while incorrect responses are randomly sampled. To increase the challenge of the task, measures are taken to ensure that there is no overlap of contexts and roles between the training, validation, and test sets. This approach guarantees the evaluation of models' generalization capabilities across diverse scenarios. Several studies [5, 9–11, 38] have also explored both non-pre-trained and pre-trained approaches on the two versions to enhance personalized dialogue retrieval and maintain consistent personality in the responses.

Metrics Hits@1 and Mean Reciprocal Rank (MRR) [16, 35] are popular metrics in multi-round personalized dialogue retrieval to assess the accuracy of relevant conversation models. They are usually employed in reference-based evaluation.

The experiments are implemented in Python 3.7 and TensorFlow 1.3.0 on a commodity server equipped with an NVIDIA TITAN Xp and Intel(R) Xeon(R) E5-2650 CPU.

4.2 Baselines

As discussed in Sec. 2, there are three main paradigms in multi-turn dialogue retrieval. Therefore, we investigate three generative paradigms of eight baselines and conduct evaluations on the persona consistency performance.

Non-pretraining The non-pretraining approach involves training the personalized dialogue retrieval model using dataset D, such as the popular *DIM* [10] and *FIRE* [9].

Pretraining and Fine-tuning Pretraining methods utilize a pre-trained model trained on a large-scale corpus. Fine-tuning initializes the personalized dialogue retrieval model with pre-trained and updates its weights using dataset D and matching function g to adapt it to the task. Therefore, the *FT-PC* [26], *TransferTransfo* [38], *BERT-CRA* [11], and BERT-Fusion-Strategies [5] are baselines of this work. It includes BERT-NoFusion with no fusion strategies and BERT-Fusion-All with interactive fusion strategies that can capture the interactions among the persona, emotion, and contexts of the dialogue. The multi-grained conversational graph network (MCGN) [35] is proposed to consider multiple levels of abstraction from dialogue histories and semantic dependencies within multi-turn dialogues for addressing.

Prompt Learning The prompt learning includes a prompt p within dataset D. In this paper, we will primarily utilize the current state-of-the-art model **P5** [16], which is the state-of-the-art methods and is along with the following four reconfigured prompt learning models: *BERT-MLM* [7], transformed from BERT_CRA into the MLM format; DialogLM-MLM, derived from the dialogue model *DialogLM* [43] and converted into the MLM format; as well as *BERT-NSP* [32] and *DialogLM-NSP*, both converted into the NSP template.

Table 1. Results of LoG-P (Ours) and baselines on two assigned personas and two different scenarios (%)

Paradigm	Model	Original Version				Revised Version			
		Self-Persona		Partner Persona		Self-Persona		Partner Persona	
		Hits@1	MRR	Hits@1	MRR	Hits@1	MRR	Hits@1	MRR
Non-Pretraining	DIM	78.8	86.7	64.0	76.1	70.7	81.2	63.9	76.0
	FIRE	81.6	-	-	-	74.8	-	-	-
Pretraining & Fine-tuning	FT-PC	-	-	-	-	60.7	-	-	-
	TransferTransfo	80.7	-	-	-	-	-	-	-
	BERT-CRA	84.3	90.3	71.2	80.9	79.4	86.9	71.8	81.5
	BERT-NoFusion	84.4	90.7	71.2	81.1	79.4	87.6	71.4	81.5
	BERT-Fusion-All	86.6	91.6	72.6	81.9	81.3	88.6	72.4	81.9
	MCGN	76.9	85.8	-	-	76.9	85.8	-	-
Generative Learning	P5	87.4	-	-	-	82.8	-	-	-
	LoG-P (Ours)	**87.7**	**92.6**	**75.2**	**84.5**	**83.7**	**90.1**	**75.5**	**84.5**

4.3 Overall Performance

The results of LoG-P and eight baselines are shown in Table 1 on original and revised persona versions.

In terms of Hits@1 and MRR, the proposed LoG-P outperforms all baseline models with up to 87.7% (Hits@1) and 92.6% (MRR) based on two persona versions, which verifies the effectiveness of persona expansion under a prompt learning-based task framework. Specifically, compared with pre-training and fine-tuning paradigms, using prompt learning to reconstruct the personalized conversation retrieval obtains the Hits@1 gains with up to 7.0% and the MRR gains with up to 4.4%. Note that the results of MCGN are from the original experiments, which is the overall performance of original and revised persona versions. All these results indicate that the prompt learning can extract the external knowledge learned in pre-training, and also help to improve the context understanding of complex relationships in multiple rounds of dialogue. By adding interaction learning and representation of the speaker persona, the retrieval ability of the personalized dialogue system is improved more effectively.

Even though P5 uses the speaker persona most similar to the response as a prompt, which helps the model find the accurate response, the proposed LoG-P is still slightly higher in Hits@1 by 0.3%-0.9%. Experimental results show that in personalized dialogue, the persona, dialogue history, and prompt learning work together to obtain more advantages.

4.4 Effectiveness of Persona Expansion

To study the localized and generalized persona expansion performance, we conduct an ablation study from two modules including interaction learning and MR. As shown in Table 2, without MR, the performance of LoG-P-MLM (BERT), LoG-P-NSP (BERT), LoG-P-MLM (DialogLM), and LoG-P-NSP (DialogLM)

Table 2. Ablation study of persona expansions (%).

Model	Self-Persona (Original)	
	Hits@1	MRR
LoG-P-MLM (BERT)	85.3	91.1
-MR (generalized)	84.2	90.2
-Multi-Level Attention (localized)	83.6	89.9
-Second Level Attention (localized)	84.1	90.2
LoG-P-NSP (BERT)	85.9	91.5
-MR (generalized)	84.3	90.3
-Multi-Level Attention (localized)	84.4	90.2
-Second Level Attention (localized)	84.8	90.7
LoG-P-MLM (DialogLM)	85.4	91.1
-MR (generalized)	85.1	90.8
-Multi-Level Attention (localized)	65.2	78.4
-Second Level Attention (localized)	72.1	85.2
LoG-P-NSP (DialogLM)	86.6	91.8
-MR (generalized)	85.7	91.2
-Multi-Level Attention (localized)	84.7	90.2
-Second Level Attention (localized)	85.1	90.7

have declined with up to 1.6% (at Hits@1) and 1.2% (at MRR), which demonstrates the effectiveness of MR. DialogLM's pre-training includes multi-party dialogue mask identification tasks, that is, the original model has a certain knowledge reserve for the potential characteristics of speakers. The two BERT-related models have a relatively large decrease, which shows the effectiveness of this module in improving the accuracy of personalized conversation retrieval tasks.

Similarly, comparing the four models with and without localized persona expansion, i.e., Multi-Level attention and second-level attention, their performance has declined by up to 1.9% at Hist@1 and 12.7% at MRR. Regarding localized persona expansion by interaction learning, the models' accuracy is significantly reduced after the ablation of this module, which may be because the MASK language model mentioned above uses both [CLS] and [MASK] tags as matching features. The results show the different importance of each kind for LoG-P performance and its effectiveness for personalized conversation retrieval tasks.

4.5 Persona Consistency Analysis

To explore the retrieval persona consistency of the proposed LoG-P, we first conduct a semantic expansion of each persona description. The previous popular document enrichment strategies include n-grams and semantic similarity [14], as well as semantic relationships based on word co-occurrence [6] and contextual information [36]. Inspired by these works, we compare the MR and synonymous sentence replacement to enrich the semantics of the persona description. Synonymous sentences have replaced the *original* persona. The performance of persona-centric multi-turn dialogue models is investigated from two aspects.

One is whether dialogue models can still give expected results under the MR in Sec. 3.2 after applying the metamorphic test. The other is whether dialogue models can be further improved with MR-based data augmentation.

Table 3. The performance of LoG-P enhanced models evaluated by violation rate (%). Lower is Better ↓.

Dialogue Model	Non-MR	MR
BERT-CRA	20.42	-
P5	14.39	-
LoG-P-MLM (BERT)	19.08	**12.94**
LoG-P-NSP (BERT)	18.64	**12.43**
LoG-P-MLM (DialogLM)	17.94	**11.21**
LoG-P-NSP (DialogLM)	17.29	**10.01**
LoG-P (ours)	15.70	**9.32**

The violation rate [3] measures the extent to which the tested model fails to meet the metamorphic relations and is applied here to evaluate Robustness-A. As shown in Table 3, the LoG-P effectively improves the robustness of dialogue retrieval by significantly decreasing the violation rate compared to BERT-CRA, P5, and four basic prompt learning models such as Prompt-MLM (BERT), and the violation rate is as low as 9.32%. Considering Robustness-B, the generalized persona expansion by MR brings perturbations to dialogue models. The similarity between speaker persona and candidate responses is traditional to expand persona with highly similar responses. Table 4 reports the retrieval performance of the two expansions. The generalized expansion by MR obtains the Hits@1 with up to 88.7% and the MRR with up to 93.2%. Therefore, generalized expansion by MR is better than similarity-based in terms of Hits@1 and MRR, which advocates for replacing the original persona with synonymous sentences.

Table 4. Results of MR *VS.* Similarity in generalized expansion based on partner persona of original version

Expansion	Model	Original		Revised	
		Hits@1	MRR	Hits@1	MRR
Similarity	BERT-MLM	86.1	91.5	80.2	87.7
	BERT-NSP	85.3	91.0	79.8	87.4
	DialogLM-MLM	85.7	91.2	79.5	87.1
	DialogLM-NSP	85.8	91.4	81.3	88.4
	LoG-P (ours)	**87.5**	**92.6**	**83.0**	**89.6**
MR	BERT-MLM	85.7	91.2	80.2	87.9
	BERT-NSP	85.7	91.4	80.0	87.6
	DialogLM-MLM	86.2	91.5	80.9	88.3
	DialogLM-NSP	87.3	92.3	82.4	89.0
	LoG-P (ours)	**87.7**	**92.6**	**83.7**	**90.1**

Comparing the LoG-P and baselines, the results show that the LoG-P effectively improves the robustness of the models, which indicates that persona expansion by MR is effective.

5 Conclusion and Future Works

The localized-generalized persona-augmented LoG-P by metamorphosis relationship is proposed to improve the robustness and accuracy of personalized dialogue retrieval. The effectiveness of the proposed LoG-P is verified through comparative experiments of accuracy performance and persona consistency improvements in two assigned persona versions. In the future, persona expansion will be designed and conducted on more data focusing on consistent personality traits.

Acknowledgments. This work is partially supported by NSFC, China (No. 62276196), Guangxi Science and Technology Major Program (No.AA24206067), and the China Scholarship Council program (No.202306950097).

References

1. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking large language models in retrieval-augmented generation. In: AAAI, February 20-27, 2024, Vancouver, Canada. pp. 17754–17762 (2024)
2. Chen, S., Jin, S., Xie, X.: Validation on machine reading comprehension software without annotated labels: a property-based method. In: ESEC/FSE, August 23-28, 2021. pp. 590–602 (2021)
3. Chen, S., Jin, S., Xie, X.: Validation on machine reading comprehension software without annotated labels: a property-based method. In: ESEC/FSE Athens, Greece, August 23-28, 2021. pp. 590–602. ACM (2021)
4. Cui, J., Li, L., Tao, X.: Be-or-not prompt enhanced hard negatives generating for memes category detection. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 174–179 (2023)
5. Das, S., Saha, S., Srihari, R.K.: Using multi-encoder fusion strategies to improve personalized response selection. In: COLING, Gyeongju, Republic of Korea, October 12-17. pp. 532–541 (2022)
6. Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M.A., Jr., W.M.: Word co-occurrence features for text classification. Inf. Syst. **36**(5), 843–858 (2011)
7. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. In: ACL/IJCNLP, Virtual Event, August 1-6, 2021. pp. 3816–3830 (2021)
8. Gu, J., Li, T., Liu, Q., Ling, Z., Su, Z., Wei, S., Zhu, X.: Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In: CIKM, Virtual Event, Ireland, October 19-23, 2020. pp. 2041–2044 (2020)
9. Gu, J., Ling, Z., Liu, Q., Chen, Z., Zhu, X.: Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In: EMNLP, Online Event, 16-20 November 2020. pp. 1412–1422. Findings of ACL (2020)
10. Gu, J., Ling, Z., Zhu, X., Liu, Q.: Dually interactive matching network for personalized response selection in retrieval-based chatbots. In: EMNLP-IJCNLP, Hong Kong, China, November 3-7, 2019. pp. 1845–1854 (2019)

11. Gu, J., Liu, H., Ling, Z., Liu, Q., Chen, Z., Zhu, X.: Partner matters! an empirical study on fusing personas for personalized response selection in retrieval-based chatbots. In: SIGIR, Virtual Event, Canada, July 11-15, 2021. pp. 565–574 (2021)
12. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: NIPS 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 2042–2050 (2014)
13. Hu, Z., Wang, L., Chen, Y., Liu, Y., Li, R., Zhao, M., Lu, X., Jiang, Z.: Dynamically retrieving knowledge via query generation for informative dialogue generation. Neurocomputing **569**, 127036 (2024)
14. Huang, Q., Chen, Z., Lu, Z., Ye, Y.: Analysis of bag-of-n-grams representation's properties based on textual reconstruction. CoRR **abs/1809.06502** (2018)
15. Jang, Y., Lim, J., Hur, Y., Oh, D., et al.: Call for customized conversation: Customized conversation grounding persona and knowledge. In: AAAI, Virtual Event, February 22 - March 1, 2022. pp. 10803–10812 (Jun 2022)
16. Lee, J., Oh, M., Lee, D.: P5: plug-and-play persona prompting for personalized response selection. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP, Singapore, December 6-10, 2023. pp. 16571–16582 (2023)
17. Lee, K., Lee, C., Kim, D., Lee, K.H.: Dialogue act-based partner persona extraction for consistent personalized response generation. Expert Systems with Applications **254**, 124380 (2024)
18. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. In: Proc. of NAACL-HLT (2016)
19. Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., Dolan, W.B.: A persona-based neural conversation model. In: ACL, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers (2016)
20. Li, L., Liu, D., Zhao, L., Zhang, J., Liu, J.: Evidence mining for interpretable charge prediction via prompt learning. IEEE Transactions on Computational Social Systems **11**(4), 4556–4566 (2024)
21. Li, Y., Feng, S., Wang, D., Zhang, Y., Yang, X.: Paper: A persona-aware chain-of-thought learning framework for personalized dialogue response generation. In: NLPCC, Hangzhou, China, November 1–3, 2024. p. 215–227 (2024)
22. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manually labelled multi-turn dialogue dataset. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing. pp. 986–995 (2017)
23. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 195:1–195:35 (2023)
24. Lowe, R., Pow, N., Serban, I., Pineau, J.: The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In: SIGDIAL, 2-4 September 2015, Prague, Czech Republic. pp. 285–294 (2015)
25. Malfa, E.L., Wu, M., Laurenti, L., Wang, B., Hartshorn, A., Kwiatkowska, M.: Assessing robustness of text classification through maximal safe radius computation. In: EMNLP, Online Event, 16-20 November 2020. pp. 2949–2968 (2020)
26. Mazaré, P., Humeau, S., Raison, M., Bordes, A.: Training millions of personalized dialogue agents. In: EMNLP, Brussels, Belgium, October 31 - November 4, 2018. pp. 2775–2779 (2018)
27. Nolasco, A., Molina, F., Degiovanni, R., Gorla, A., et al.: Abstraction-aware inference of metamorphic relations. Proc. ACM Softw. Eng. **1**(FSE), 450–472 (2024)

28. Peng, S., Qu, D., Zhang, W., Zhang, H., Li, S., Xu, M.: Easy and effective! data augmentation for knowledge-aware dialogue generation via multi-perspective sentences interaction. Neurocomputing **614**, 128724 (2025)
29. Ren, D., Cai, Y., Lei, X., Xu, J., Li, Q., Leung, H.f.: A multi-encoder neural conversation model. Neurocomputing **358**, 344–354 (2019)
30. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: Behavioral testing of NLP models with checklist (extended abstract). In: IJCAI, Virtual Event / Montreal, Canada, 19-27 August 2021. pp. 4824–4828 (2021)
31. Sun, X., Tang, X., Ali, A.E., et al.: Chain-of-strategy planning with llms: Aligning the generation of psychotherapy dialogue with strategy in motivational interviewing. CoRR **abs/2408.06527** (2024)
32. Sun, Y., Zheng, Y., Hao, C., Qiu, H.: NSP-BERT: A prompt-based few-shot learner through an original pre-training task - - next sentence prediction. In: COLING, Gyeongju, Republic of Korea, October 12-17, 2022. pp. 3233–3250 (2022)
33. Tonmoy, S.M.T.I., Zaman, S.M.M., et al.: A comprehensive survey of hallucination mitigation techniques in large language models. CoRR **abs/2401.01313** (2024)
34. Tu, Q., Tao, C., Yan, R.: Multi-grained conversational graph network for retrieval-based dialogue systems. In: LREC/COLING, 20-25 May, 2024, Torino, Italy. pp. 11756–11765 (2024)
35. Tu, Q., Tao, C., Yan, R.: Multi-grained conversational graph network for retrieval-based dialogue systems. In: LREC-COLING. pp. 11756–11765 (2024)
36. Viegas, F., Cunha, W., Gomes, C., et al.: Cluhtm - semantic hierarchical topic modeling based on cluwords. In: ACL, July 5-10. pp. 8138–8150 (2020)
37. Wang, L., Zhao, M., Ji, H., et al.: Dialogue summarization enhanced response generation for multi-domain task-oriented dialogue systems. Inf. Process. Manag. **61**(2), 103668 (2024)
38. Wolf, T., Sanh, V., Chaumond, J., Delangue, C.: Transfertransfo: A transfer learning approach for neural network based conversational agents. CoRR **abs/1901.08149** (2019)
39. Yuan, C., Zhou, W., Li, M., Lv, S., Zhu, F., Han, J., Hu, S.: Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In: EMNLP-IJCNLP, Hong Kong, China, November 3-7, 2019. pp. 111–120 (2019)
40. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? In: ACL, Melbourne, Australia, July 15-20, 2018. pp. 2204–2213 (2018)
41. Zhao, M., Wang, L., Jiang, Z., et al: From easy to hard: Improving personalized response generation of task-oriented dialogue systems by leveraging capacity in open-domain dialogues. Knowl. Based Syst. **295**, 111843 (2024)
42. Zheng, Y., Chen, G., Huang, M., Liu, S., Zhu, X.: Personalized dialogue generation with diversified traits. CoRR **abs/1901.09672** (2019)
43. Zhong, M., Liu, Y., Xu, Y., Zhu, C., Zeng, M.: Dialoglm: Pre-trained model for long dialogue understanding and summarization. In: AAAI, Virtual Event, February 22 - March 1, 2022. pp. 11765–11773 (2022)
44. Zhong, P., Zhang, C., Wang, H., Liu, Y., Miao, C.: Towards persona-based empathetic conversational models. In: EMNLP, Online, November 16-20, 2020. pp. 6556–6566 (2020)
45. Zhou, X., Li, L., Dong, D., Liu, Y., Chen, Y., Zhao, W.X., Yu, D., Wu, H.: Multi-turn response selection for chatbots with deep attention matching network. In: ACL 2018, Melbourne, Australia, July 15-20. pp. 1118–1127 (2018)