A Persona-Aware LLM-Enhanced Framework for Multi-Session Personalized Dialogue Generation

Dongshuo Liu[1], Zhijing Wu[1,†], Dandan Song[1], Heyan Huang[1]
[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China,
{dsliu0817, zhijingwu, sdd, hhy63}@bit.edu.cn

Abstract

Multi-session personalized dialogue generation is one of the most important topics in open-domain dialogue. It aims to generate responses consistent with the dialogue history and personality information across multiple sessions to engage users' interest in the dialogue. Recent approaches focusing on history modeling and persona modeling have advanced the development of this field. However, they overlook the importance of dialogue structure in helping large language models (LLMs) understand the dialogue context. Moreover, these methods do not efficiently expand and utilize personality information, reducing the responses' consistency. In this paper, we propose a Persona-Aware LLM-enAnCEd(PALACE) framework for multi-session personalized dialogue generation. Specifically, the framework consists of three components: a topic-aware memory bank, a persona prompt learning module, and VAE-LoRA. The topic-aware memory bank works by retrieving historical information that possesses a certain dialogue structure and relevant topics. The persona prompt learning module enhances the LLM's persona-aware capabilities by utilizing a persona commonsense knowledge graph and a query-driven graph neural network. Furthermore, to enhance the generative capabilities of the LLM and obtain more useful prior knowledge, we combine VAE with LoRA to propose VAE-LoRA. Experimental results on the MSC and DuLeMon dataset demonstrate that our framework outperforms the state-of-the-art methods in automatic and human evaluation metrics*.

1 Introduction

Personalized dialogue generation has become one of the crucial tasks in open-domain dialogue systems (Zhang et al., 2018). It aims to generate responses consistent with personality information and dialogue history. Personality information can make generated responses more specific and comprehensive, thereby improving user engagement (Kwon et al., 2023). Most existing personalized dialogue models focus exclusively on the dialogue history of the current session (Dinan et al., 2019). It remains unable to establish long-term interactions with humans in multi-session dialogue (Chen et al., 2024), especially in multi-session personalized dialogue generation. In real dialogue scenarios, it is often necessary for dialogue models to possess the ability for long-term companionship and familiarity (Li et al., 2024a) .As shown in Figure 1, in authentic dialogues, multi-session personalized dialogue encompasses dialogue history from multiple sessions, significantly surpassing the constraints of traditional personalized dialogue limited to 16 turns (Huang et al., 2020).

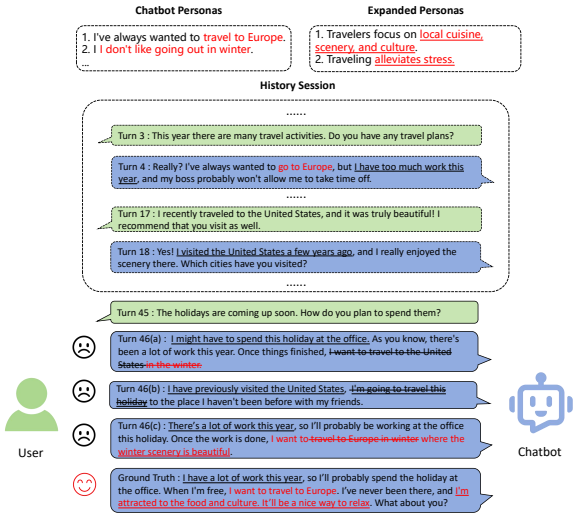


Figure 1: An example of varying response outcomes from LLMs in multi-session personalized dialogue. Text related to dialogue history is underlined, and personas are highlighted in red. Strikethrough text indicates inconsistency with history or persona. Text that is both highlighted in red and underlined represents expansions on persona.

[†] Corresponding authors.
*Code: https://github.com/Dreamer-learning/PALACE_

Findings of the Association for Computational Linguistics: ACL 2025, pages 103–123

Existing methods for multi-session personalized dialogue generation primarily focus on **history modeling** and **persona modeling**. History modeling can be further divided into summary-based, expanded context window, and retrieval-based methods (Zhong et al., 2023; Kim et al., 2024b; Wang et al., 2024b; Lee et al., 2023a). Summary-based methods typically use prompts or instructions to enable generative models to summarize dialogue history. Expanded context window methods increase input capacity by altering the generative model's architecture (Li et al., 2024b) but may risk **losing critical historical information or introducing noise**. For example, in "Turn 46(a)" of Figure 1, the chatbot infers "I might have to spend this holiday at the office" but loses the historical detail "I visited the United States". To direct the model's attention to critical historical information, many studies have explored the use of retrieval-based methods. They utilize retrieval models to obtain relevant dialogue information (Liu et al., 2023b). While these approaches allow access to pertinent historical information, they can **disrupt the dialogue structure**, hindering the model's understanding of dialogue history (Yin et al., 2023). As shown in "Turn 46(b)" in Figure 1, although the chatbot can retrieve that "I visited the United States", it overlooks the statement "I have too much work this year" mentioned in "Turn 4", resulting in a response inconsistent with the dialogue history.

Persona modeling aims to expand the sparse personality information in multi-session personalized dialogue (Zheng et al., 2019; Li et al., 2016; Cao et al., 2022). Existing expansion methods primarily involve using external knowledge bases (Lim et al., 2023; Liu et al., 2022) or mining new personality data (Zhou et al., 2021; Tang et al., 2023a). The former provides more explicit character knowledge (Zhou et al., 2023), while the latter can reveal implicit personality present in the dialogue (Huang et al., 2023; Liu et al., 2023a, 2020). However, a large volume of dialogue can result in an overwhelming amount of expanded personality information for the model. This poses challenges for the input size of generative models and **may lead the model to focus on personas that are inconsistent with the dialogue context** due to excessive persona information. As illustrated in "Turn 46(c)" in Figure 1. Although the model can generate traveler-related attributes like "where the winter scenery is beautiful" from the expanded personas, it overlooks the statement "I don't like going out in winter" from "Chatbot Personas" resulting in responses that are inconsistent with the personas.

In this paper, we utilize the large language model (LLM) as the dialogue generator. To address the aforementioned challenges, we propose a Persona-Aware LLM-enhAnCEd(PALACE) framework for multi-session personalized dialogue generation. Specifically, our framework consists of three main components: topic-aware memory bank, persona prompt learning, and VAE-LoRA (see the framework in Figure 2). The **topic-aware memory bank** aims to obtain both relevant history and dialogue structural information simultaneously. It retrieves information relevant to the current query topic while maintaining the dialogue structure from long-term dialogue history. We also introduce a topic detector to obtain information consistent with the current query topic in the short-term history while preserving the original dialogue structure. The **persona prompt learning module** aims to enhance the LLM's persona-aware capabilities and extract deeper personality information relevant to the current dialogue context. It first constructs a unique persona graph for each person in the dialogue with a knowledge graph and a triples extractor, and then designs a query-driven graph neural network and a persona prompt learning mechanism to lead the model to focus on character attributes consistent with the current dialogue context. Therefore, it can utilize relevant knowledge to extract deeper personality information. Finally, to further enhance the generative capabilities of the LLM and incorporate useful prior knowledge, we propose **VAE-LoRA**. VAE-LoRA injects hidden dialogue information from the query and golden response during the training process and introduces additional prior knowledge by maximizing mutual information. Armed with these three components, PALACE can generate the ground truth response shown in Figure 1.

We conduct experiments on MSC dataset (Xu et al., 2021) and DuLeMon dataset(Xu et al., 2022). Our framework consistently outperforms the compared baselines across various backbone LLMs on both automatic evaluation and human evaluation. Furthermore, ablation studies confirmed that components of topic-aware memory bank, persona prompt learning, and VAE-LoRA contribute to the performance improvement of our framework on the MSC dataset and the DuLeMon dataset.

Our contributions can be summarized as follows:

- We propose a Persona-Aware LLM-enhAnCEd framework for multi-session personalized dialogue generation(PALACE), which effectively enables the LLMs to generate responses consistent with the dialogue history and personality information.
- We introduce the persona prompt learning method that enhances the LLM's persona-aware capabilities while alleviating the issue of personality sparsity in dialogue.
- We propose VAE-LoRA, which effectively provides the LLM with useful prior knowledge to enhance its dialogue generation capabilities.
- Experimental results on the MSC dataset and the DuLeMon dataset demonstrate that our framework consistently outperforms the state-of-the-art baseline in both automatic evaluation metrics and human evaluation metrics.

2 Methodology

We present detailed descriptions of our framework in this section. As shown in Figure 2, the **topic-aware memory bank** is presented to retrieve history relevant to the current query's topic while preserving the dialogue structural information (Section 2.2). **Persona prompt learning** is designed to effectively alleviate the persona sparsity problem and uncover deeper personality information (Section 2.3). Moreover, **VAE-LoRA** is proposed to effectively incorporate latent prior knowledge from dialogues into the LLM (Section 2.4).

2.1 Problem Formalization

The goal of multi-session personalized dialogue generation is to generate the response r consistent with context C and the provided personality information P. Formally, we denote the dataset D by a list of N dialogues in format (C, P, r). Context C consists of the short-term dialogue history from the current session X and the long-term dialogue history from previous sessions H. Here, $X = \{q_1, r_1, ..., q_t\}$, q_i and r_i represent the query from the user and the response from the chatbot in the i-th round of dialogue in the current session respectively, and q_t is the query waiting for a response from the chatbot in the current round of dialogue. $H = \{H_1, H_2, ..., H_M\}$ denotes M dialogue sessions where $H_i = \{h_1^i, h_2^i, ..., h_{n_i}^i\}$ indicates that there are n_i utterances in the i-th dialogue session. $P = \{P_1, P_2, ..., P_T\}$ denotes T persona sentences. r is the golden truth response to q_t. The generation of our method can be formulated as

$$r_t = LM_\Theta(C, P), \tag{1}$$

where LM is the language model and Θ is the learnable parameters.

2.2 Topic-aware Memory Bank

To retrieve memories that are more relevant to semantic and topic information from dialogue history while preserving the original dialogue structure to the greatest extent, we propose the topic-aware memory bank. In our task, the dialogue history consists of the long-term and the short-term dialogue history. For long-term dialogue history, we use the DPR model (Karpukhin et al., 2020) to retrieve the top-k relevant histories based on similarity. Following Cheng et al. (2024), we calculate Conversation Edit Distance (Lavi et al., 2021) between utterances in retrieved histories. We re-rank them based on this score to preserve as much structural information in the dialogue as possible.

For short-term dialogue history, we introduce a benchmark for topic-shift aware dialog modeling named by TIAGE (Xie et al., 2021). We trained a topic shift detector on this dataset to assess short-term dialogue history. If an utterance in the short-term dialogue is unrelated to the current query's topic, it is discarded. This approach preserves the structural information of the original dialogue sequence while filtering out irrelevant content.

Overall, our memory bank retrieves relevant history for LLMs based on semantic and thematic relevance while maintaining the dialogue structure. Compared to existing methods, this approach enhances LLM's understanding of the conversational context through structural information and allows it to focus on more pertinent semantic information.

2.3 Persona Prompt Learning

To address personality sparsity, we introduce a persona commonsense knowledge graph PeaCoK (Gao et al., 2023) to expand the personality information with necessary commonsense knowledge. To aggregate and extract deeper personality representations to enhance the persona-aware capabilities of the LLMs while filtering out inconsistent noise present in the knowledge graph and avoiding the issue of excessively long input caused by providing all persona sentences or triples, we propose

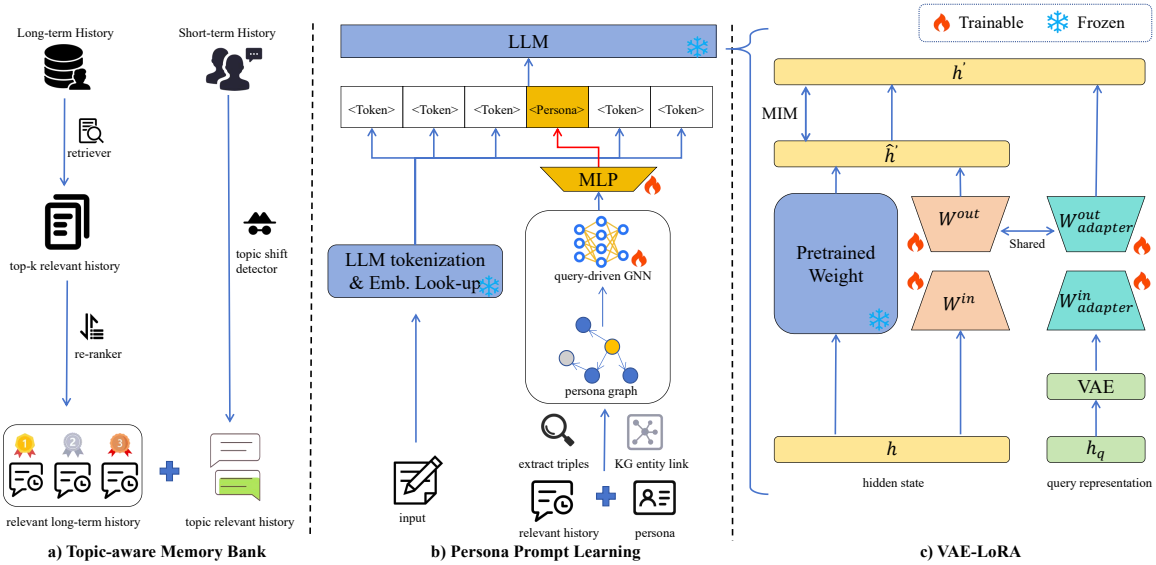


Figure 2: The overall framework of our model, comprising three components: topic-aware memory bank, persona prompt learning, and VAE-LoRA. “Emb." denotes “Embedding", “MIM" stands for “Mutual Information Maximization"

persona prompt learning, which stores all relevant personalized information in vector form. As shown in Figure 2(b), persona prompt learning consists of a persona graph and a query-driven GNN. The persona graph includes two steps: persona triples extraction and persona graph construction.

Persona triples extraction. For personalized information, unstructured text often fails to capture an individual's personality traits fully and typically contains numerous meaningless words, making it challenging to summarize the relationships between different persona attributes. We first transform the explicit personas P and the implicit personas present in C from the dataset into a unified structured format. Following Li et al. (2023), we train a model (RoBERTa-large (Liu et al., 2019)) on DNLI dataset (Welleck et al., 2019) to extract persona triples in dialogues where the format of triples is $(e1, r, e2)$. $e1$ and $e2$ represent the head entity and tail entity, respectively, which include the persona subject and persona attributes. r is the relationship between persona subject(e.g. I) and persona attributes derived from the 61 relationship categories defined in the DNLI dataset. Similar to the format of the DNLI dataset, the data provided by PeaCoK is also in the form of triples $(e1, r, e2)$. In contrast, the head entity $e1$ also includes persona sentences(e.g. *I'm a freelance programmer*), and the relationships r from the knowledge graph are categorized into eight types. More details about persona triples can be found in Appendix C.

Persona graph construction. To accurately link the head, tail entities and relationships of triples in the knowledge graph simultaneously, we convert the triples into their corresponding text forms during the entity linking process and utilize the DPR model (Karpukhin et al., 2020) for matching. For example, triples from knowledge graph *“(i am a bass player, characteristic, enjoys music)”* is converted into *“I am a bass player here is my character trait enjoys music”*. Based on triples extracted in dialogues and linked triples from the knowledge graph, we construct a persona graph for each person in dialogues. Formally, persona graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, $\mathcal{V}$ is the vertex of the graph including persona sentences and persona attributes and $\mathcal{E}$ is the edge of graph including relationships between persona sentences and persona attributes.

Query-driven GNN. Due to the numerous persona attributes contained within the persona graph, although they are related to the current persona, they do not necessarily aid in responding to the current query. For training, We modify the message-passing process based on existing GNNs. First, followingd Li et al. (2020) We initialize the representations of the nodes in the graph using the average of all token vectors from the first and last layers of the language model. We incorporate a query-driven attention mechanism into the graph neural network and the representation of the i-th

node at the $(l+1)$-th layer is given by:

$$h_i^{(l+1)} = \sigma(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\alpha_{i,j,r}^l W_r^l h_j^{(l)} + W_0^{(l)} h_i^{(l)}), \tag{2}$$

$$Q^l = tanh(H_q W_q^l), \tag{3}$$

$$\alpha_{i,j,r}^l = \frac{exp(Q^l \cdot (W_k^l h_j^l))}{\sum_{v\in\mathcal{N}_i^r} exp(Q^l \cdot (W_k^l h_v^l))}, \tag{4}$$

where $\mathcal{N}_i^r$ denotes the set of neighbor indices of node i under relation $r \in \mathcal{R}$, $\alpha_{i,j,r}^l$ is the attention score between node i and neighbor j under relation r, H_q is the representation of query in current utterance and it is initialized in the same way as the nodes in the graph. W_r^l, W_0^l, W_q^l and W_k^l are learnable parameters. After aggregating the representations of all nodes in the graph, we utilize a pooling operation to obtain the persona prompt.

Formally, the tokenization result of the input sentence is denoted as $S = [t_1, t_2, ..., u, ..., t_K]$ where t_i is the i-th token in input sentence and u is the special token to be replaced by persona prompt, denoted as "<Persona>". Due to the gap between the vector space encoded by GNNs and the semantic space of LLMs, we employ a multilayer perceptron (MLP) to map the persona prompt into the semantic space of LLMs. We then encode the prompt into a sequence of embeddings E:

$$E = [e_{t_1}, e_{t_2}, ...e_u, ..., e_{t_K}], \tag{5}$$

$$e_u = MLP(f_\phi(\mathcal{G}; q)), \tag{6}$$

where $e_{t_i}, e_u \in \mathcal{R}^{1\times d}$ represent token embedding for token t_i and persona prompt respectively, d is the dimension of language model, f_ϕ corresponds to the GNN described in equation 2 to equation 4.

2.4 VAE-LoRA

To improve the language model's generation ability by providing more effective prior knowledge, while ensuring its performance in downstream tasks, we creatively integrate the concepts of Variational Autoencoders (VAE) (Kingma and Welling, 2022) and Low-Rank Adaptation (LoRA) (Hu et al., 2021) named by VAE-LoRA, with the specific framework illustrated in the Figure 2(c).

VAE-LoRA necessitates not only the hidden state but also the query and response. The introduction of the query aims to extract latent themes, tone, emotions, and other relevant information through the VAE module. It also compels LLMs to focus on the content that needs to be addressed, thereby balancing attention to some extent. The inclusion of the response is intended to provide relevant prior knowledge, enabling LLMs to generate more meaningful responses.

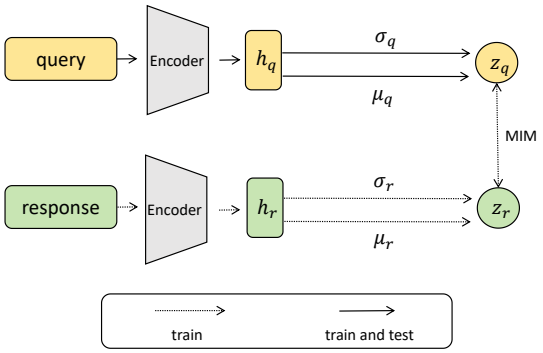


Figure 3: VAE module in VAE-LoRA. The dashed lines represent the training phase, while the solid lines indicate both the training and testing phases. "MIM" stands for "Mutual Information Maximization".

VAE-LoRA takes hidden states h, query q, and golden response r as input. We first utilize encoders such as BERT (Devlin et al., 2019) to encode the query and golden response, obtaining their corresponding representations:

$$h_q = encoder(q); h_r = encoder(r). \tag{7}$$

As illustrated in Figure 3, to sample the latent vector of query and response, we use the reparameterization trick to make all the processes derivable:

$$z_q = \mu_q + \sigma_q \times \epsilon_q, \epsilon_q \sim \mathcal{N}(0, I), \tag{8}$$

$$z_r = \mu_r + \sigma_r \times \epsilon_r, \epsilon_r \sim \mathcal{N}(0, I), \tag{9}$$

where $\mu_q, \sigma_q, \mu_r, \sigma_r$ are fitted by prior networks and posterior networks with h_q and h_r respectively. We use KL distance to approximate them. The loss of the VAE module is defined as follows:

$$\begin{aligned}\mathcal{L}_{VAE} = \alpha KL(q_\phi(z_r|q,r)||p_\phi(z_q|q)) + \\ \beta KL(\mathcal{N}(\mu_q, \sigma_q^2 I)||\mathcal{N}(0,1)),\end{aligned} \tag{10}$$

where $q_\phi(z_r|q,r)$ and $p_\phi(z_q|q)$ are approximate posterior distribution and approximate prior distribution respectively. α and β are hyperparameters. And we define output h' of VAE-LoRA as follows:

$$\begin{aligned}h' &= hW + hW^{in}W^{out} + z_q W_{adapter}^{in} W_{adapter}^{out} \\ &= hW + (hW^{in} + z_q W_{adapter}^{in}) W^{out},\end{aligned} \tag{11}$$

where W is the parameters of the original language model. W^{in} and W^{out} are weight matrices of the

original LoRA. $W^{in}_{adapter}$ and $W^{out}_{adapter}$ are the parameter matrix of the adapter designed to bridge the gap between the implicit vectors extracted by the VAE and the original model's vector space. Following Zhang et al. (2024), we share W^{out} with $W^{out}_{adapter}$ to better adapt to downstream tasks.

For training, it is essential not only to incorporate the response but also to ensure that the latent information in the query and in the response have mutual information maximization (MIM), thereby guaranteeing the consistency of themes, emotions, and other latent information. Moreover, inspired by Zhang et al. (2023), we align the distance between the task-specific and latent vectors from the VAE module with MIM. Task-specific representation $\hat{h'}$ in VAE-LoRA is formulated as:

$$\hat{h'} = hW + hW^{in}W^{out}. \tag{12}$$

While for inference, only the hidden state and the query need to be passed in.

The total loss during training of VAE-LoRA is defined as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{GEN} + \mathcal{L}_{VAE} + \gamma\mathcal{L}_{MIM} \\ &= -E_{q_\phi(z_r|q,r)}[log(LM_\Phi(r|z,C,P))] \\ &\quad + \alpha KL(q_\phi(z_r|q,r)||p_\phi(z_q|q)) \\ &\quad + \beta KL(\mathcal{N}(\mu_q,\sigma_q^2 I)||\mathcal{N}(0,1)) \\ &\quad + \gamma \sum_{\forall(\hat{h'},h')} MIM(\hat{h'},h'),\end{aligned} \tag{13}$$

where $\mathcal{L}_{GEN}$ is the generation loss of the language model, $\mathcal{L}_{MIM}$ is the loss between task-specific representation $\hat{h'}$ in VAE-LoRA and latent vectors h' from the VAE module, MIM is the method to get maximum mutual information, in practice, both mean square error and KL divergence can be applied and γ is the hyperparameter.

3 Experimental Settings

Due to space limitations, additional details regarding the experimental setup (such as dataset statistics, baseline descriptions, human evaluation details, and parameter settings.) can be found in the Appendix D.

Dataset. We conduct experiments on the Multi-Session Chat (MSC) dataset(Xu et al., 2021), the largest high-quality English dataset of multi-session personalized dialogues. We also conduct experiments on the large-scale Chinese dataset of multi-session personalized dialogues DuLeMon (Xu et al., 2022) to demonstrate the generalizability of our method, further details can be found in the Appendix F due to space limitations.

Metrics. Following previous works (Li et al., 2024b,a), we adopt BLEU-1/2/3, ROUGE-L, BERTScore (Zhang et al., 2020), G-Eval (Liu et al., 2023c)for automatic evaluation. For human evaluation metrics, we evaluate the generated responses for fluency, consistency, sensibleness, and engagingness, assessing whether they are fluent, consistent with the dialogue history and persona information, meaningful, and capable of engaging the user to encourage continued conversation.

Baselines. We introduce two categories of baselines, pre-trained-based approaches including BlenderBot (Roller et al., 2021) and HAHT (Zhang et al., 2022) and LLM-based approaches including vanilla LLM, ChatGPT, CPD (Fan et al., 2024), StreamingDialogue (Li et al., 2024b) and LD-Agent (Li et al., 2024a) to compare the performance of our method.

4 Results and analysis

We present the experimental results of the automatic evaluation, human evaluation, and ablation study in this section. More experimental results can be found in Appendix B (case study), F (experiments on DuLeMon), E (ablation study on different sessions), G(analysis on VAE-LoRA), H (analysis on dialogue structure) I (evaluation of different hidden states in the GNN) and J (Performance validation of Query-driven GNN). Furthermore, due to space constraints in the figures and the consistent trends between BERTScore and G-Eval across all experiments, G-Eval is omitted from some of the tables.

4.1 Automatic Evaluation

The automatic evaluation results of different models on the MSC dataset can be found in Table 1. The results indicate that (1) our framework outperforms baselines on all automatic evaluation metrics across different backbones, demonstrating its effectiveness. (2) LLM-based models outperform pre-trained models, with only a small gap between base LLMs and fine-tuned models, demonstrating the significant potential of LLMs for this task. (3) Comparing StreamingDialogue with other methods, it is evident that the context window extension method has significantly lower R-L scores than the

Table 1: Automatic evaluation results of different models on MSC dataset. B-1, B-2, B-3, R-L, BS denote the average BLEU-1, BLEU-2, BLEU-3, ROUGE-L and BertScore scores across all sessions on the MSC respectively. The best results are in **bold** and the second-best results are in underlined.

Model	B-1	B-2	B-3	R-L	BS	G-Eval
Pre-trained model						
BlenderBot	-	4.91	1.53	16.05	-	-
HAHT	-	5.10	1.59	16.58	-	-
LLM-based model						
ChatGLM	19.20	5.54	1.50	16.49	48.60	3.25
Llama2	17.34	4.37	1.21	10.29	45.25	3.11
ChatGPT	18.96	5.77	1.51	16.84	50.47	3.70
CPD	12.45	4.41	-	12.14	-	-
StreamingDialogue	19.33	-	-	15.86	-	-
LD-Agent	19.54	7.31	2.51	18.44	52.36	3.84
PALACE (ChatGLM)	20.17	7.92	**2.73**	**20.27**	**54.68**	**3.97**
PALACE (Llama2)	**21.04**	**8.31**	2.70	18.98	54.09	3.94

retrieval-based methods. As it introduces excessive noise by providing all dialogue context, making it difficult for the LLM to discern relevant information and recall the correct answers. (4) Our method achieves the greatest improvement in R-L scores, indicating its ability to effectively recall relevant information from lengthy dialogue histories and abundant personality information. The improvement in B-3 is relatively low, as in open-domain dialogue, the range of model responses is quite broad. Additionally, our framework expands personality information and incorporates more prior knowledge, making complete alignment with the gold responses challenging.

4.2 Evalutaion on different sessions

To further validate the feasibility of our framework with varying dialogue history scales, we present the experimental results on various sessions on the MSC, as shown in Table 2. The data from the first session lacks effective dialogue history, and the personality information is relatively subtle, which does not align well with the characteristics of multi-session personalized dialogue generation. Therefore, we evaluate our method with the last four sessions of the dataset, specifically sessions 2 to 5.

From the results, we can draw the following conclusions: (1) Under different scales of dialogue history, our framework shows significant improvements across all metrics compared to baselines. (2) Methods that are neither fine-tuned nor attentive to dialogue structure exhibit high sensitivity to dialogue history. For example, ChatGLM, Llama2, and LD-Agent show a decline in performance as the scale of dialogue history increases. (3) It can be observed that within the range of sessions 2 to 4, our method exhibits a steady performance increase with the addition of sessions, while the evaluation results for session 5 show a decline compared to those of sessions 2 to 4. This indicates our method effectively retrieves historical information relevant to the current query while integrating additional persona information and prior knowledge from these histories. As the dialogue history increases, the amount of useful information utilized by our framework also increases. However, this capability has limits, when dialogue interactions exceed 60 turns, lengthy histories can introduce noise that adversely affects model performance.

4.3 Human Evaluation

We measure the inter-rater reliability with Fleiss' Kappa (Fleiss and Cohen, 1973). Our annotations obtain "good agreement" for Fluency (0.573) and Consistency (0.526) and "moderate agreement" for Sensibleness (0.611) and Engagingness (0.693) among 6 annotators. Table 3 presents the results of human evaluations for different models on the MSC dataset. From the results, we can observe that (1) our framework outperforms the compared baselines on all human evaluation metrics, with Llama2 demonstrating improved performance after being trained with our framework. (2) Our method shows the most significant improvements in consistency and sensibleness, indicating that it effectively enables LLMs to focus on relevant dialogue history and personality information. In contrast, the improvement in fluency is the least pronounced, as LLMs inherently possess strong dialogue generation capabilities, enabling them to produce fluent responses. (3) Our method achieves the highest scores in engagingness, signifying that our model can effectively expand personality information and enrich the generated content, leading users to have a strong willingness to continue the conversation.

4.4 Ablation Study

To demonstrate the effectiveness and generality of our proposed method, we design ablation experiments for different modules within the framework using two distinct backbone models. We conduct ablation experiments on the three key modules: topic memory bank, persona prompt learning, and VAE-LoRA, with the results presented in Table 4.

From the results of the ablation experiments, we observe that (1) all modules contribute positively

Table 2: Automatic evaluation results of different models on MSC dataset. B-1, B-2, B-3, R-L, BS denote BLEU-1 BLEU-2, BLEU-3, ROUGE-L and BertScore respectively. The best results are in **bold** and the second-best results are in underlined.

Model	Session 2					Session 3					Session 4					Session 5				
	B-1	B-2	B-3	R-L	BS	B-1	B-2	B-3	R-L	BS	B-1	B-2	B-3	R-L	BS	B-1	B-2	B-3	R-L	BS
pre-trained model																				
BlenderBot	-	4.76	1.51	16.18	-	-	5.03	1.61	16.39	-	-	4.78	1.49	15.56	-	-	4.98	1.48	16.10	-
HAHT	-	5.07	1.57	16.90	-	-	5.27	1.67	16.72	-	-	5.00	1.55	15.97	-	-	5.16	1.60	16.42	-
LLM-based model																				
ChatGLM	19.29	5.44	1.49	16.76	48.77	19.21	5.18	1.55	15.51	48.63	19.15	5.74	1.52	16.68	48.55	19.02	5.92	1.45	16.63	48.07
Llama2	17.39	4.47	1.21	10.43	45.02	17.34	4.39	1.22	10.33	45.48	17.32	4.32	1.21	10.25	45.28	17.16	4.20	1.17	10.03	44.97
ChatGPT	19.29	5.85	1.50	16.83	50.60	18.96	5.74	1.45	16.61	50.52	18.79	5.62	1.43	16.76	50.48	18.35	5.63	1.62	17.00	49.74
StreamingDialogue	18.33	-	-	13.53	-	19.27	-	-	15.67	-	19.33	-	-	15.86	-	19.16	-	-	15.21	-
LD-Agent	19.51	7.38	2.63	18.84	52.97	19.30	7.40	2.57	18.31	52.46	19.25	7.16	2.31	18.08	52.05	19.10	7.11	2.31	17.68	51.32
PALACE (ChatGLM)	20.10	7.91	**2.68**	**20.28**	**54.68**	20.40	7.93	2.67	**20.51**	**54.80**	20.69	8.24	**2.93**	**20.94**	54.47	19.78	7.65	**2.67**	**19.33**	**54.46**
PALACE (Llama2)	**20.95**	**8.13**	2.66	18.96	53.25	**20.87**	**8.29**	**2.70**	19.22	53.63	**21.45**	**8.64**	2.77	19.32	**55.28**	**21.29**	**8.54**	2.66	18.83	53.28

Table 3: Human evaluation results of different models on MSC dataset. Flu., Con., Sen., Eng. denote Fluency, Consistency, Sensibleness, and Engagingness respectively. The best results are in **bold** and the second-best results are in underlined.

Model	Flu.	Con.	Sen.	Eng.
ChatGLM	3.50	3.00	3.13	3.31
Llama2	3.47	3.02	3.20	3.29
ChatGPT	3.80	3.21	3.17	3.66
LD-Agent	3.74	3.29	3.38	3.73
PALACE (ChatGLM)	**3.95**	3.61	3.86	3.88
PALACE (Llama2)	3.86	**3.93**	**3.94**	**3.97**

Table 4: Ablation experiments on the MSC dataset. TMB stands for Topic-Aware Memory Bank, PPL refers to Persona Prompt Learning. B-1/2/3, R-L, BS denote the average BLEU-1/2/3, ROUGE-L and BERTScore scores across all sessions on the MSC, respectively. The best results for each backbone model are in **bold**.

Model	B-1	B-2	B-3	R-L	BS
ChatGLM (Base)	19.20	5.54	1.50	16.49	48.60
w/ TMB	19.46	7.45	2.37	19.36	52.33
w/ PPL	19.44	7.51	2.52	19.67	52.17
w/ VAE-LoRA	19.87	7.71	2.55	19.87	53.86
w/ TMB & PPL	19.95	7.78	2.59	20.08	54.27
w/ TMB & VAE-LoRA	20.04	7.86	2.64	20.23	54.41
w/ PPL & VAE-LoRA	19.93	7.75	2.57	19.99	54.29
PALACE (ChatGLM)	**20.20**	**7.93**	**2.73**	**20.27**	**54.68**
Model	**B-1**	**B-2**	**B-3**	**R-L**	**BS**
Llama2 (Base)	17.34	4.37	1.21	10.29	45.25
w/ TMB	17.92	4.90	1.51	12.83	46.63
w/ PPL	20.56	8.05	2.53	18.72	53.83
w/ VAE-LoRA	20.51	8.16	2.59	18.77	53.77
w/ TMB & PPL	20.70	8.25	2.62	18.96	53.96
w/ TMB & VAE-LoRA	20.67	8.28	2.64	18.98	53.78
w/ PPL & VAE-LoRA	20.56	8.20	2.60	18.88	53.81
PALACE (Llama2)	**21.06**	**8.35**	**2.70**	**19.01**	**54.09**

to the model's performance. (2) The impact of the topic-aware memory bank on performance is relatively low, as LLMs used in the ablation experiment for this module are not fine-tuned. In contrast, VAE-LoRA has the most significant impact on performance, as effective prior knowledge plays a crucial role in dialogue generation. (3) Efficient historical information enhances the fine-tuning potential of LLMs. On ChatGLM and LLaMA, w/ TMB & PPL and w/ TMB & VAE-LoRA outperform w/ PPL & VAE-LoRA. (4) w/ VAE-LoRA can capture useful prior information, such as personality information in dialogues. Comparison between w/ VAE-LoRA and w/ VAE-LoRA & PPL shows a slight improvement, indicating that VAE-LoRA has partially acquired relevant prior information. The ablation experimental results across different sessions can be found in Appendix E.

5 Related Work

To enhance the consistency and distinctiveness of dialogue responses, multi-session personalized dialogue generation has been proposed (Xu et al., 2021, 2022). Existing methods primarily focus on history modeling and persona modeling. We review classic works from these two categories of methods in the introduction, a comprehensive discussion of related work is available in Appendix A.

6 Conclusion

In this paper, we propose a persona-aware LLM-enhanced framework for multi-session personalized dialogue generation including three components, topic-aware memory bank, persona prompt learning, and VAE-LoRA. Topic-aware memory bank retrieves history while preserving dialogue structure. To enhance LLM's persona-aware capabilities, persona prompt learning. VAE-LoRA is

employed to obtain more useful prior knowledge, enhancing the generative capabilities of LLMs. Experimental results on two datasets show our framework outperforms the state-of-the-art methods in automatic and human evaluation metrics.

7 Limitations

One limitation is that we conducted experiments solely on the MSC and DuLeMon datasets. However, to the best of our knowledge, these are currently the only two datasets available for multi-session personalized dialogue generation. In the future, we plan to construct more datasets for multi-turn personalized dialogues in English and test our framework with them.

8 Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62302040), the China Postdoctoral Science Foundation (No. 2022TQ0033), and the Beijing Institute of Technology Research Fund Program for Young Scholars.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. Plato-2: Towards building an open-domain chatbot via curriculum learning. *Preprint*, arXiv:2006.16779.

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

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665, Torino, Italia. ELRA and ICCL.

Chuanqi Cheng, Quan Tu, Wei Wu, Shuo Shang, Cunli Mao, Zhengtao Yu, and Rui Yan. 2024. "in dialogues we learn": Towards personalized dialogue without pre-defined profiles through in-dialogue learning. *Preprint*, arXiv:2403.03102.

Eric Chu, Prashanth Vijayaraghavan, and Deb Roy. 2018. Learning personas from dialogue with attentive memory networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2638–2646, Brussels, Belgium. Association for Computational Linguistics.

Alexandra DeLucia, Mengjie Zhao, Yoshinori Maeda, Makoto Yoda, Keiichi Yamada, and Hiromi Wakaki. 2024. Using natural language inference to improve persona extraction from dialogue in a new domain. *Preprint*, arXiv:2401.06742.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason D. Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098.

Shixuan Fan, Wei Wei, Wendi Li, Xian-Ling Mao, Wenfeng Xie, and Dangyang Chen. 2024. Position debiasing fine-tuning for causal perception in long-term dialogue. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6261–6269. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613 – 619.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023. PeaCoK: Persona commonsense knowledge for consistent and engaging narratives. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6569–6591, Toronto, Canada. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *Preprint*, arXiv:1905.05709.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and H Tang. 2023. Personalized dialogue generation with persona-adaptive attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12916–12923.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. *Preprint*, arXiv:2112.08619.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. Persona expansion with commonsense knowledge for diverse and consistent response generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1139–1149, Dubrovnik, Croatia. Association for Computational Linguistics.

Hana Kim, Kai Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. 2024a. Commonsense-augmented memory construction and management in long-term conversations via context-aware persona refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 104–123, St. Julian's, Malta. Association for Computational Linguistics.

Minju Kim, Beong woo Kwak, Youngwook Kim, Hong in Lee, Seung won Hwang, and Jinyoung Yeo. 2022. Dual task framework for improving persona-grounded dialogue dataset. *Preprint*, arXiv:2202.05435.

Seo Hyun Kim, Keummin Ka, Yohan Jo, Seung won Hwang, Dongha Lee, and Jinyoung Yeo. 2024b. Ever-evolving memory by blending and refining the past. *Preprint*, arXiv:2403.04787.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Diederik P Kingma and Max Welling. 2022. Auto-encoding variational bayes. *Preprint*, arXiv:1312.6114.

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

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1169–1177, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023a. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Joosung Lee, Minsik Oh, and Donghun Lee. 2023b. P5: Plug-and-play persona prompting for personalized response selection. *Preprint*, arXiv:2310.06390.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024a. Hello again! llm-powered personalized agent for long-term dialogue. *Preprint*, arXiv:2406.05925.

Jia-Nan Li, Quan Tu, Cunli Mao, Zhengtao Yu, Ji-Rong Wen, and Rui Yan. 2024b. Streamingdialogue: Prolonged dialogue learning via long context compression with minimal losses. *Preprint*, arXiv:2403.08312.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *Preprint*, arXiv:1603.06155.

Yunpeng Li, Yue Hu, Yajing Sun, Luxi Xing, Ping Guo, Yuqiang Xie, and Wei Peng. 2023. Learning to know myself: A coarse-to-fine persona-aware training framework for personalized dialogue generation. In *AAAI Conference on Artificial Intelligence*.

Jungwoo Lim, Myunghoon Kang, Yuna Hur, Seungwon Jung, Jinsung Kim, Yoonna Jang, Dongyub Lee, Hyesung Ji, Donghoon Shin, Seungryong Kim,

and Heuiseok Lim. 2023. You truly understand what i need: Intellectual and friendly dialogue agents grounding knowledge and persona. *Preprint*, arXiv:2301.02401.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Annual Meeting of the Association for Computational Linguistics*.

Pingsheng Liu, Zhengjie Huang, Xiechi Zhang, Linlin Wang, Gerard de Melo, Xin Lin, Liang Pang, and Liang He. 2023a. A disentangled-attention based framework with persona-aware prompt learning for dialogue generation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Shuai Liu, Hyundong J. Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023b. Recap: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. *Preprint*, arXiv:2306.07206.

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

Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. Improving personality consistency in conversation by persona extending. volume 39 of *CIKM '22*, page 1350–1359. ACM.

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

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *Preprint*, arXiv:1706.02861.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. TVShowGuess: Character comprehension in stories as speaker guessing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4267–4287, Seattle, United States. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *Preprint*, arXiv:2208.03188.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

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

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8878–8885.

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023a. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. *Preprint*, arXiv:2305.11482.

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023b.

Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5456–5468, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2023. Focused transformer: Contrastive training for context scaling. *Preprint*, arXiv:2307.03170.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024a. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *CoRR*, abs/2401.13256.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2024b. Recursively summarizing enables long-term dialogue memory in large language models. *Preprint*, arXiv:2308.15022.

Zhilin Wang, Xuhui Zhou, Rik Koncel-Kedziorski, Alex Marin, and Fei Xia. 2022. Extracting and inferring personal attributes from dialogue. *Preprint*, arXiv:2109.12702.

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

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

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. TIAGE: A benchmark for topic-shift aware dialog modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1684–1690, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *Preprint*, arXiv:2107.07567.

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

Congchi Yin, Piji Li, and Zhaochun Ren. 2023. Ctrl-struct: Dialogue structure learning for open-domain response generation. In *Proceedings of the ACM Web Conference 2023*, volume 995, page 1539–1550. ACM.

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? *Preprint*, arXiv:1801.07243.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Tong Zhang, Yong Liu, Boyang Li, Zhiwei Zeng, Pengwei Wang, Yuan You, Chunyan Miao, and Lizhen Cui. 2022. History-aware hierarchical transformer for multi-session open-domain dialogue system. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3395–3407, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2023. Domain generalization via switch knowledge distillation for robust review representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12812–12826, Toronto, Canada. Association for Computational Linguistics.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024. Personalized lora for human-centered text understanding. *Preprint*, arXiv:2403.06208.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2019. A pre-training based personalized dialogue generation model with persona-sparse data. *Preprint*, arXiv:1911.04700.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States. Association for Computational Linguistics.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. Memorybank: Enhancing large language models with long-term memory. *Preprint*, arXiv:2305.10250.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *Preprint*, arXiv:2311.16832.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2021. Learning to predict persona information for dialogue personalization without explicit persona description. *Preprint*, arXiv:2111.15093.

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

A Related Work

History modeling can be categorized into memory mechanism and context window expansion. An efficient memory mechanism has a significant positive impact on understanding dialogue history. Early memory networks (Chu et al., 2018; Shuster et al., 2022) were designed to enhance representational capacity for better context modeling. With the continuous breakthroughs of LLMs in dialogue generation, some works have attempted to transfer the concepts of memory mechanisms, resulting in the categorization into retrieval-based methods and summary-based methods. Retrieval-based methods primarily involve training a retriever to extract relevant historical information from the memory bank (Kim et al., 2024b; Lee et al., 2023a). While, summary-based methods typically provide prompts to summarize pertinent information from the memory bank (Zhong et al., 2023; Wang et al., 2024b). Although both approaches effectively enable the model to access long-term historical details, they disrupt the dialogue structure and are prone to losing critical information. Expanding the context window is primarily achieved by modifying the internal structure of the model or the attention computation mechanism (Li et al., 2024b; Tworkowski et al., 2023). This approach allows the model to access all historical information with a dialogue structure but also introduces significant noise.

To enrich the personality information and better model human reasoning abilities in dialogues, persona modeling aims to expand the sparse personality information in multi-session personalized dialogue (Wu et al., 2021; Qian et al., 2017; Song et al., 2021, 2019; Kim et al., 2022). Existing expansion methods primarily involve the introduction of external knowledge bases (Jang et al., 2022; Ma et al., 2021) or the extraction of new personality information (DeLucia et al., 2024; Lee et al., 2023b; Tang et al., 2023b). Commonsense knowledge graph is one of the most important external knowledge bases(Kim et al., 2023, 2024a; Majumder et al., 2020). PeaCoK (Gao et al., 2023) is a large persona commonsense knowledge graph that contains approximately 100K manually verified character facts. PeaCoK knowledge graph provides rich persona commonsense reasoning for downstream systems, aiding in generating more consistent and engaging narratives. Commonsense knowledge can effectively enhance a model's ability to reason about personality, how to filter relevant knowledge remains a key challenge. To effectively explore potential persona information within dialogues, many studies have applied persona extraction to construct or expand a persona's knowledge grap-h (Zhong et al., 2022). Persona extraction is similar to traditional extraction tasks, typically involving the extraction of persona triples or related information present in dialogue or personality data (Zhu et al., 2023; DeLucia et al., 2024). Previous work has introduced various generative models to achieve persona extraction (Sang et al., 2022; Wang et al., 2022; Zhu et al., 2023). Additionally, some studies have attempted to incorporate Welleck et al. (2019) to ensure the consistency of extraction results through natural language inference (Wang et al., 2022; Song et al., 2020).

In this paper, our work not only retrieves relevant historical information for the model but also effectively preserves the original dialogue structure of the dialogue history, enhancing the model's understanding of the dialogue context. Furthermore, unlike existing methods that explicitly expand personas, our approach utilizes prompt learning to store the expanded character information in im-

plicit vectors and employs graph neural networks to obtain character attributes consistent with the current dialogue context. Finally, unlike current fine-tuning methods for LLMs on dialogue data, we propose VAE-LoRA, which provides effective prior knowledge to the LLM and demonstrates its efficacy. Our method not only focuses on history modeling but also emphasizes personalized representations. Therefore, we selected multi-session personalized dialogue datasets such as MSC and DuLeMon, rather than a single multi-session dialogue dataset.

B Case Study

Table 5 presents an example of the generated results from different models on the MSC dataset. Compared to the results of the previous start-of-the-art method, the LLMs trained using our proposed framework not only understand and utilize relevant information from the dialog history but also accurately present the pertinent personality information. From the examples in the table, we can observe that when the user asks for running advice, our method recalls relevant historical information and incorporates it into the generated response. For instance, *"a short walk"* and *"slowly add"* correspond to *"start out slowly"* and *"a short walk"* from the previous dialog history. Moreover, our method not only generates accurate and relevant personality information but also produces pertinent persona knowledge. *"Protein"* corresponds to *"I eat a protein-heavy diet"*, while *"5 miles"* corresponds to *"I could run 5 miles"*. *"Give you more motivation"* and *"boost your energy"* serve as expansions of the personality information related to *"protein-heavy diet"* and *"Regular runners often warm up"* expands on the personality information associated with *"I could run 5 miles"*.

C Persona triples extractor

Following Li et al. (2023), we also train a RoBERTa model[†] on DNLI to implement a persona triples extractor, where the extracted triples are in the form of $(e1, r, e2)$. Since we need to construct a unique user graph for each person in every dataset, $e1$ is typically in the first person, while triples that are not in the first person are ignored to prevent introducing noise. The relation r denotes the relationship between $e1$ and $e2$. In the DNLI dataset, some relations are in the form of $r1\_r2$, where $r1$ is a verb and $r2$ is a noun. In our work, we select 46 relations for training based on their frequency of occurrence in the DNLI dataset, resulting in the following set of relations: { *place_origin, live_in_citystatecountry, live_in_general, nationality, employed_by_company, employed_by_general,*

[†]The fine-tuned RoBERTa-large model achieves 79.8% micro-F1 on DNLI.

Table 5: A case study of the MSC dataset. Historical relevant content is in **bold**, personality-related information is in **bold and underlined**, and valid personality expansion information is highlighted in red.

Chatbot's Personas	
1. I try to run five miles per day.	
2. It took me 3 months of training before I could run 5 miles	
3. I eat a protein-heavy diet	
...	
History Sessions	
[Session 1]	
...	
User:	I wish I could exercise more, I've got the kids running me around all day.
Chatbot:	How is your diet? I try to stay with meat and **proteins**.
...	
[Session 2]	
...	
User:	...Do you have any advice on how to start a running routine?
Chatbot:	**Start out slowly**, if you are not used to exercising, the best thing to do is just **walking**
...	
Current Session	
...	
User:	...How long did it take you to get up to that long of a run?
Chatbot:	We have been running **consistently** for about 7 months. I feel that consistency is the key to being successful in running. You should start **running a little when you go for a walk**.
User:	Maybe I will start! Any other tips? How long should we shart?
Responses from different models	
LD-Agent:	I think it is important to start with a short walk and then build up to longer walks. I would recommend starting with a 10 minutes walk.
PALACE(ChatGLM):	You can start with **a short walk** before running and gradually increase your running time. If you **stick with it**, you can run **5 miles** like me. On runn -ing days, eating some **high-protein** foods can give you more motivation.
PALACE(Llama2)	I would start with a 20 minute walk and then **slowly add** on a minute each day. I think that is the best way to get. Regular runners often warm up this way. You can also boost your energy by increasing your **protein** intake.

has_profession, previous_profession, job_status, teach, school_status, has_degree, attend_school, like_general, like_food, like_drink, like_animal, like_movie, like_music, like_read, like_sports, like_watching, like_ac-tivity, like_goto, dislike, has_hobby, has_ability, member_of, want_do, want_job, want, favorite, favorite_food, favorite_color, favorite_book, favorite_movie, favorite_music, favorite_music artist, favorite_activity, favorite_drink, favorite_show, favorite_place, favorite_hobby, favorite_season, favorite_animal, favorite_sport, own, have, have_pet, have_sibling, have_children, have_family }.

For the persona commonsense knowledge graph PeaCok, the dataset includes eight types of relations. However, not all relations contribute to the expansion of persona information in MSC. PeaCok contains many persona-specific rather than general attributes. For example, the triple (*"I am a famous pianist", "experience", "win a Grammy award"*) does not effectively expand the persona of a *"pianist"*, as not all pianists *"win a Grammy award"*. Therefore, to filter out this noise, we manually selected five general relations. They are *"characteristic", "characteristic_relationship", "routine_habit_relationship", "goal_plan_relationship", "experience_relationship"*.

D Detailed experimental settings

D.1 Dataset

We conduct our experiments on the Multi-Session Chat(MSC) dataset (Xu et al., 2021) which is the largest high-quality English dataset of long-term personalized dialogues so far. The dataset was collected by co-workers chatting according to specified scenarios and personality information. The personality information is provided as a series of sentences describing characteristics, events, and opinions. The training set of MSC contains 4 sessions and the test set comprises 5 sessions. Each session includes a maximum of 14 utterances, with intervals between conversations ranging from a few hours to several days. The statistics of the dataset are shown in Table 6. Because session 1 has no session dialogue history, we mainly evaluate our method in sessions 2-5.

D.2 Baselines

We introduce two categories of baselines, Pre-trained-based approaches, and LLM-based approaches to compare the performance of our model.

Table 6: The statistics of MSC dataset. We show the number of dialogues(#Dialog), and utterances(#Utts) on the train set, the valid set, and the test set for each session. Session number i indicates that i-1 history conversation sessions happened before the current session.

Session Number	Train		Vaild		Test	
	#Dialog	#Utts	#Dialog	#Utts	#Dialog	#Utts
Session 1	8,939	131,438	1,000	7,801	1,015	6,634
Session 2	4,000	46,420	500	5,897	501	5,939
Session 3	4,000	47,259	500	5,890	501	5,924
Session 4	1,001	11,870	500	5,904	501	5,940
Session 5	-	-	500	5,964	501	-

Pre-trained-based approaches:

- BlenderBot (Roller et al., 2021): BlenderBot is an advanced open-domain dialogue model developed through large-scale pre-trained on large-scale datasets with a retrieval-refinement mechanism and optimized decoding strategies.
- HAHT (Zhang et al., 2022): It is a model that employs hierarchical encoding and attention mechanisms to maintain and utilize long-term historical dialogue memory, generating contextually relevant responses through a history-aware response generator.

LLM-based approaches:

- Vanilla LLM: We directly employ the LLM as the chatbot where we concatenate dialogue history and personality information as prompt. We utilize ChatGLM-6B (GLM et al., 2024) and Llama2-7B-Chat(Touvron et al., 2023) as vanilla LLMs.
- ChatGPT: ChatGPT is a closed-source large language model based on the GPT architecture, and we utilize the API services of OpenAI's 'gpt-3.5-turbo' model.
- CPD (Fan et al., 2024): CPD is a causal perception multi-turn dialogue framework that employs a perturbation-based causal variable discovery method to extract statements with high causal relevance from historical dialogues, thereby enhancing the causal perception capabilities of LLMs.
- StreamingDialogue (Li et al., 2024b): It is a model that effectively handles long-context dialogues by compressing dialogue history into

"conversation attention sinks" and employing short-term memory reconstruction (SMR) and long-term memory activation (LMR) learning strategies.

- LD-Agent (Li et al., 2024a): LD-Agent is a dialogue agent framework that supports coherent dialogue by integrating event memory and personalized role modeling. It is the previous state-of-the-art method on MSC dataset.

D.3 Metrics

Following previous works (Li et al., 2024b,a; Wang et al., 2024b), we conduct automatic evaluation metrics and human evaluation metrics to measure the effectiveness of our method. For automatic evaluation metrics, we adopt BLEU-1, BLEU-2, BLEU-3 (Papineni et al., 2002) and ROUGE-L (Lin and Och, 2004) to measure word overlaps between the golden response and the generated response. We also use BERTScore(Zhang et al., 2020) and G-Eval(Liu et al., 2023c) to evaluate the semantic similarity and consistency between the golden response and generated response. For human evaluation metrics, following (Lee et al., 2023a; Zhang et al., 2022), we recruit 6 professional annotators to label the generated results. They are all experts in the fields of computer science or artificial intelligence, familiar with the task of multi-session personalized dialogue generation, and are under the age of 35. They score 12 randomly selected samples from each session, totaling 60 samples on fluency, consistency, sensibleness, and engagingness. Fluency measures whether the generated response is fluent and human-like. Consistency measures whether the generated response is consistent with dialogue history and personality information. Sensibleness measures whether the generated response makes sense. Engagingness measures whether the user is engaged and would want to continue the dialogue. All human evaluation scores range from 0 to 5.

D.4 Model Settings

In this work, we utilize ChatGLM-6B (GLM et al., 2024) and Llama2-7B-Chat (Touvron et al., 2023) as the backbone. The maximum input length is set to 2048. For graph neural networks, we set the number of layers to 2 and the hidden layer channels to 2048. For model training, we use the Adam-ax optimizer (Kingma and Ba, 2017) with a learning rate of 5e-5, batch size of 32, dropout ratio of 0.1, α of 0.5, β of 0.5 and γ of 1. All the fine-tuned models are trained with a maximum of two 48GB GPUs(NVIDIA A6000). In the training process, we first train the topic detector, followed by training the VAE-LoRA, and finally, we train the persona prompt learning. In the experiments, for the HAHT, CPD, and StreamingDialogue methods with unpublished source code, we use the results reported in the original papers. We obtained the results of BlenderBot fine-tuned on the MSC dataset from the HAHT paper. For LD-Agent, we utilize the publicly available code for reproduction. The HAHT paper does not provide experimental results on all sessions. We calculate the average of the results from each session as a reference. The CPD paper does not provide experimental results on different sessions. Therefore, we do not report its results on different sessions in our paper.

System prompt
As a communication expert with outstanding communication habits, you embody the role of CHATBOT throughout the following dialogues. Here are some of your distinctive personal traits: [CHATBOTPERSONA].

User prompt
<CONTEXT>
Drawing from your recent conversation with USER:
[CONTEXT]

Now, please role-play as CHATBOT to continue the dialogue between CHATBOT and USER.
USER just said: [QUERY]

Please respond to USER's statement using the following format (maximum 50 words, must be in English):

RESPONSE:

Figure 4: Prompt in both the main experiments and the ablation studies.

D.5 Prompt used in experiments

In all our experiments, we employed a total of three types of prompts. As illustrated in Figure 4 which displays the prompt provided to the LLM in both the main experiments and the ablation studies. [CHATBOTPERSONA] can be replaced by the vector output from persona prompt learning, [CONTEXT] is derived from the topic-aware memory bank, and [QUERY] represents the current query.

In the experiment of validating the effectiveness of Query-driven GNN, we employs query-related LLM prompts for commonsense reasoning on the graph driven by the queries, as illustrated in the Figure 5.

Finally, In the evaluation using G-Eval, Figure 6 illustrates the prompts inputted into the LLM for assessment.

System prompt
You are an expert with personality perception abilities. You can filter out personality triples related to the current user's query from the candidate personality triples, which are {}. The user's current query is {}. Please output the triples relevant to the query (note that the triples you output must be from the candidate triples):

User prompt
The user's current query is {}. Please output the triples relevant to the query (note that the triples you output must be from the candidate triples)

Outputs:

Figure 5: Query-related LLM prompts for commonsense reasoning.

System prompt
You are an expert in evaluating text generation and specialize in assessing systems from various perspectives. You will be provided with a dialogue history, the personality information of both participants, the user's current query, and the corresponding response. Your task is to evaluate the response in terms of consistency with personality information, consistency with the dialogue history, and other relevant dimensions. Finally, you will provide an overall score.

User prompt
You will be given a conversation between two individuals. You will then be given one potential response for the next turn in the conversation. The response concerns an interesting fact, which will be provided as well.
Your task is to rate the responses on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Crieteria:
Consistency (1-5) Is the response consistent with personas and dialogue history?
- A score of 1 (Completely Inconsistent) means that the response is entirely unrelated to the context, diverging from the conversation topic and it completely misaligned with the character's profile, with no sense of adherence
- A score of 2 (Clearly Inconsistent) means the response has low relevance, with logical conflicts or difficulty maintaining coherence and it significantly deviates from the character's profile, with mismatched tone and content.
- A score of 3 (Partially Consistent) means the response is limited relevance to the context, with some logical gaps or deviations and it partially aligns with the character, but there are noticeable inconsistencies in tone or behavior.
- A score of 4 (Mostly Consistent) means the response is generally relevant to the context, with slight deviations that do not affect understanding and it overall matches the character profile, with minor mismatches in tone or phrasing.
- A score of 5 (Fully Consistent) means the response seamlessly connects with the context, with no logical contradictions and it fully aligns with the established tone, personality, and knowledge scope.

Evaluation Steps:
1. Read the conversation, the corresponding fact and the response carefully.
2. Rate the response on a scale of 1-5 for Consistency, according to the criteria above.
3. Provide a brief explanation for your rating, referring to specific aspects of the response and the conversation.
Example:
Conversation History:{}, Corresponding Personas:{}, Current Query:{}, Response:{}
Evaluation Form (scores ONLY):

- Consistency:

Figure 6: G-Eval prompt.

E Ablation Study on different sessions

We conducted ablation experiments across different sessions on the MSC dataset. The experimental results using ChatGLM and Llama2 are shown in Table 8. We found that (1) regardless of the scale of the dialogue history, each module in our framework has a positive impact on different backbones across various sessions. (2) The topic-aware memory bank significantly enhances the model's ability to handle long textual history and is a dominant factor in the performance variations of the model across different history scales. The ablation experiments indicate that, with the addition of the topic-aware memory bank, the model's performance improves with the increase in dialogue history size within a certain range; however, beyond this range, the model's performance declines.

Table 7: Statistics of DuLeMon

Category	SELF	BOTH
# Dialogues	24500	3001
# Utterances	400472	48522
Avg. # turns	16.3	16.2
Avg. length of utterances	19.7	21.2
Avg. # bot persona	4.0	4.0
Avg. # user persona (seen)	0	4.4
Avg. # user persona (unseen)	4.0	1.3

F Experiments on DuLeMon

To demonstrate the generalizability of our method, we select an additional dataset, DuLeMon (Xu et al., 2022), for experimentation. DuLeMon is a large-scale dataset for the multi-session personalized dialogue task in Chinese. There are two versions of DuLeMon: *Self* and *Both*, where the persona information comes from only the self side (user side) or both side (both user and chatbot side). Its data statistics are presented in the Table 7. To further demonstrate the effectiveness of our method, we select several strong baselines for comparison with our approach:(1) vanilla LLM: we directly employ the LLM as the chatbot the same as section D.2, we utilize ChatGLM-6B (GLM et al., 2024), Llama2-7B-Chat(Touvron et al., 2023) and Llama2-Chinese[‡] as vanilla LLMs. (2) PLATO-FT: The PLATO-2 (Bao et al., 2021) model fine-tuned on our proposed DuLeMon dataset.(3) UniMS-RAG(Wang et al., 2024a): UniMS-RAG is a retrieval-augmented generation framework that integrates multiple knowledge sources for personalized dialogue.

From the table 9, we can observe that our method outperforms nearly all baselines on DuLeMon, including the currently leading models PLATO-FT and UniMS-RAG on the DuLeMon dataset. Llama2-Chinese outperforms Llama2 because it utilizes large-scale Chinese data, which enhances the Chinese capabilities of the Llama2 model from the pre-training phase.

[‡]https://github.com/LlamaFamily/Llama-Chinese

Table 8: Ablation experiments using ChatGLM and Llama2 as the backbone models on the MSC dataset. TMB stands for Topic-Aware Memory Bank, PPL refers to Persona Prompt Learning. B-1, B-2, B-3, R-L denote BLEU-1, BLEU-2. BLEU-3 and ROUGE-L respectively. The best results for each backbone model are in **bold**.

Model	Session2				Session3				Session4				Session5			
	B-1	**B-2**	**B-3**	**R-L**	**B-1**	**B-2**	**B-3**	**R-L**	**B-1**	**B-2**	**B-3**	**R-L**	**B-1**	**B-2**	**B-3**	**R-L**
ChatGLM (Base)	19.29	5.44	1.49	16.76	19.21	5.18	1.55	15.51	19.16	5.74	1.52	16.68	19.02	5.92	1.45	16.63
ChatGLM w/ TMB	19.41	7.43	2.38	19.23	19.54	7.57	2.48	19.63	19.62	7.59	2.49	19.71	19.32	7.24	2.21	19.01
ChatGLM w/ PPL	19.47	7.56	2.64	19.94	19.48	7.59	2.54	19.77	19.35	7.43	2.42	19.47	19.41	7.42	2.35	19.20
ChatGLM w/ VAE-LoRA	19.86	7.67	2.56	20.06	19.92	7.76	2.58	19.90	19.99	7.93	2.64	20.03	19.78	7.53	2.40	19.28
ChatGLM w/ TMB & PPL	19.88	7.63	2.49	20.13	20.04	7.78	2.58	20.23	20.17	8.00	2.65	20.27	19.67	7.52	2.46	19.27
ChatGLM w/ TMB & VAE-LoRA	19.93	7.73	2.63	20.27	20.23	7.83	2.61	20.39	20.30	8.17	2.70	20.77	19.73	7.59	2.55	19.30
ChatGLM w/ PPL & VAE-LoRA	19.90	7.69	2.58	20.10	19.95	7.76	2.58	20.07	20.03	7.99	2.68	20.29	19.75	7.53	2.43	19.30
PALACE (ChatGLM)	**20.10**	**7.91**	**2.68**	**20.28**	**20.40**	**7.93**	**2.67**	**20.51**	**20.69**	**8.24**	**2.93**	**20.94**	**19.78**	**7.65**	**2.67**	**19.33**
Llama2 (Base)	17.39	4.47	1.21	10.43	17.34	4.39	1.22	10.33	17.32	4.32	1.21	10.25	17.16	4.20	1.17	10.03
Llama2 w/ TMB	17.83	4.89	1.47	12.32	18.09	4.98	1.66	13.09	18.76	5.04	1.67	13.83	18.13	4.89	1.32	12.57
Llama2 w/ PPL	19.90	7.70	2.49	18.67	20.78	8.19	2.55	18.89	21.17	8.29	2.59	18.68	21.05	8.38	2.54	18.67
Llama2 w/ VAE-LoRA	19.96	7.93	2.58	18.80	20.47	8.05	2.59	18.89	21.19	8.48	2.64	18.94	20.95	8.43	2.58	18.44
Llama2 w/ TMB & PPL	20.10	8.00	2.59	18.88	20.83	8.21	2.63	19.13	21.26	8.57	2.65	19.20	21.10	8.47	2.60	18.72
Llama2 w/ TMB & VAE-LoRA	20.07	8.07	2.63	18.88	20.81	8.23	2.63	19.15	21.30	8.56	2.70	19.24	21.15	8.50	2.63	18.80
Llama2 w/ PPL & VAE-LoRA	19.99	7.96	2.59	18.83	20.53	8.09	2.58	19.07	21.23	8.53	2.62	18.99	21.03	8.44	2.58	18.53
PALACE (Llama2)	**20.95**	**8.13**	**2.66**	**18.96**	**20.87**	**8.29**	**2.70**	**19.22**	**21.45**	**8.64**	**2.77**	**19.32**	**21.29**	**8.54**	**2.66**	**18.83**

G Analysis on VAE-LoRA

We creatively combined VAE and LoRA within our framework to propose VAE-LoRA, providing effective prior knowledge necessary for dialogue generation in LLMs. To demonstrate the effectiveness of this module, we first used ChatGLM and Llama2 as backbones to compare the performance of VAE-LoRA with the original LoRA across different sessions in MSC.

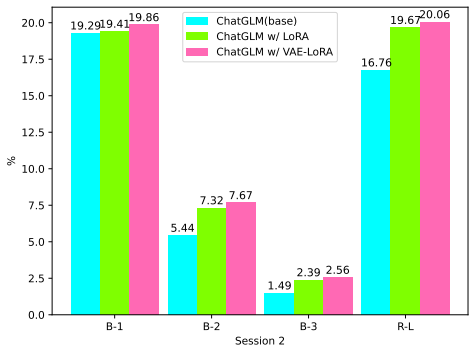
The results from Figures 7 and 8 demonstrate that, regardless of whether ChatGLM or Llama2 is used as the backbone, VAE-LoRA consistently has a positive impact on performance across all sessions in the MSC dataset and outperforms the original LoRA. This indicates that our proposed VAE-LoRA effectively provides useful prior knowledge to the LLM compared to the original LoRA.

To further investigate and analyze the specific types of prior knowledge that VAE-LoRA extracts, We randomly select thousands of dialogues from the dataset and perform t-SNE visualization of the original representations of the queries extracted by the LLM, as well as the query representations extracted by VAE-LoRA, as shown in Figure 9.

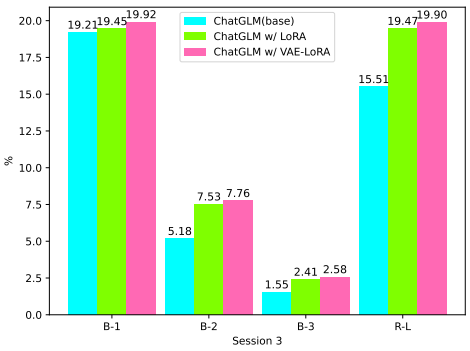
From the figure, it is evident that the query representations extracted by the LLM are uniformly distributed across an approximately circular plane, whereas the query representations extracted by VAE-LoRA exhibit distinct clustering centers, resulting in a clustered distribution. This indicates that VAE-LoRA can aggregate the queries to some extent based on the personalized implicit information presented within them. Such representations not only establish connections between different queries to enhance the LLM's understanding of various dialogue contexts, but also allow for the extraction of diverse personalized information from different clusters, thereby improving the LLM's sensitivity to personalization and ultimately enhancing its performance in multi-session personalized generation.

Table 9: Automatic evaluation results of different models on DuLeMon dataset. B-1, B-2, R-L, BS denote BLEU-1, BLEU-2, ROUGE-L and BERTScore respectively. The best results are in **bold** and the second-best results are in underlined.

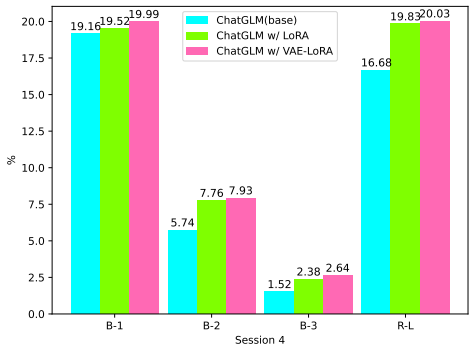
Model	B-1	B-2	R-L	BS
ChatGLM(Base)	13.34	4.49	14.91	62.08
Llama2(Base)	11.99	4.24	12.55	61.78
Llama2-Chinese	14.23	4.53	14.83	62.61
ChatGPT	14.73	4.96	16.67	63.07
PLATO-FT	19.40	8.10	-	-
UniMS-RAG	18.95	-	20.87	-
PALACE(ChatGLM)	19.48	8.66	**22.00**	64.71
PALACE(Llama2)	19.21	8.21	20.59	62.80
PALACE(Llama2-Chinese)	**19.93**	**8.89**	21.71	**64.81**



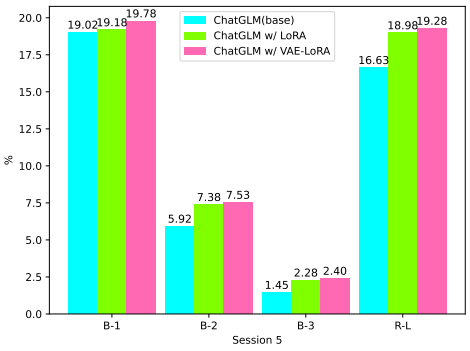
(a) The performance on Session 2 of the MSC.



(b) The performance on Session 3 of the MSC.

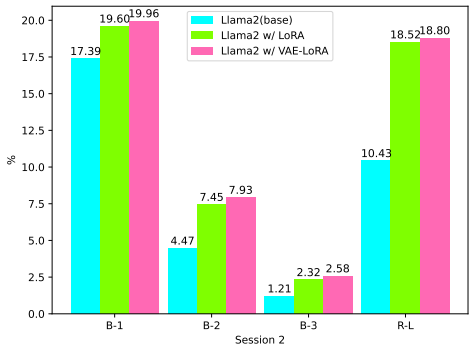


(c) The performance on Session 4 of the MSC.

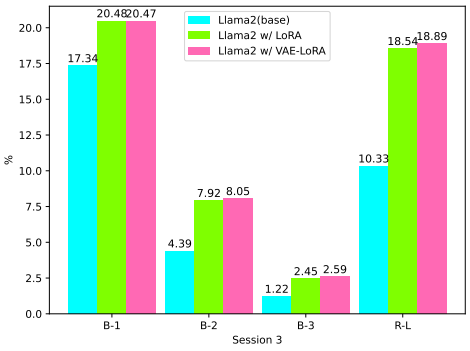


(d) The performance on Session 5 of the MSC.

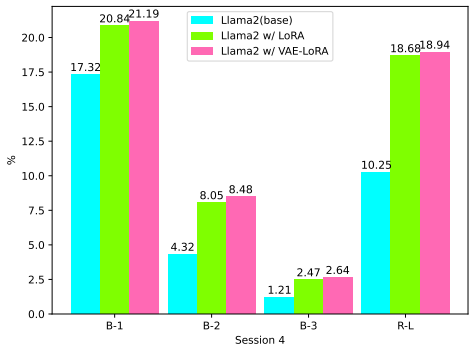
Figure 7: The performance of various LoRA fine-tuning methods on different sessions of MSC, utilizing ChatGLM as the backbone. B-1, B-2, R-L denote BLEU-1, BLEU-2 and ROUGE-L respectively



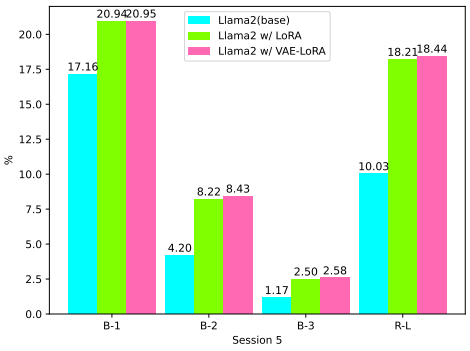
(a) The performance on Session 2 of the MSC.



(b) The performance on Session 3 of the MSC.

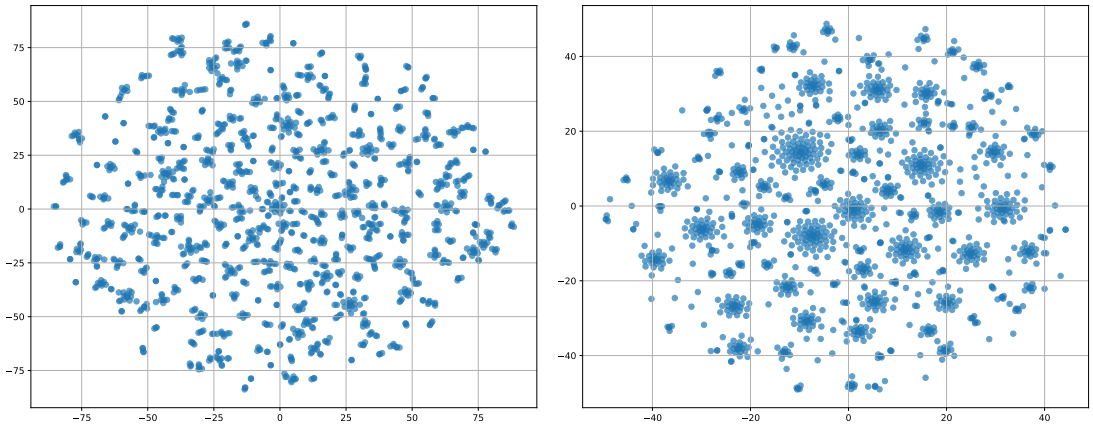


(c) The performance on Session 4 of the MSC.



(d) The performance on Session 5 of the MSC.

Figure 8: The performance of various LoRA fine-tuning methods on different sessions of MSC, utilizing Llama2 as the backbone. B-1, B-2, R-L denote BLEU-1, BLEU-2 and ROUGE-L respectively.



(a) Visualization of the representations extracted by the LLM.

(b) Visualization of the representations extracted by VAE-LoRA.

Figure 9: T-SNE visualization of query representations.

H Analysis on dialogue structure

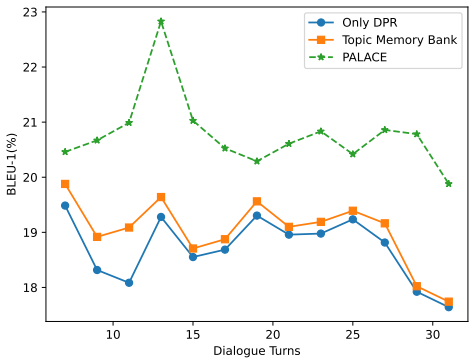
To demonstrate the effectiveness of the Topic Memory Bank within the framework, we partition the MSC dataset based on the number of dialogue turns and evaluate the results using only DPR as well as those using the Topic Memory Bank. In this experiment, we define a turn as a pair consisting of a query and a response. Due to the insufficient dialogue history in conversations with fewer than 8 turns and the lack of data in conversations with more than 32 turns, these results are not statistically significant; therefore, we have excluded them from the figure. From Figure 10, we can observe that: (1) Across BLEU-1, BLEU-2, BLEU-3, and ROUGE-L metrics, Topic Memory Bank consistently improves the model's performance in dialogue generation, regardless of the number of dialogue turns. (2) As the number of dialogue turns increases, the model's performance shows a decreasing trend. This decline may be attributed to the continuous increase in dialogue history, which can result in diminished retrieval performance or an overload of relevant history, making it difficult to retrieve all dialogue details, ultimately affecting the model's dialogue generation capabilities. (3) The addition of the Topic Memory Bank still shows a significant gap compared to PALACE, which is attributable to the critical roles played by VAE-LoRA and personalized prompt learning in enhancing dialogue generation.

Table 10: Automatic evaluation results of different hidden states on the MSC dataset. B-1, B-2, R-L denote BLEU-1, BLEU-2 and ROUGE-L respectively. The best results are in **bold** and the second-best results are in underlined.

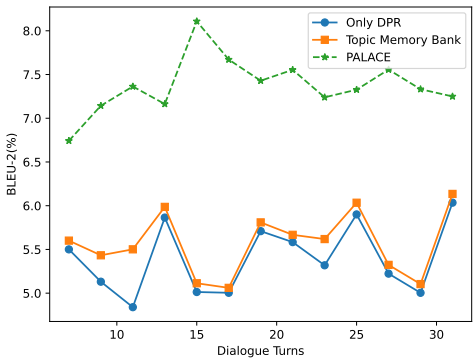
Hidden State	B-1	B-2	B-3	R-L
1024	20.07	7.85	2.48	**20.64**
2048	**20.17**	**7.92**	**2.73**	20.27
4096	19.55	7.72	2.53	20.24

I Evaluation of different hidden states in the GNN

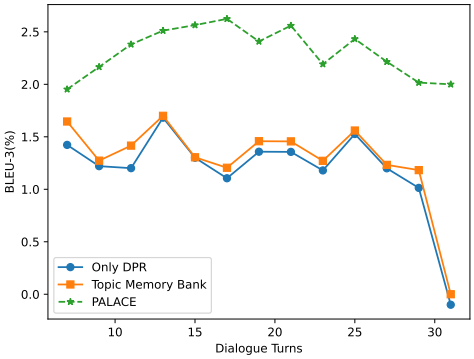
We designed additional experiments to illustrate why we set the hidden state of the GNN to 2048 in our experimental setup. We present the evaluation results using different hidden states on the MSC dataset with ChatGLM as the backbone in Table 10. The figure indicates that when the hidden state is set to 2048, BLEU-1, BLEU-2 and BLEU-3 are significantly superior to those of other hidden states. From the table, it is evident that although ROUGE-L is not optimal when the hidden state is set to 2048, BLEU-1, BLEU-2 and BLEU-3 significantly outperform those of other hidden states.



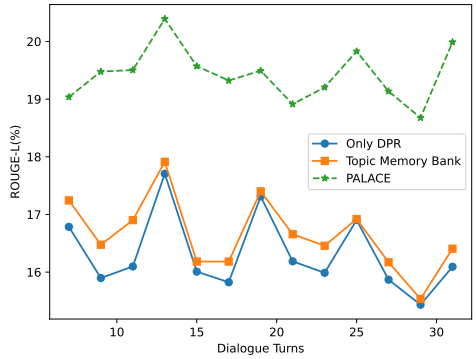
(a) BLEU-1 evaluation results.



(b) BLEU-2 evaluation results.



(c) BLEU-3 evaluation results.



(d) ROUGE-L evaluation results.

Figure 10: Performance results of the Topic Memory Bank and only DPR across different dialogue turn counts in the MSC dataset.

J Performance validation of Query-driven GNN

We conduct experiments using ChatGLM and Llama2 as backbones to further demonstrate the effectiveness of Query-driven GNN. We use the vanilla LLM, the RGCN-based commonsense reasoning approach, and the query-related prompt-based reasoning method as baselines. For query-related prompt-based reasoning method, we provide the LLM with the instruction, the graph, and the current query, prompting it to output nodes or subgraphs related to the query. The complete prompt can be found in D.5. The results are shown in the table below.

Table 11: Performance validation of Query-driven GNN on MSC dataset. B-1, B-2, B-3, R-L and BS denote the average BLEU-1, BLEU-2 BLEU-3, ROUGE-L and Bertscore scores across all sessions on the MSC, respectively. The best results for each backbone model are in **bold**.

Model	B-1	B-2	B-3	R-L	BS
ChatGLM(Base)	19.20	5.54	1.50	16.49	48.60
w/ Query-related prompt	19.34	5.93	1.71	17.84	50.29
w/ RGCN	19.37	7.48	2.50	19.34	51.82
w/ Query-driven GNN	**19.44**	**7.51**	**2.52**	**19.67**	**52.17**
Model	**B-1**	**B-2**	**B-3**	**R-L**	**BS**
Llama2(Base)	17.34	4.37	1.21	10.29	45.25
w/ Query-related prompt	18.01	5.11	1.48	12.15	46.91
w/ RGCN	20.18	7.94	2.51	18.70	53.77
w/ Query-driven GNN	**20.56**	**8.05**	**2.53**	**18.72**	**53.83**