"In Dialogues We Learn": Towards Personalized Dialogue Without Pre-defined Profiles through In-Dialogue Learning

Anonymous ACL submission

Abstract

Personalized dialogue systems have gained significant attention in recent years for their ability to generate responses in alignment with different personas. However, most existing approaches rely on pre-defined personal profiles, which are not only time-consuming and labor-intensive to create but also lack flexibility. We propose **I**n-**D**ialogue **L**earning (IDL), a fine-tuning framework that enhances the ability of pre-trained large language models to leverage dialogue history to characterize persona for completing personalized dialogue generation tasks without pre-defined profiles. Our experiments on three datasets demonstrate that IDL brings substantial improvements, with BLEU and ROUGE scores increasing by up to 200% and 247%, respectively. Additionally, the results of human evaluations further validate the efficacy of our proposed method.

1 Introduction

Recently, there has been growing interest in building personalized dialogue systems (Tang et al., 2023; Chen et al., 2023c; Huang et al., 2023; Chen et al., 2023a). Such systems are often adept at incorporating special personal characteristics into responses. Consequently, personalized dialogue systems offer enhanced flexibility, enabling them adapt more effectively to a wide range of conversational scenarios, such as role-playing games (Park et al., 2023).

To achieve personalized dialogues, a common practice is to condition a dialogue model on a profile that explicitly depicts the personality traits one aims to portray with a textual description (Song et al., 2021; Liu et al., 2022; Chen et al., 2023b). While a profile can effectively delineate the desired personality traits, creating an accurate profile is nevertheless time-consuming and arduous.

In this work, we attempt to develop a model capable of performing personalized dialogue generation without the need of profiles designed in advance.

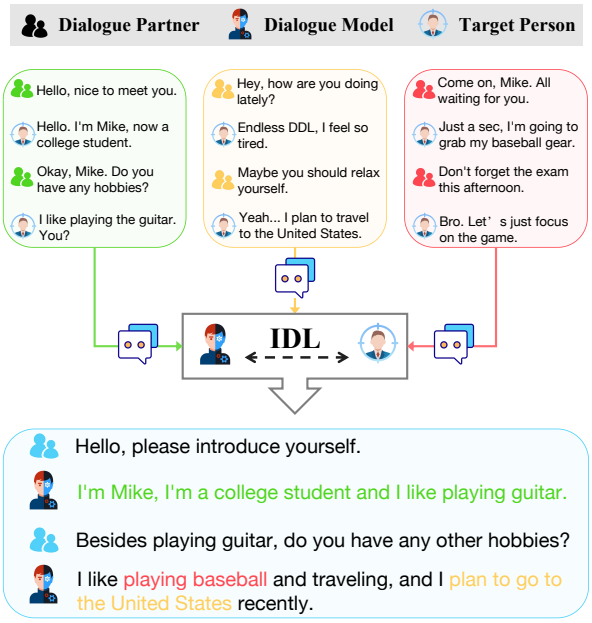

Figure 1: An example of profile-free personalized dialogue generation by In-Dialogue Learning. Persona information in different dialogues is marked with corresponding colors.

To this end, we introduce **I**n-**D**ialogue **L**earning (IDL), a two-stage framework that directly learns persona information from dialogue sessions, and leverages the learnt insights to synthesize responses that exhibit explicit personality characteristics (cf., Figure 1).

IDL comprises a Mutual Supervised Learning (MSL) stage and a Deep Personalized Alignment (DPA) stage. The objective of MSL is to equip a dialogue model with persona knowledge conveyed in dialogue sessions. To this end, one can simply select one dialogue as the target and take the remaining as the reference to perform few-shot learning to optimize the dialogue model. Such a straightforward implementation, however, suffers from two major problems: (1) unified reference dialogues normally contain abundant irrelevant information to the target dialogue, which increases

the difficulty of learning; and (2) incoherent transition in multiple dialogues could cause disruption in the dialogue structure. To address the problems, we propose Static Persona Identification (SPI) and Dynamic Persona Identification (DPI) to cluster and re-order dyadic dialogues between a target person and the other interlocutors for effective IDL. SPI divides the dialogues of the person into multiple persona-relevant clusters, ensuring that the target dialogue can easily access inter-session personalized information from reference dialogues from each cluster. DPI further re-orders the reference dialogues by minimizing the gaps in these dialogues, which is measured by conversational edit distance (convED) (Lavi et al., 2021).

To enhance the alignment of responses with the target persona, we adopt reinforcement learning through the Deep Personalized Alignment stage. We introduce Direct Preference Optimization with Criterion (DPOC), an optimization method derived from DPO (Rafailov et al., 2023) to mitigate preference degradation problem with a criterion-based penalty. This approach ensures that responses are more closely aligned with the target persona learned from reference dialogues.

We conducted experiments on several personalized dialogue datasets to evaluate the effectiveness of IDL. Evaluation results show that IDL achieves performance comparable to very strong profile-based methods, without utilizing any pre-defined profile information and supervision. In comparison to traditional personalized dialogue approaches, IDL demonstrates significant improvements, highlighting the benefits of leveraging large language models for personalized dialogue. Furthermore, IDL shows significant improvement over ICL when both utilize large language models, with BLEU and ROUGE scores increasing up to 200% and 247%, respectively. This suggests that, unlike ICL, which primarily learns from data samples, IDL is more effective at incorporating persona information within dialogues.

Our contributions are threefold:

(1) We introduce In-Dialogue Learning (IDL) as the first effort to create a personalized dialogue system using large language models without pre-defined user profiles, enabling response generation using persona information directly learned from dialogue sessions.

(2) We introduce methods for static and dynamic persona identification to improve data organization for IDL and enhance the use of persona information from dialogues. Additionally, we present DPOC, a novel reinforcement learning approach, to address preference degradation problem and align responses more precisely with the persona indicated in reference dialogues.

(3) We conduct extensive experiments on multiple datasets, showing the superior performance of IDL on personalized dialogue generation. As a profile-free method, it achieves comparable performance with profile-based methods and significantly outperforms other profile-free methods.

2 Related Work

2.1 Personalized Dialogue Systems

Personalized dialogue methods are classified into three types based on persona information acquisition. The first type uses structured databases (e.g., tables) (Zhang et al., 2018; Song et al., 2019; Wolf et al., 2019; Liu et al., 2020; Bao et al., 2019; Song et al., 2021) but faces limitations in response diversity due to data sparsity. The second type uses plain text profiles for richer information (Qian et al., 2018; Song et al., 2020; Zheng et al., 2020; Song et al., 2021; Tang et al., 2023), yet struggles to completely capture personality and requires significant effort, affecting scalability.

Different from these methods, the third type mines persona information from dialogue sessions. For example, DHAP (Ma et al., 2021) uses a transformer-based approach to analyze dialogue history for generating responses, but it ignores partner utterances, missing key persona details. MSP (Zhong et al., 2022) improves upon DHAP by using a retrieval method to collect similar dialogues from various users, yet it only selects limited tokens from these dialogues, affecting their coherence. Our method, in a broad sense, belongs to the third type. The stark difference is that we make good use of the capabilities of large language models, and significantly enhance the performance of personalized dialogue systems when no profiles are available.

2.2 In-Context Learning

In-context learning (ICL) emerges as language models scale (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023), enabling them to perform complex tasks by learning from a few contextual demonstrations (Wei et al., 2022). The ICL ability of LLMs can be enhanced by using super-

vised fine-tuning methods, involving in-context data construction and multitask learning (Chen et al., 2022; Min et al., 2021), since pre-training objectives aren't designed for ICL. Researches also show that the effectiveness of ICL relies on the choice and arrangement of demonstrations (Zhao et al., 2021; Lu et al., 2021; Chen et al., 2023a).

Our method, while looks similar to ICL, is tailored for personalized dialogue generation by organizing sessions and learning persona-related information, differing from typical supervised in-context fine-tuning. It also uniquely incorporates reinforcement learning to enhance personalized dialogue capabilities beyond ICL methods.

3 Method

We present technique details of In-Dialogue Learning (IDL) in this section. As shown in Figure 2, IDL involves two stages: Mutual Supervised Learning (MSL) and Deep Personalized Alignment (DPA). In the MSL stage, we propose static and dynamic persona identification to cluster and reorder the dialogues of the target person, and then organize these dialogues into an end-to-end form to perform supervised learning, endowing the model with the ability to leverage persona information within previous dialogues. In the DPA stage, we further extend the DPO algorithm with Criterion (abbreviated as DPOC) to address the issue of preference degradation through the incorporation of criterion examples and penalty terms, facilitating fine-grained personalized learning.

3.1 Problem Formalization

The goal of IDL is to generate responses that reflect the personality of a target person u based on his/her previous dialogues $\mathbb{D}^u$. Formally, $\forall d^{(u,v)} = (q_1, r_1, \ldots, q_t, r_t) \in \mathbb{D}^u$, $d^{(u,v)}$ represents a dialogue between u and another participant v where (q_i, r_i) is the i-th turn with q_i the utterance from v and r_i the response from u, respectively. Given the current dialogue context $C_i = (q_1, r_1, \ldots, q_i)$, the generation of IDL can be formulated as

$$r_i = \text{LM}_\Theta(C_i, \mathbb{D}^u), \tag{1}$$

where LM represents the language model, and Θ is the learnable parameters. Following the common practice, we concatenate $\mathbb{D}^u$ and C_i as the input of the LM.

3.2 Mutual Supervised Learning

IDL represents learning the personalized response generation ability conditioned on the previous dialogues. If we deem the dialogues of the target person as nodes in a graph, each of them can utilize the remaining dialogues as the reference, which can be imagined as a *complete graph*. This property induces the concept of Mutual Supervised Learning (MSL). However, the straightforward *complete graph* usage suffers from two challenges: (1) over messy historical information and (2) incoherent transition relationship. The former denotes that the messy historical information will cause the misuse of persona information when dialogues with unrelated persona knowledge are used as the reference. The latter means that the improper order of these dialogues as the reference will cause incoherent cross-dialogue transition, harming the dialogue structure. To overcome these two challenges, we propose **static and dynamic persona identification** for personalized dialogue clustering and re-ordering (as shown in the left part of Figure 2).

3.2.1 Static Persona Identification

Learning dialogue generation from a wide variety of reference dialogues is not always effective (Bao et al., 2019), especially when we aim to capture the personality characteristics embedded in the dialogues. To enhance the efficacy of the process, static persona identification partitions the dialogues of a target person into multiple persona-relevant clusters (cf., Figure 2 left). Hence, within each persona-relevant cluster, IDL can learn more meaningful mapping from reference dialogues to target dialogues. The challenge then lies in how to measure the distance between the dialogues across persona dimensions for effective dialogue clustering.

We employ a public dataset PersonaExt (Zhu et al., 2023) and train a persona extractor to recognize persona-intensive utterances in a dialogue corpus. PersonaExt segregates persona information within dialogues into triples of <*subject, relationship, object*>. The dataset defines 105 types of relationships. Based on the dataset, we develop the persona extractor (abbreviated as Ext) to directly extract the triples from the dialogue. Then, the extracted *objects* are used to locate the persona-intensive utterances. We formulate the extraction process as

$$\{p_j^{(u,v)}\}_{j=1}^n = \text{Ext}(d^{u,v}), \tag{2}$$

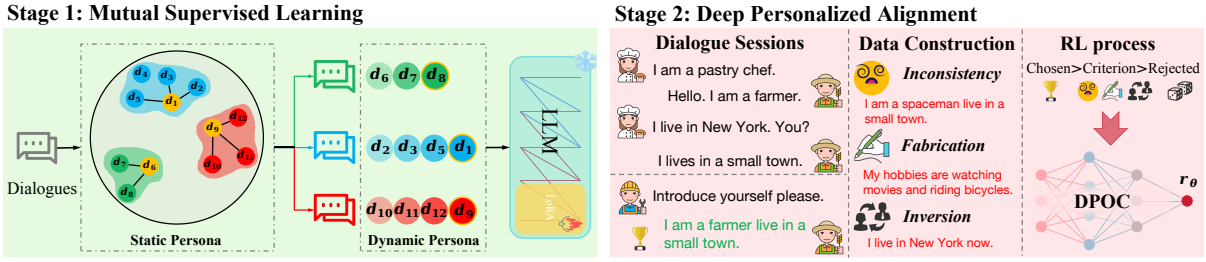

Figure 2: The framework of IDL. **Left**: the MSL stage that fine-tunes the dialogue model using data organized by static persona and dynamic persona identification. **Right**: the DPA stage in which we collect three types of criterion examples and conduct DPOC to further optimize the model to align with the target persona in a better way.

where $p_j^{(u,v)}$ denotes a persona-intensive utterance in dialogue $d^{u,v}$. The extracted utterances are then transformed to a vector $z^{(u,v)}$ by

$$\begin{aligned} p^{(u,v)} &= \text{Concat}(p_1^{(u,v)}, \ldots, p_n^{(u,v)}), \\ z^{(u,v)} &= \text{Enc}(p^{(u,v)}), \end{aligned} \tag{3}$$

where we utilize the sentence-embedding model as the Enc[1]. Based on $\{z^{(u,v)}\}$ and the euclidean metric, $\mathbb{D}^u$ is clustered by k-means algorithm:

$$K^u = \text{KMeans}(\{z^{(u,i)}\}, c), \tag{4}$$

where c is the number of clusters. Subsequently, within each cluster $K_j^u \in K^u, j = 1, 2, \ldots, c$, we randomly select a dialogue as the *target dialogue* while the closest top-k in the remaining dialogues are regarded as the *reference dialogues*.

3.2.2 Dynamic Persona Identification

Following static persona identification, we gather persona-relevant reference dialogues along with a target dialogue for optimization within each cluster. While we could directly concatenate these reference dialogues as input for the model, determining the optimal sequence remains a challenge. Our goal is to merge these dialogues into a cohesive long-term conversation, as we recognize that an inappropriate sequence could negatively affect the structure of the dialogue (Chen et al., 2023b).

To achieve the goal, we compute the optimal order which could minimize the overall semantic distance between adjacent dialogue sessions in the long-term conversation. This approach ensures a smoother transition in the ongoing dialogue.

To quantify the semantic distance between dialogues, we introduce Conversation Edit Distance (convED) (Lavi et al., 2021). The convED metric is akin to the traditional edit distance, but it modifies the basic unit of editing from characters to sentences within a dialogue. The metric aligns one dialogue with another through the processes of inserting, deleting, and substituting sentences. Detailed formulations of convED are presented in Appendix A.2.

Given a pair of dialogues (d_i, d_j), the distance $dist_{i,j} = \text{convED}(d_i, d_j)$ measures the cost of aligning d_i to d_j. Hence, by computing paired convED, we obtain a semantic distance matrix between reference dialogues in a cluster. Subsequently, we introduce Dijkstra's minimum distance algorithm (Dijkstra, 2022) to re-order the reference dialogues based on the semantic distance matrix and compute the optimal order.

In each cluster of K^u, we concatenate the reference dialogues according to the optimal order and split the target dialogue with the last utterance as a response and the remaining as the context. These data elements satisfy Equation 1, and we can optimize the LM by minimizing the negative likelihood loss. Above processes endow the model with basic IDL ability, which could generate personalized responses based on reference/historical dialogues.

Note that we utilize two kinds of distance in static and dynamic persona identification, where the former measures the personalized relevance and clusters the relevant dialogues of a target person, while the latter measures the semantic distance and re-orders the reference dialogues in a cluster.

3.3 Deep Personalized Alignment

The model after MSL initially exhibits the ability of personalized response generation by referencing some dialogues. However, due to hallucinations of LLMs (Kalai and Vempala, 2023) and complexity of long context, it still fall short in generating personalized response in a more precise manner. Consequently, we introduce the preference align-

[1]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

ment technique into IDL for Deep Personalized Alignment (DPA).

3.3.1 DPOC

The first consideration for preference alignment is Direct Preference Optimization (DPO) (Rafailov et al., 2023). DPO distinguishes itself from conventional reinforcement learning algorithms by bypassing the need for reward models, thereby conserving time and computational resources and reducing the complexity typically associated with reinforcement learning. However, DPO encounters a challenge in the form of unstable training outcomes. This instability arises because the primary objective of DPO is to widen the gap between chosen and rejected examples, while it overlooks the diminishing rewards of the chosen examples. Thus, even when the disparity between chosen and rejected examples increases, it may be caused by a concurrent decrease in rewards for both chosen and rejected examples, ultimately leading to a diminished efficacy of the optimized model. This issue is referred as **preference degradation**.

To address this problem, DPOC incorporates a corrective measure by adding a penalty term $\mathcal{P}$:

$$\mathcal{P}(r_w, r_l) = -\min\left(0, \log r_w - \log r_l\right), \quad (5)$$

where r_w is the reward of the better sample y_w and r_l is the reward of the worse sample y_l. In most cases, $r_w > r_l$ and $\mathcal{P}(r_w, r_l) = 0$. However, when $r_l > r_w$, $\mathcal{P}(r_w, r_l)$ functions as the penalty term. This inclusion ensures that the optimized model does not significantly deviate from the initial model. Building upon the foundation of DPO, the loss function of DPOC is formulated as

$$\begin{aligned}\mathcal{L}_{DPOC}(r_{cho}, r_{rej}, r_{crt}) = \mathcal{L}_{DPO}(r_{cho}, r_{rej}) \\ + \mathcal{P}(r_{cho}, r_{crt}) \\ + \mathcal{P}(r_{crt}, r_{rej})\end{aligned} \quad (6)$$

The criterion sample reward r_{crt} typically serve as intermediary benchmarks between chosen sample reward r_{cho} and rejected sample reward r_{rej}. They offer a reference point for the optimization process in DPOC. Specifically, if the reward from a chosen sample falls below that of a criterion sample, or if the reward of a rejected sample's reward is unexpectedly high compared to criterion examples, the current model incurs a penalty, which is represented by $\mathcal{P}(r_{cho}, r_{crt})$ and $\mathcal{P}(r_{crt}, r_{rej})$, respectively. This mechanism contributes to alleviating the preference degradation problem.

3.3.2 Data Construction

In the context of personalized dialogue, we identify three distinct types of criterion examples (cf., Figure 2 right). Each of them utilizes persona information with inaccuracies. (1) Inconsistency: includes information conflicting with the persona established in the dialogue sessions. (2) Fabrication: introduces personality details not mentioned in the dialogue sessions. (3) Inversion: adopts the persona information of the other participant. Given dialogue sessions $\mathbb{D}^u$, the context of on-going dialogue context C and a chosen sample h_{cho} of the current response, the construction of the three types of criterion examples are detailed as follows:

Inconsistency. We employ the personality extraction model introduced in §3.2.1, and utilize the personality triplet randomly extracted from $\mathbb{D}^u$ to substitute a triplet in h_{cho} to formulate h_{crt}. For example, h_{cho} *"I am a farmer live in a small town"* is transformed into h_{crt} *"I am a spaceman live in a small town"* by replacing *<I, job, farmer>* with *<I, job, spaceman>*, which is extracted from $\mathbb{D}^u$.

Fabrication. We encode sentences in the dataset, selecting top-m candidates with highest semantic similarity to h_{cho}. A candidate, h_{crt}, is randomly chosen ensuring $\text{Ext}(h_{crt}) \cap \text{Ext}(\mathbb{D}^u) = \emptyset$. For example, from the utterance *"My hobbies are watching movies and riding bicycles"*, we extract triples *<I, hobby, watching movies>* and *<I, hobby, riding bicycles>*. As the triples are not involved in $\text{Ext}(D^u)$, we can adopt this utterance as h_{crt}.

Inversion. In $\mathbb{D}^u$ and C, utterances are divided into R for the target person u and Q for the other participant v, then the most semantically similar utterance in Q to a chosen r_{cho} is identified as h_{crt}. For instance, for r_{cho} *"I am a farmer living in a small town"*, *"I live in New York"* from Q is selected as h_{crt}.

4 Experiments

4.1 Datasets

ConvAI2 (Dinan et al., 2020) is a high-quality English dataset focused on personalized dialogues. Each dialogue revolves around a specific profile. The dataset is expanded from the classic PersonaChat (Zhang et al., 2018) by crowd workers.

Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011) contains over $220,000$ dialogues collected from more than 600 movies with rich meta-data, offering a diverse

range of dialogues between $10,000$ pairs of characters.

LIGHT (Urbanek et al., 2019) is a large-scale crowdsourced fantasy text adventure game research platform. We extract dialogues of each character to form the dataset used in the experiments.

Note that profiles are only available in ConvAI2 and not in Cornell Movie-Dialogs Corpus and LIGHT. Implementation details are presented in Appendix A.1.

4.2 Baselines

Profile-based Approaches utilize persona information extracted from the given profiles. Along this research line, we consider the following models: *GPT-2* (Radford et al., 2019) is known for its proficiency in a variety of text generation tasks. *PerCVAE* (Zhao et al., 2017) processes the persona information as a conditional representation and employs CVAE to produce personalized responses. *BoB* (Song et al., 2021) leverages BERT for personalized dialogues by combining consistency generation task and consistency inference tasks. *CLV* (Tang et al., 2023) categorizes persona descriptions into distinct groups to enhance personalized response generation with historical queries.

Profile-free Approaches perform personalized dialogue generation without profiles. We employ *DHAP* (Ma et al., 2021) and *MSP* (Zhong et al., 2022) as baselines.

Large Language Models have made great progress in recent years. We select LLaMA-2-7B-Chat and LLaMA-2-13B-Chat (Touvron et al., 2023) as the backbones of IDL, and name the models LLaMA-2-7B IDL and LLaMA-2-13B IDL, respectively. Besides, *Vicuna*[2] and *WizardLM* (Xu et al., 2023) are involved in comparison, where the former is an open-source chatbot developed by fine-tuning LLaMA with user-shared conversations sourced from ShareGPT, and the latter is fine-tuned from LLaMA-2, starting with a basic set of instructions.

Since profiles are available in ConvAI2, we compare IDL with the profile-based approaches as well as the the profile-free approaches on this dataset. As existing profile-based approaches are not based on LLMs, we further fine-tune LLaMA-2-7B-Chat and LLaMA-2-13B-Chat with the gold profiles in ConvAI2 for fair comparison, and name the models LLaMA-2-7B gold and LLaMA-2-13B gold, respectively. On Movie and LIGHT, we assess the transferability of IDL by comparing LLaMA-2-7B IDL and LLaMA-2-13B IDL, both fine-tuned on ConvAI2, against other LLMs utilizing in-context learning method.

[2]https://lmsys.org/blog/2023-03-30-vicuna/

4.3 Evaluation Metrics

We employ various metrics to evaluate the performance of the dialogue models from the following aspects:

Coherence. BLEU-1/2 (Papineni et al., 2002) and ROUGE-L (Lin and Och, 2004) are typical word overlap-based metrics for measuring the similarity between model responses and the ground-truth.

Diversity. Distinct-1/2 (Li et al., 2015) consider the number of uni- or bi-grams in model responses, which are commonly used for evaluating diversity of dialogue generation.

Persona. Since our goal is to leverage persona information in dialogue sessions, we adopt P-F1 (Ma et al., 2021) to measure the uni-gram F1 score between the model response and the latest utterance in the context. Inspired by (Zhong et al., 2022), we use P-Co (Persona Cosine Similarity) as a supplement to the word overlap metrics to evaluate the semantic similarity between model responses and the ground-truth. Besides, following (Tang et al., 2023), we also adopt Con.Score and Coh-Con.Score to measure the consistency between model responses and the given profiles in ConvAI2.

4.4 Main Results

4.4.1 Automatic Evaluation

In Table 1, we compare the proposed method with existing personalized dialogue generation methods on ConvAI2. From the results, we can conclude that (1) when equipped with IDL, an open-source LLM can significantly outperform the existing methods in terms of almost all metrics, implying that IDL offers an effective way for leveraging LLMs in the task of personalized dialogue generation. (2) IDL can successfully recover personality characteristics from dialogue sessions. This is supported by the comparison between LLaMA-2 IDL and LLaMA-2 gold. Even without any hints from the profiles, IDL can still achieve comparable performance to the models fully supervised by the profiles.

In Table 2, we present results of IDL and other LLMs of comparable size on Movie and LIGHT. All the baseline models engage in personalized

Dataset	Model	Coherence		Diversity		Persona	
		BLEU-1	ROUGE-L	Dist-1	Dist-2	Coh.	Coh-Con.
ConvAI2	GPT-2	6.77	10.96	68.22	88.81	56.71	13.29
	PerCVAE	6.89	10.54	67.48	89.46	53.26	12.95
	BoB	7.85	12.46	63.85	85.02	62.47	15.97
	DHAP	7.21	9.90	69.86	90.23	64.27	16.04
	MSP	8.19	11.67	65.79	89.43	65.81	15.45
	CLV	11.85	15.1	71.24	92.89	71.72	**23.01**
	LLaMA-2-7B IDL	52.4†	18.98†	86.13†	96.97	96.86†	13.26†
	LLaMA-2-7B gold	**54.56**	**20.98**	**87.02**	**97.33**	**98.15**	**18.72**
	LLaMA-2-13B IDL	54.48	20.05†	87.78†	97.45†	**98.48**†	**19.63**†
	LLaMA-2-13B gold	**55.32**	**21.58**	**88.49**	**97.78**	98.1	17.77

Table 1: Automatic evalution compared to profile-based methods on ConvAI2. All of these models are trained on this dataset. The best results are in **bold** and the second best results are underlined. "†" indicates that our model passed the t-test with p-value < 0.05 in comparison to the best baseline.

Dataset	Size	Model	Coherence			Diversity		Persona	
			BLEU-1	BLEU-2	ROUGE-L	Dist-1	Dist-2	P-F1	P-Co
Movie	7B	Vicuna	14.76	5.53	5.44	71.45	63.58	11.13	17.05
		LLaMA-2 ICL	6.12	3.07	5.95	65.38	91.10	11.70	18.95
		LLaMA-2 IDL	**31.60**†	**11.74**†	**10.86**†	**89.86**†	**95.81**†	**19.95**†	**21.07**†
	13B	Vicuna	12.82	4.01	3.88	75.37	60.53	6.54	14.22
		WizardLM	29.60	10.45	9.75	87.55	94.62	18.67	20.92
		LLaMA-2 ICL	15.04	7.00	8.21	75.26	94.55	14.38	20.71
		LLaMA-2 IDL	**32.56**†	**13.00**†	**10.62**	**90.31**†	**97.24**†	**19.67**	**22.88**
LIGHT	7B	Vicuna	36.07	17.37	10.52	83.27	90.56	16.53	23.40
		LLaMA-2 ICL	15.41	8.92	9.88	67.74	93.24	16.78	**31.99**
		LLaMA-2 IDL	**46.32**†	**22.01**†	**13.45**†	**83.90**†	**94.70**†	**20.18**†	28.00†
	13B	Vicuna	19.68	8.87	5.87	59.85	58.07	8.27	16.11
		WizardLM	44.59	21.45	11.13	83.11	95.15	18.28	28.01
		LLaMA-2 ICL	24.31	13.47	10.55	75.07	96.24	17.69	**31.48**
		LLaMA-2 IDL	**49.69**†	**24.64**†	**13.24**	**87.53**†	**97.54**	**20.28**	30.95

Table 2: Automatic evalution compared to pre-trained large language models on Movie and LIGHT. The best results are in **bold** and the second best results are underlined. "†" indicates that our model passed the t-test with p-value < 0.05 in comparison to the best baseline.

dialogue through ICL. Based on the results, we observe that (1) ICL underperforms in personalized dialogue generation, indicating that while ICL can handle the textual structure of dialogue sessions, it fails to effectively utilize persona information within these dialogues and (2) LLaMA-2-7B IDL and LLaMA-2-13B IDL fine-tuned on ConvAI2 also perform well on Movie and LIGHT. This confirms that the success of IDL is not due to the optimization for a particular dataset; rather, it stems from the ability to effectively utilize persona information in dialogues.

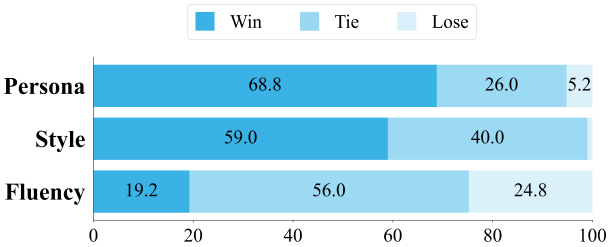

Figure 3: Human evaluation results for IDL compared to ICL. Both methods adopt LLaMA-2-13B-Chat.

4.4.2 Human Evaluation

We incorporate human evaluation to more accurately assess the quality of dialogues on three subjective dimensions: **(1) Persona**: evaluators will assess whether the response accurately and consistently reflects the persona information of the target person. **(2) Style**: evaluators will judge if the response aligns with the expected wording and tone for the target person. **(3) Fluency**: evaluators will examine the smoothness of the dialogue flow, con-

sidering both linguistic and logical fluency. We arranged the generated responses into pairs and conducted pairwise comparisons across these three dimensions.

Human evaluation results on ConvAI2 are shown in Figure 3. We sampled 500 pairs and engaged a professional evaluation group to perform the assessments. The two responses within each pair are produced from identical dialogue sessions and contexts, and the order of these two responses is randomized in the evaluation system. For each of the three dimensions mentioned previously, evaluators are required to assign a judgment of Win, Tie, or Lose based on the quality of these two responses.

The results show that IDL has brought significant improvements in both persona and style, with winning rates of 68.8% and 59.0% respectively, which demonstrates that the model using IDL can more effectively simulate the personality and tone of the target person. Regarding fluency, there is a slight decline in performance when using IDL, possibly attributed to the model's increased focus on aligning with persona information.

4.5 Discussions

4.5.1 Ablation Study

Model	BLEU	ROUGE	P-F1	P-Co
IDL	**32.56**	**13.00**	**19.67**	**22.88**
w/o Criterion	31.58	10.55	17.76	21.79
w/o DPA	31.25	10.89	18.98	21.12
w/o SPI	29.94	10.93	19.02	21.14
w/o DPI	28.8	9.60	18.46	21.01

Table 3: Ablation study on Movie.

Table 3 shows the ablation study results on Movie. In order to clarify the contribution of each IDL process to the overall effect, we gradually remove each process and get a list of variants: **(a)** *w/o* Criterion removes the criterion samples and uses standard DPO for persona alignment. **(b)** *w/o* DPA removes the whole persona alignment process. **(c)** *w/o* SPI further removes the static persona identification in the MSL stage on the basis of (b). **(d)** *w/o* DPI removes the dynamic persona identification on the basis of (c).

From the results, we observe that (1) DPOC plays a crucial role in enhancing the acquisition of better persona information, and the elimination of criterion samples significantly diminishes the model's effectiveness. This is because the model can pay more attention to persona-related tokens after deep personalized alignment. Relevant case study can be found in Appendix A.3. Additionally, the findings suggest that merely employing DPO falls short in substantially improving the overall performance of models. This is because the preference alignment of DPO is not optimized for problems that can arise from personalized dialogue generation task, as illustrated in § 3.3.2. Furthermore, the diminished effectiveness observed upon removing static and dynamic persona identifiers underscores the importance of reorganizing training data before the supervised fine-tuning process.

4.5.2 Effect of Sessions

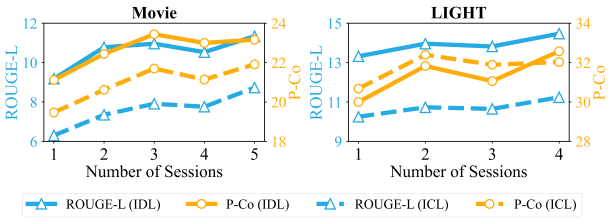

Figure 4: Experiments with different numbers of dialogue sessions on the Movie and LIGHT.

In this work, we make the model learn personality-related information from the dialogue sessions and generate personalized responses. We present the performance of IDL and ICL under different demonstrations (dialogue sessions) to compare the learning efficiency of them. Figure 4 illustrates that similar to ICL, with the increase in the number of dialogue sessions, there is a general improvement in the quality of responses of IDL. However, as a specialized learning method for dialogue, IDL exhibits a faster learning ability under different dialogue sessions than ICL, indicating the effectiveness of our proposed mutual supervised learning and deep personalized alignment. Benefits from these advancements, IDL paves a new road to develop and update dialogue systems in an online manner.

5 Conclusion

In this study, we introduce a framework In-Dialogue Learning (IDL) designed for personalized dialogue generation task. Unlike previous approaches, our framework directly derives persona information from dialogues without the need of pre-defined profiles and is widely applicable to LLMs. The efficacy of IDL in producing personalized responses is validated through both automatic and human evaluation results.

Limitations

First, given the complexity of large-scale experiments, we limited our research to the more representative LLaMA-2 series models. This approach does not ensure favorable outcomes across all pre-trained large language models. Moreover, the capacity of IDL to manage highly diverse or conflicting persona traits within dialogue sessions has not been examined, which may restrict its use in situations involving non-coherent or changing user identities. Additionally, while the datasets employed in our study consistently includes personality information within dialogues, this may not hold true in real-world applications.

Ethics Statement

Dialogues and persona information often contain sensitive information about individuals, which could result in breaches of privacy. We took measures to ensure that the datasets utilized in our experiments were strictly confined to the scope of the study and did not include any sensitive personal information.

The datasets employed in this research are publicly available, and the models we utilize adhere to their licenses, meeting both academic standards and ethical guidelines.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Liang Chen, Hongru Wang, Yang Deng, Wai-Chung Kwan, Zezhong Wang, and Kam-Fai Wong. 2023b. Towards robust personalized dialogue generation via order-insensitive representation regularization. *arXiv preprint arXiv:2305.12782*.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. *arXiv preprint arXiv:2205.01703*.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.

Edsger W Dijkstra. 2022. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: His Life, Work, and Legacy*, pages 287–290.

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.

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.

Adam Tauman Kalai and Santosh S Vempala. 2023. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*.

Ofer Lavi, Ella Rabinovich, Segev Shlomov, David Boaz, Inbal Ronen, and Ateret Anaby-Tavor. 2021. We've had this conversation before: A novel approach to measuring dialog similarity. *arXiv preprint arXiv:2110.05780*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. *arXiv preprint arXiv:2004.05388*.

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 & Knowledge Management*, pages 1350–1359.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

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.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

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 O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Ijcai*, pages 4279–4285.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

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. *arXiv preprint arXiv:2106.06169*.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020. Profile consistency identification for open-domain dialogue agents. *arXiv preprint arXiv:2009.09680*.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. *arXiv preprint arXiv:1905.12188*.

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. *arXiv preprint arXiv:2305.11482*.

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*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. *arXiv preprint arXiv:1903.03094*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

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

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

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. *arXiv preprint arXiv:2204.08128*.

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.

A Appendix

A.1 Implementation Details

We experimented with a range of parameter combinations in our study. We adopt LLaMA-2-7B-Chat[3] and LLaMA-2-13B-Chat[4] as the backbones. The parameters utilized to obtain the experimental results presented in this chapter are as follows: In the MSL stage, the maximum number of clusters c is set to 3 and the maximum number of nearest neighbor k is set to 5. Scaling coefficient λ is set to 5. We adopt Lora for training. The batch size is 4 and the learning rate is $5e-5$. In the DPA stage, the penalty of DPOC is set to 2. The batch size is set to 1 and the learning rate is $1e-5$. The model used as persona extractor is LLaMA-2-7B fine-tuned on PersonaExt.

A.2 convED

Similar to Edit distance, convED also employs three operations: Insertion, Deletion, and Substitution. It calculates the shortest distance using Dynamic Programming (DP). However, unlike Edit distance, convED operates on sentences within dialogues, resulting in a distinct approach to distance calculation.

Assuming dialogue A comprises m sentences and dialogue B comprises n sentences, we obtain an $m \times n$ matrix lev, where $\text{lev}(i,j)$ represents the shortest edit distance between the first i sentences of dialogue A and the first j sentences of dialogue B. The costs of the three operations of convED are as follows:

Insertion Insert B_j into dialogue A. The edit distance lev_{ins} is updated as:

$$\text{lev}_{ins}(i,j) = \text{lev}(i,j-1) + 1 \tag{7}$$

Deletion Delete A_i from dialogue A. The edit distance lev_{del} is updated as:

$$\text{lev}_{del}(i,j) = \text{lev}(i-1,j) + 1 \tag{8}$$

Substitution Substitute sentence A_i to align with B_j. The edit distance lev_{sub} is updated as:

$$\text{lev}_{sub}(i,j) = \text{lev}(i-1,j-1) + \lambda \cdot w_{sub}(A_i, B_j) \tag{9}$$

The scale parameter λ regulates the substitution cost, with both insertion and deletion costs being fixed at 1. w_{sub} is a function that calculates the semantic similarity of two sentence vectors:

$$w_{sub}(s_1, s_2) = \begin{cases} \infty & \text{if } r(s_1) \neq r(s_2) \\ 1 - \cos(Enc(s_1), Enc(s_2)) \end{cases} \tag{10}$$

where Enc is the encoder, used to encode sentences into vector space. It's important to highlight that sentences uttered by different individuals in a conversation, even if they share semantic similarities, cannot be aligned through substitution. Consequently, the function $r(*)$ is employed to identify the speaker of a sentence. Cosine similarity is then calculated for sentences from the same speaker, while the substitution cost between sentences from different speakers is considered infinite.

Finally, $\text{lev}(i,j)$ is the minimum cost of these three operations:

$$\text{lev}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \text{lev}_{ins}(i,j) \\ \text{lev}_{del}(i,j) \\ \text{lev}_{sub}(i,j) \end{cases} & \text{otherwise} \end{cases}$$

A.3 Case Study

To investigate the specific content within dialogue sessions that a model trained with IDL focuses on when crafting responses, we conducted an analysis of the attention weights during the reply generation process, as illustrated in Figure 5. We identified the top 100 tokens receiving the highest attention within the dialogue sessions and examined their correspondence with the personality-related keywords found in the gold profile. The experimental findings indicate that the LLaMA-2-13B-Chat model typically concentrates on an average of 9 keywords. However, the same model, once implemented with IDL, shows an enhanced focus on 13 keywords. This improvement suggests that IDL significantly enhances the model's ability to precisely leverage persona information within dialogues.

[3] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[4] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

Profile

I love to go to disney world every year.
I love to sing songs from the movie frozen.
I play with my friends on the playground.
I'm in the third grade.
Mickey mouse is my favorite character.

Dialogue Sessions

Q: What are you doing tonight. i am listening to music.
R: I am singing **songs** from **frozen** my favorite movie .
Q: I liked the movie too. how old are you?
R: 8 in third **grade**. i like to **play** on the **playground** with my **friends**. you?
Q: I watch kids on the playground. i teach third grade .
R: Awesome! have you ever been to **disney** **world**? i **go** every **year**.
Q: I haven't had the pleasure but i am sure i'll soon!
R: You have to meet **mickey** **mouse** he's my **favorite**. Do you have kids?
Q: I don't have kids yet except in class and i love playing games with them .
R: What games do you play?
Q: I like scrabble, stratego, risk and clue a lot plus many others, you?

Response

Gold: I like sorry and monopoly. do you have pets?
IDL: I like monopoly and chess. we play games on the playground too.

Figure 5: A case study. Keywords in the profile are marked in red, while the corresponding keywords that have high attention weight within dialogue sessions are bolded and highlighted with a yellow background.