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

Anonymous ACL submission

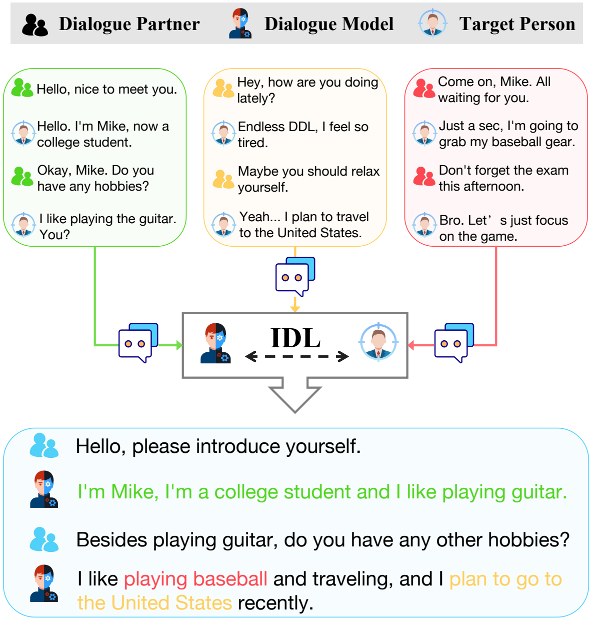
Abstract

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 **In-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 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 personalized dialogue systems (Tang et al., 2023; Chen et al., 2023c; Huang et al., 2023; Chen et al., 2023a; Tu et al., 2022). Such systems are often adept at incorporating special personal characteristics into responses. Consequently, they offer enhanced flexibility, enabling them to adapt more effectively to a wide range of conversational scenarios, such as personal assistants or chatbots [1].

A common practice in personalized dialogues is to condition a dialogue model on a pre-defined profile that explicitly depicts the personality traits one aims to portray with a textual description. While there have been extensive studies along this line (Song et al., 2021; Liu et al., 2022; Chen et al., 2023b), we explore the problem from a different angle: instead of using a brief profile to describe a person’s personality, we leverage multiple conversations between the individual and others to build a personalized dialogue model. Consequently, the model can generate personalized dialogues without the need for pre-designed profiles which could be both time-consuming and labor-intensive. Furthermore, as dialogue history accumulates, past conversations may provide more personalized information than a static profile.



Hello, nice to meet you.
Hello. I'm Mike, now a college student.
Okay, Mike. Do you have any hobbies?
I like playing the guitar. You?

Hey, how are you doing lately?
Endless DDL, I feel so tired.
Maybe you should relax yourself.
Yeah... I plan to travel to the United States.

Come on, Mike. All waiting for you.
Just a sec, I'm going to grab my baseball gear.
Don't forget the exam this afternoon.
Bro. Let’ s just focus on the game.

Hello, please introduce yourself.
I'm Mike, I'm a college student and I like playing guitar.
Besides playing guitar, do you have any other hobbies?
I like playing baseball and traveling, and I plan to go to the United States recently.

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

We introduce **In-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

[1] https://character.ai/

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 better align responses with the target persona (Ouyang et al., 2022; Yuan et al., 2023; Song et al., 2023; Hong et al., 2023), 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 conduct 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. 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 In-Context Learning (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 (single-turn), IDL is more effective at incorporating persona information within dialogues (multi-turn).

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 predefined 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 supervised 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

As shown in Figure 2, In-Dialogue Learning (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 re-order 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, $\forall d^{(u,v)} = (q_1, r_1, \dots, q_t, r_t) \in \mathbb{D}^u$, $d^{(u,v)}$ represents a dialogue between user u and the other 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 $C_i = (q_1, r_1, \dots, q_i)$, the generation of IDL can be formulated as

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

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.

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) that can directly extract the above-mentioned triples from a dialogue by fine-tuning an LLM. More details about training of the persona extractor are presented in

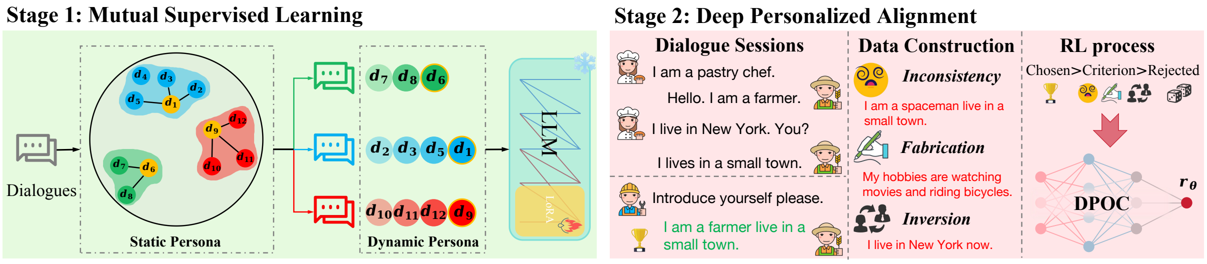

Stage 1: Mutual Supervised Learning
Dialogues
d_4 d_3 d_2 d_5 d_1
d_9 d_{12} d_{10} d_{11}
d_7 d_6 d_8
Static Persona
d_7 d_8 d_6
d_2 d_3 d_5 d_1
d_{10} d_{11} d_{12} d_9
Dynamic Persona
LLM
LoRA
Stage 2: Deep Personalized Alignment
Dialogue Sessions
I am a pastry chef.
Hello. I am a farmer.
I live in New York. You?
I lives in a small town.
Introduce yourself please.
I am a farmer live in a small town.
Data Construction
Inconsistency
I am a spaceman live in a small town.
Fabrication
My hobbies are watching movies and riding bicycles.
Inversion
I live in New York now.
RL process
Chosen>Criterion>Rejected
DPOC
r_θ

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.

Appendix A.2. The persona extractor identifies and extracts persona-intensive utterances from a dialogue by recognizing utterances that contain at least one object in the extracted triples. Formally, the extraction process can be formulated as

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

where $p_j^{(u,v)}$ is a persona-intensive utterance in dialogue $d^{(u,v)}$, with u and v as the participants. $\{p_j^{(u,v)}\}_{j=1}^{n}$ are then encoded as the persona representation $z^{(u,v)}$ of $d^{(u,v)}$ 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 Enc$(\cdot)$ is an off-the-shelf sentence encoder[2]. Based on $\{z^{(u,v)}\}$, dialogues $\{d^{(u,v)}|d^{(u,v)} \in \mathbb{D}^u\}$ are clustered by k-means algorithm[3]:

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

where $z^{(u,v)}$ serves as the index of dialogue $d^{(u,v)}$ and 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). It 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.3.

Given a pair of dialogues (d_i, d_j), the distance $dist_{i,j}$ = 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. The above processes endow the model with basic IDL ability, which could generate personalized responses based on reference dialogues.

Note that we utilize two kinds of distance in static and dynamic persona identification, where the former measures the personalized relevance and

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

[3] We tested various clustering algorithms, including k-means, BSCAN, Mean Shift, WARD, and BIRCH, but observed no significant differences. Therefore, we chose k-means for its simplicity and widespread use.

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), it still falls short in generating more precise personalized responses. Consequently, we propose Deep Personalized Alignment (DPA) in IDL.

3.3.1 DPOC

Previous DPO (Rafailov et al., 2023) method for preference alignment encounters a challenge in the form of unstable training outcomes. This instability is aroused by the primary objective of DPO, which maximizes the generation probability gap between chosen and rejected examples. This objective will overlook 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 $\mathcal{P}$:

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

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

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

The criterion sample reward r_{crt} typically serves as intermediary pivots between chosen sample reward r_{cho} and rejected sample reward r_{rej}. 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. Detailed formulations are presented in Appendix A.4.

3.3.2 Data Construction

To perform DPOC, we need to specify the criterion samples. The intuition of criterion sample construction comes from analysis of the model after the MSL stage, where we observe three major problems, including responses revealing fictitious persona information, conflicting with the persona set by the context, and confusing the partner's persona with the target person's. Based on the analysis, we consider the following three types of criterion samples (cf., Figure 2 right): (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 ongoing dialogue 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 the 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 dialoigues 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.2.

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 recently. 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[4] and WizardLM (Xu et al., 2023) are involved in comparison, where the former is an open-source chatbot developed by fine-tuning LLaMA with ShareGPT, and the latter is fine-tuned from LLaMA-2, starting with a basic set of instructions.

In the ConvAI2 dataset, we compare our IDL models with both profile-based and profile-free approaches. Unlike existing profile-based methods that don't use Large Language Models (LLMs), we fine-tune LLaMA-2 models (7B and 13B versions) with ConvAI2's profiles for a fair comparison, naming them LLaMA-2-7B gold and LLaMA-2-13B gold. We also include two other LLM baselines: LLaMA-2 System, which uses profiles directly in system instructions without further training, and LLaMA-2 FT, which fine-tunes on ConvAI2 treating each conversation as a separate example.

For the Movie and LIGHT datasets, we test the adaptability of our IDL models (LLaMA-2-7B IDL and LLaMA-2-13B IDL, both fine-tuned on ConvAI2) against other LLMs using ICL.

4.3 Evaluation Metrics

We employ BLEU (Papineni et al., 2002) and ROUGE-L (Lin and Och, 2004) metrics to assess the coherence of the text.[5] For evaluating diversity, Distinct-1/2 (Li et al., 2015; Lv et al., 2023) metrics are utilized. Additionally, P-F1 (Ma et al., 2021), P-Co (Persona Cosine Similarity) (Zhong et al., 2022) are used to measure persona consistency, while Con.Score, and Coh-Con.Score are used to measure the consistency between model responses and the given profiles in ConvAI2 (Tang et al., 2023).

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 discover persona information from dialogue sessions, comparing LLaMA-2 IDL with 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 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

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

[5] We use NLTK to calculate both metrics.

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	15.10	71.24	92.89	71.72	**23.01**
	LLaMA2-7B System	7.22	9.56	72.21	94.39	98.87	**22.32**
	LLaMA2-7B FT	50.23	18.04	**88.32**	**97.45**	97.41	8.63
	LLaMA2-7B IDL	52.40†	18.98†	86.13†	96.97†	96.86†	13.26†
	LLaMA2-7B gold	**54.56**	**20.98**	87.02	97.33	**98.15**	18.72
	LLaMA2-7B System	11.80	10.39	76.46	94.88	98.92	19.30
	LLaMA2-7B FT	51.80	18.14	88.29	**97.80**	97.64	9.71
	LLaMA2-13B IDL	54.48†	20.05†	87.78†	97.45†	**98.48**†	**19.63**†
	LLaMA2-13B gold	**55.32**	**21.58**	**88.49**	97.78	98.10	17.77

Table 1: Automatic evaluation on ConvAI2. All models are trained on this dataset. The best results are in **bold** and the second best results are underlined. "†" indicates that our model passed t-test with p-value < 0.05 in comparison to the best baseline. Results on BLEU-2/3/4 are presented in A.1.

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 evaluation on Movie and LIGHT. The best results are in **bold** and the second best results are underlined. "†" indicates that our model passed t-test with p-value < 0.05 in comparison to the best baseline. Results on BLEU-3 and BLEU-4 are presented in Appendix A.1.

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.

4.4.2 Human Evaluation

We hire 5 well-educated volunteers as annotators, and require them to judge the quality of model responses from three aspects: **(1) Persona**: the annotators assess whether a response accurately and consistently reflects the persona information of the target person. **(2) Style**: the annotators judge if the response aligns with the expected wording and tone for the target person. **(3) Fluency**: the annotators examine the smoothness of the dialogue

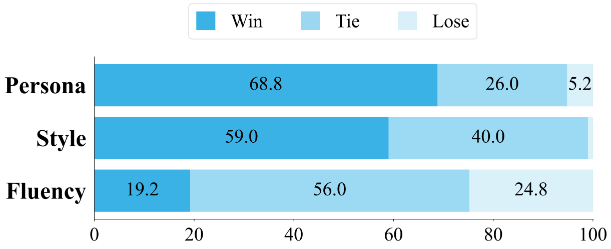


Figure 3: Human evaluation results for IDL compared to ICL. Both methods adopt LLaMA-2-13B-Chat.

flow, considering both linguistic and logical fluency. We sample 500 dialogues associated with demonstrations from the test set of ConvAI2, and obtain responses for each dialogue from IDL and ICL (based on LLaMA-2-13B-Chat), respectively. Each time, a pair of responses are randomly shuffled and presented to the 5 annotators. Each annotator assign labels from {Win, Tie, Lose} to a pair according to Persona, Style, and Fluency, and in total, each pair obtains 5 labels on each of the three aspects. Figure 3 shows the evaluation results. IDL significantly improves upon ICL on both persona and style, with winning rates of 68.8% and 59.0%, respectively, demonstrating 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. We calculate Cohen's kappa, and the values for persona, style, and fluency are 0.53, 0.56 and 0.51, respectively, indicating moderate agreement among the annotators.

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	28.80	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.5. 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

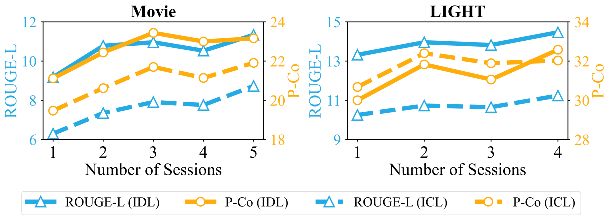


Figure 4: Experiments with different numbers of dialogue sessions on the Movie and LIGHT.

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 pretrained 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.

Jixiang Hong, Quan Tu, Changyu Chen, Xing Gao, Ji Zhang, and Rui Yan. 2023. Cyclealign: Iterative distillation from black-box llm to white-box models for better human alignment.

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*.

Ang Lv, Jinpeng Li, Yuhan Chen, Gao Xing, Ji Zhang, and Rui Yan. 2023. DialoGPS: Dialogue path sampling in continuous semantic space for data augmentation in multi-turn conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1280, Toronto, Canada. Association for Computational Linguistics.

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*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

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.

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*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*.

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*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.

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*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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 Supplementary Results

We present the evaluation results in terms of BLEU-1/2/3/4 in Table 4. The prompt used in LLaMA2-System is “Here are your persona settings, your reply must be consistent with the persona: {profile}”. From the results, we can conclude that IDL holds consistent advantages over baseline methods on all BLEU metrics.

A.2 Implementation Details

In-Dialogue Learning. In the Mutual Supervised Learning stage, the maximum cluster number c is set to 3 and the maximum number of neighbors k is set to 5. Besides, the scaling coefficient λ is set to 5. We use LoRA for training. The rank is set to 8 and the lora_alpha is set to 8. We adopt AdamW as the optimizer. We set adam_beta1, adam_beta2 and adam_epsilon to 0.9, 0.999 and $1e^{-8}$, respectively. We use cosine schedule to warm up. The batch size is set to 4 and the learning rate is set to $5e^{-5}$. We finetune our model on ConvAI2 dataset for 2 epochs. Each epoch takes around 40 minutes. The training of this process is completed on one Nvidia A100 GPU.

In the Deep Personalized Alignment stage, we set the penalty of DPOC to 2. The batch size is set to 1 and the learning rate is set to $1e^{-5}$. We use LoRA for training. The rank is set to 8 and the lora_alpha is set to 8. We adopt AdamW as the optimizer. We set adam_beta1, adam_beta2 and adam_epsilon to 0.9, 0.999 and $1e^{-8}$, respectively. We use cosine schedule to warm up.

We fine-tune our models on the DPOC dataset, which is built based on ConvAI2. For each sample, we use the ground truth of ConvAI2 as the chosen sample, and select the one with the worst quality among the candidate responses (the candidate responses has been sorted by quality in the ConvAI2 provided by parlai) as the rejected sample. As for the criteria sample, we randomly select a type (line 386) and build it according to its corresponding construction method. The number of training epochs is set to 1. Each epoch takes around 12 hours. The training of this process is completed on one Nvidia A100 GPU.

Persona Extractor. The original personaExt dataset contains 35K samples, with each sample containing a sentence and a triple <subject, relation, object>, which is a description of the persona

Dataset	Size	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ConvAI2	124M	MSP	8.19	2.34	1.56	0.73
	125M	CLV	11.85	4.29	2.41	1.02
	7B	LLaMA-2 IDL	52.40	25.43	12.74	8.40
	7B	LLaMA-2 gold	54.56	27.98	15.37	10.80
	13B	LLaMA-2 IDL	54.48	28.42	14.77	9.89
	13B	LLaMA-2 gold	55.32	28.71	16.00	11.33
Movie	7B	Vicuna	14.76	5.53	2.36	1.29
	7B	LLaMA-2 ICL	6.12	3.07	1.36	0.72
	7B	LLaMA-2 IDL	31.60	11.74	4.89	2.87
	13B	Vicuna	12.82	4.01	1.43	0.70
	13B	WizardLM	29.60	10.45	6.03	3.83
	13B	LLaMA-2 ICL	15.04	7.00	3.65	2.07
	13B	LLaMA-2 IDL	32.56	13.00	5.56	3.32
LIGHT	7B	Vicuna	36.07	17.37	7.42	3.83
	7B	LLaMA-2 ICL	15.41	8.92	3.99	2.02
	7B	LLaMA-2 IDL	46.32	22.01	9.63	5.27
	13B	Vicuna	19.68	8.87	3.53	1.75
	13B	WizardLM	44.59	21.45	10.30	5.53
	13B	LLaMA-2 ICL	24.31	13.47	6.00	3.08
	13B	LLaMA-2 IDL	49.69	24.64	10.90	6.03

Table 4: Evaluation results w.r.t. BLEU 1-4.

information in the sentence. For example, the triple of the sentence "I have an apple, it is juicy." is <I, have, apple>. In order to adapt to the format of the conversation, we simply modified the original dataset by stacking multiple samples together to simulate the form of multiple sentences in a conversation. The modified dataset has 4K samples.

We select LLaMA-2 7B as the base model of Persona Extractor, and formalize the learning task as sequence-to-sequence generation (i.e., the model decodes the triples from a given sentence). Therefore, as for the input data, we concatenate sentences to form the input sequence, and use \n as the separator. The form of output data is similar to the input data. We treat the triples corresponding to the statements as strings "<object, relation, subject>", and concatenate them to form the output sequence with \n as the separator. The order of triples in the output sequence is the same as the order of their corresponding sentences in the input sequence.

We use 90% of the data as the training set and the remaining 10% as the validation set. To test the accuracy of the trained Persona Extractor, for each sample in validation set, we concatenate sentences to form the input sequence. After that, we can get the output sequence of the Persona Extractor. We use the regex "<.*?,.*?,.*?>" to parse the output sequence. If one element of the triple is different from the ground truth, then the sample is judged "fals". The accuracy of the Persona Extractor reached 87% in validation.

A.3 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 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.4 Completed Loss Function for DPOC

Please refer to Equation 11.

A.5 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.

A.6 Low-resource Scenarios

We hope that the current conversation and its historical conversations are similar, so that the model can get more relevant information available from historical conversations. Therefore, in the Mutual Supervised Learning stage, we cluster input conversations so that they share more persona information (Static Persona Identification) and minimize the distance between the current conversation and historical conversations (Dynamic Persona Identification).

Of course, even so, we still can not guarantee that current conversation is similar to its historical conversations. Therefore, we add the following experiments. For each sample in original ConvAI2 test set, we replace the historical conversations with those in other clusters, so that the similarity between the historical conversations and the target conversation in this example is reduced. Results are shown in Table 5. The model used in this experiment is LLaMA-2 IDL 13B.

We can observe that the performance of the model has declined slightly. This is because the similarity between the historical conversations and the current conversation is the basic guarantee for IDL to be effective. When the current conversation involves a certain topic, IDL will focus on similar parts in historical conversations, thus completing the simulation of the persona. Therefore, when the similarity between the historical conversations and the current conversation decreases, the performance of IDL will also be affected.

A.7 Generalizability

To assess the generalizability of IDL, we also utilize LLaMA as the base model. Experimental results are presented in Table 6 and Table 7.

We repeated the training process of LLaMA-2 IDL using LLaMA and obtained LLaMA IDL. According to the results, LLaMA IDL surpasses Vicuna in multiple metrics, which further illustrates the effectiveness of IDL.

Size	Similarity	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	Dist-1	Dist-2	Coh.	Coh-Con
13B	Original	54.48	28.42	14.77	9.89	20.05	87.78	97.45	98.48	19.63
13B	Out-of-Cluster	52.37	25.93	11.69	6.97	16.49	90.84	98.95	97.82	6.60
7B	Original	52.40	25.43	12.74	8.40	18.98	86.13	86.97	96.86	13.26
7B	Out-of-Cluster	51.45	24.53	11.56	6.84	16.47	87.70	97.99	95.27	7.15

Table 5: Results for low-resource scenario.

Model	BLEU-1	BLEU-2	ROUGE-L	Dist-1	Dist-2	P-F1	P-Co
Vicuna 7B	14.76	5.33	5.44	71.45	63.58	11.13	17.05
LLAMA 7B ICL	11.14	3.70	4.90	53.13	55.85	10.25	15.33
LLAMA 7B IDL	22.55	9.01	8.79	59.58	70.53	20.63	20.11
Vicuna 13B	12.82	4.01	3.88	75.37	60.53	6.54	14.22
LLAMA 13B ICL	11.27	3.74	4.42	45.85	46.49	8.83	14.85
LLAMA 13B IDL	24.11	9.69	8.79	68.02	78.43	18.25	20.91

Table 6: Generalizability experiment results on Movie dataset

Model	BLEU-1	BLEU-2	ROUGE-L	Dist-1	Dist-2	P-F1	P-Co
Vicuna 7B	36.07	17.37	10.52	83.27	90.56	16.53	23.4
LLAMA 7B ICL	19.39	7.80	6.83	61.36	64.83	10.75	17.01
LLAMA 7B IDL	46.89	23.18	13.87	80.68	93.48	24.07	31.21
Vicuna 13B	19.68	8.87	5.87	59.85	58.07	8.27	16.11
LLAMA 13B ICL	22.57	9.18	7.19	57.30	60.78	12.25	17.98
LLAMA 13B IDL	48.77	24.25	13.99	83.42	95.70	23.52	31.75

Table 7: Generalizability experiment results on LIGHT dataset

$$
\begin{aligned}
\mathcal{L}_{DPOC}(\pi_\theta;\pi_{ref}) = -\,\mathbb{E}_{(x,y_w,y_r,y_l)\sim\mathcal{D}}[\log\sigma(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)})\\
-\min(0,\lambda\log\frac{\pi_{\theta(y_w|x)}}{\pi_{ref}(y_w|x)} - \lambda\log\frac{\pi_{\theta(y_r|x)}}{\pi_{ref}(y_r|x)})\\
-\min(0,\lambda\log\frac{\pi_{\theta(y_r|x)}}{\pi_{ref}(y_r|x)} - \lambda\log\frac{\pi_{\theta(y_l|x)}}{\pi_{ref}(y_l|x)})],
\end{aligned}
\tag{11}
$$

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.