Hello Again! LLM-powered Personalized Agent for Long-term Dialogue

Anonymous ACL submission

Abstract

Open-domain dialogue systems have seen remarkable advancements with the development of large language models (LLMs). Nonetheless, most existing dialogue systems predominantly focus on brief single-session interactions, neglecting the real-world demands for long-term companionship and personalized interactions with chatbots. Crucial to addressing this real-world need are event summary and persona management, which enable reasoning for appropriate long-term dialogue responses. Recent progress in the human-like cognitive and reasoning capabilities of LLMs suggests that LLM-based agents could significantly enhance automated perception, decision-making, and problem-solving. In response to this potential, we introduce a model-agnostic framework, the Long-term Dialogue Agent (**LD-Agent**), which incorporates three independently tunable modules dedicated to event perception, persona extraction, and response generation. For the event memory module, long and short-term memory banks are employed to separately focus on historical and ongoing sessions, while a topic-based retrieval mechanism is introduced to enhance the accuracy of memory retrieval. Furthermore, the persona module conducts dynamic persona modeling for both users and agents. The integration of retrieved memories and extracted personas is subsequently fed into the generator to induce appropriate responses. The effectiveness, generality, and cross-domain capabilities of LD-Agent are empirically demonstrated across various illustrative benchmarks, models, and tasks. The code is released at `https://anonymous.4open.science/r/LDA-D7B6`.

1 Introduction

Open-domain dialogue systems aim to establish long-term, personalized interactions with users via human-like chatbots (Xu et al., 2022a; Zhang et al., 2022; Xu et al., 2022b). Unlike most existing studies (Li et al., 2017; Zhang et al., 2018; Rashkin et al., 2019) that are limited to brief, single-session interactions spanning 2-15 turns, real-life scenarios often necessitate a chatbot's capability for long-term companionship and familiarity (Xu et al., 2022a; Zhang et al., 2022; Jang et al., 2023). Achieving this requires the chatbot not only to understand and remember extensive dialogue histories but also to faithfully reflect and consistently update both the user's and its personalized characteristics (Xu et al., 2022a; Jang et al., 2023; Zhang et al., 2023b).

Motivated by real-life demands, the core challenge of open-domain dialogue systems is to simultaneously maintain long-term event memory and preserve persona consistency (Gu et al., 2019; Cao et al., 2022; Zhao et al., 2023; Xu et al., 2022b). Existing research often addresses these aspects separately—focusing either on event memory or persona extraction—thereby hindering long-term consistency. Current strategies for event memory typically involve constructing a memory bank that stores historical event summaries, complemented by retrieval-augmented approaches to access relevant information for response generation (Chen et al., 2019; Zhang et al., 2019). Studies on persona-based dialogue rang from unidirectional user modeling (Chen et al., 2023a) to bidirectional agent-user modeling (Wu et al., 2020; Liu et al., 2020; Xu et al., 2022b), enhancing personalized chat abilities by leveraging profile information. Worse still, the aforementioned methods are highly dependent on specific model architectures, making them challenging to adapt to other models. Additionally, These dialogue models largely lack zero-shot generalization capabilities, essential for effective deployment across various real-world domains (Zhang et al., 2022; Xu et al., 2022b). We conjecture that an optimal long-term dialogue framework should be model-agnostic, deployable in various real-world domains, and capable of autonomously integrating comprehensive data from both event memories and personas, as illustrated in Figure 1. However, developing such a model-agnostic, cross-domain, and autonomous framework remains unexplored and challenging.

Benefiting from the excellent human-like cognitive and reasoning abilities of large language models (LLM), there is an increasing trend (Deng et al., 2023; Wang et al., 2023a; Qian et al., 2023; Park et al., 2023; Zhang et al., 2023a) to employ LLMs as the cores of agent-based simulation systems to automate the process of perception, decision-making, and problem-solving. While recent studies have developed LLM-powered agents in various fields, such as economics (Cheng and Chin, 2024), politics (Hua et al., 2023), sociology (Xu et al., 2024), and recommendation (Zhang et al., 2023a), its application in open-domain dialogue remains unexplored. To effectively support long-term open-domain dialogue, an LLM-powered dialogue agent framework should exhibit broad generality, cross-domain adaptability, and the ability to dynamically refine information across dimensions like events, user personalities, and agent personalities.

In this paper, we propose **LD-Agent**—a model-agnostic **L**ong-term **D**ialogue **Agent** framework consisting of three principal components: an event memory perception module, a persona extraction module, and response generation module (see the framework of LD-Agent in Figure 2). The event memory perception module is designed to enhance coherence across sessions by separately maintaining long-term and short-term memory banks. The long-term memory bank stores vector representations of high-level event summaries from previous sessions, refined through a tunable event summary module. The short-term memory bank maintains contextual information for ongoing conversations. The persona extraction module, designed to facilitate personalized interactions, incorporates a disentangled, tunable mechanism for accurate user-agent modeling. Extracted personas are continuously updated and stored in a long-term persona bank. These personas, along with relevant memories, are then integrated into the response generation module, guiding the generation of appropriate responses, as depicted in Figure 1.

We conduct comprehensive experiments on two illustrative long-term multi-session daily dialogue datasets, MSC (Xu et al., 2022a) and Conversation Chronicles (CC) (Jang et al., 2023), to evaluate the effectiveness, generality, and cross-domain capabilities of the proposed framework. In terms of effectiveness, LD-Agent achieves state-of-the-art performance on both benchmarks, significantly outperforming existing methods (Zhang et al., 2022; Zeng et al., 2023; Roller et al., 2021). To assess generality, we examine the framework from both model and task perspectives. From the model perspective, LD-Agent is evaluated across a range of both online and offline models, including LLMs (Zeng et al., 2023) and non-LLMs (Roller et al., 2021). From the task perspective, we extend our evaluation to multiparty dialogue tasks (Hu et al., 2019), where LD-Agent also demonstrates substantial improvements, showcasing its adaptability across different models and tasks. Regarding the method's cross-domain capabilities, we design two cross-domain settings: tuning the model on the MSC dataset and testing it on the CC dataset, and vice versa. In both scenarios, LD-Agent shows competitive performance, nearly matching the results of in-domain training.

Our contributions can be summarized as follows:

- We develop LD-Agent, a general long-term dialogue agent framework, considering both historical events, ensuring dialogue coherence across sessions, and personas, ensuring character consistency.

- We introduce a disentangled, tunable approach for long-term dialogue to ensure the accuracy of each module. The highly modular framework enables it to adapt to various dialogue tasks through module re-training.

- We confirm the superiority of our proposed framework through rigorous experiments across multiple challenging benchmarks, diverse illustrative models, and various tasks. Extensive insightful ablation studies further highlight its effectiveness and generalization.

2 Method

In this section, we introduce the LD-Agent in detail with the framework shown in Figure 2. We first introduce the task definition of long-term dialogue in Section. 2.1. Consequently, we separately introduce the mechanism of event perception (Section. 2.2), dynamic personas extraction (Section. 2.3), and response generation (Section. 2.4).

2.1 Task Definition

The goal of the long-term multi-session dialogue task is to generate an appropriate response r, by

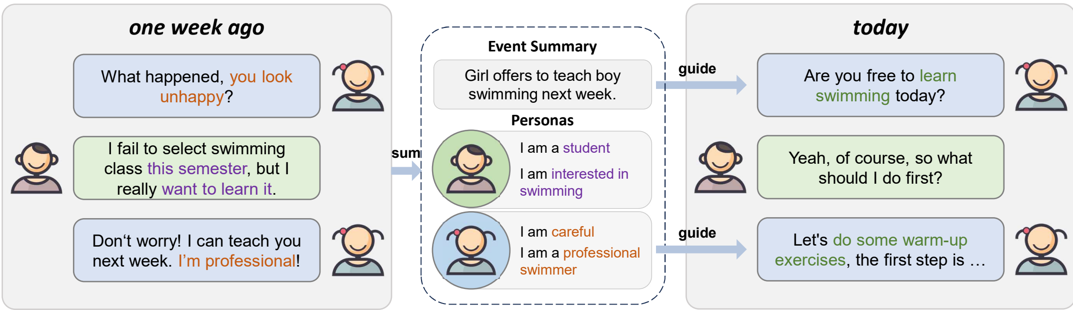


Figure 1: The illustration of how event memory and personas guide long-term dialogue. The event summary and personas are extracted from a conversation that occurred one week ago. In today's interaction, the event memory prompts the girl to inquire about the swimming lesson they scheduled last week. The personas, indicating that she is careful and professional in swimming, guide her to offer detailed and professional advice.

utilizing the context of the current session C, along with selected information extracted from historical session H. In this task, the current conversation session C is defined as $\{u_1, u_2, \ldots, u_{d_c-1}, u_{d_c}\}$, where each u_i represents i-th utterance, and d_c represents d_c turns of the current session. Each historical session within H in N historical sessions is denoted as H^i, containing $\{h_1^i, h_2^i, \ldots, h_{d_i}^i\}$, where d_i is the number of utterances of the i-th conversational session. Distinct from single-session dialogue models, a long-term multi-session dialogue system integrates both current and long-term historical conversational cues to generate contextually appropriate responses.

2.2 Event Perception

The event memory module is designed to perceive historical events to generate coherent responses across interval time. As shown in Figure 2, this event memory module is segmented into two major sub-modules that focus separately on long-term and short-term memory.

2.2.1 Long-term Memory

Memory Storage. The long-term memory module aims to extract and encode events from past sessions. Specifically, this involves recording the occurrence times t and brief summaries o into representations that are stored in a low-cost memory bank $M_L = \{\phi(t_j, o_j) \mid j \in \{1, 2, \ldots, l\}\}$. Here, $\phi(\cdot)$ indicates the text encoder (*e.g.*, MiniLM (Wang et al., 2020)), and l specifies the length of the memory bank. The encoded representations are then efficiently retrieved through an embedding-based mechanism, which enhances the accessibility of the stored memory.

Event Summary. Different from previous agent approaches (Park et al., 2023; Zhang et al., 2023a; Zhong et al., 2024) that entirely rely on LLM's zero-shot ability to excavate and summarize events, we apply instruction tuning (Wei et al., 2022a) to the event summary module, which can directly improve the event summary quality. Specifically, we rebuild the DialogSum dataset (Chen et al., 2021), a large-scale dialogue summarization dataset, into the following format: (1) an introduction to the task background, (2) the related conversations that need to be understood, and (3) detailed summarization requests. These three parts serve as input prompts (see Appendix. E.1 for more details), combined with the original summaries from DialogSum as answers, and are jointly used to fine-tune the event summary module, thereby directly improving the quality of event summarization.

Memory Retrieval. To improve retrieval accuracy, we employ a retrieval mechanism that comprehensively considers semantic relevance, topic overlap, and time decay. Optimizing the retrieval accuracy of agent memory is challenging due to the difficulty in obtaining accurate memory retrieval data. Most existing methods (Park et al., 2023; Zhang et al., 2023a) use event summaries as keys and context as queries, calculating the query-key semantic relevance score s_{sem} to find relevant memories, which inevitably results in significant errors. To enhance retrieval reliability, we extract nouns from corresponding conversations with the summaries to construct a topic library V and calculate topic overlap score s_{top} by:

$$s_{\text{top}} = \frac{1}{2}\left(\frac{|V_q \cap V_k|}{V_q} + \frac{|V_q \cap V_k|}{V_k}\right), \quad (1)$$

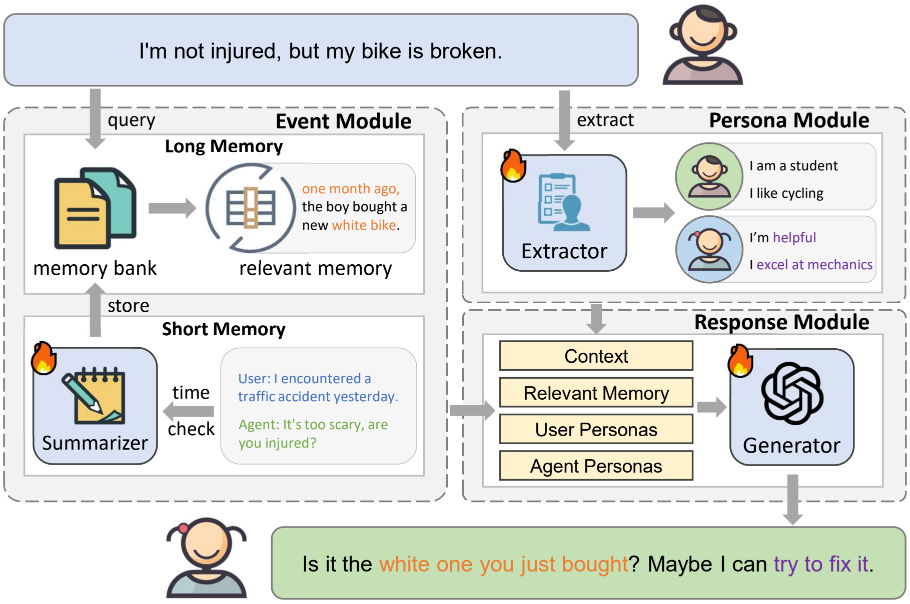


Figure 2: The Framework of LD-Agent. The event module stores historical memories from past sessions in long-term memory and current context in short-term memory. The persona module dynamically extracts and updates personas for both users and agents from ongoing utterances, storing them in a persona bank for each character. The response module then synthesizes this data to generate informed and appropriate responses.

where V_q, V_k denote the topic noun set of query and key. Additionally, we apply a time decay coefficient $\lambda_t = e^{-t/\tau}$ to reweight the overall retrieval score s_r, signified as:

$$s_{\text{overall}} = \lambda_t(s_{\text{sem}} + s_{\text{top}}). \tag{2}$$

To avoid retrieving inappropriate memory due to no suitable memories existing, we implement a semantic threshold γ. Only memories with semantic score s_{sem} greater than γ could be retrieved. If no appropriate memories are retrieved, "No relevant memory" will be returned. Eventually, the process of retrieving relevant memory m is denoted as:

$$m = \psi(M_L, \gamma). \tag{3}$$

2.2.2 Short-term Memory

The short-term memory module actively manages a dynamic dialogue cache $M_S = \{(t_i, u_i) | i = \{1, 2, 3, \dots, r_c\}\}$ with timestamps to preserve the detailed context of the current session. Upon receiving a new utterance u', the module first evaluates the time interval between the current time t' and the last recorded time t_{r_c} in the cache. If this interval exceeds a threshold β, the module triggers the long-term memory module to summarize the cached dialogue entries, creating new event records for storage in the long-term memory bank. Simultaneously, the short-term memory cache is cleared, and the new dialogue record (t', u') is added to the cache. The mathematical expression of this process is given by:

$$\begin{aligned} M_L' &= M_L \cup \{(\phi(t_{r_c}, A(M_S))\}, \\ M_S &= \{(t', u')\}. \end{aligned} \tag{4}$$

where M_L' denotes the updated long-term memory bank, $o = A(\cdot)$ is the event summary function, which process the accumulated dialogue in M_S.

2.3 Dynamic Personas Extraction

The persona module is pivotal in maintaining long-term persona consistency for both participants in a dialogue system. Drawing inspiration from prior work (Xu et al., 2022b), we adopt a bidirectional user-agent modeling approach, utilizing a tunable persona extractor to manage long-term persona bank P_u and P_a for the user and agent, respectively. Specifically, we develop an open-domain, utterance-based persona extraction dataset derived from MSC (Xu et al., 2022a). We enhance the persona extractor with LoRA-based instruction tuning, which allows for the dynamic extraction of personality traits during conversations. These traits are subsequently stored in the corresponding character's persona bank. For utterances devoid of personality traits, the module outputs "No Trait". Additionally, we employ a tuning-free strategy that harnesses the zero-shot capabilities of LLM models to directly extract personas based on prompts (see Appendix. E.2). To further improve the ability to excavate user personas without training, we adjust our reasoning strategy from direct reasoning to a Chain-of-Thought reasoning (Wei et al., 2022b).

2.4 Response Generation

Upon receiving a new user utterance u', the agent integrates various inputs: retrieved relevant memories m, short-term context M_S, and the personas P_u and P_a for the user and agent, respectively. These combined inputs are fed into a response generator to deduce an appropriate response r, formulated as

$$r = G(u', m, M_S, P_u, P_a). \tag{5}$$

To enhance the agent's ability for coherent and contextually appropriate responses, we develop a long-term, multi-session dialogue dataset, featuring dynamic retrieval memories, context, and personas sourced from the MSC and CC datasets for generator tuning. Specifically, for each sample, covering five sessions, we dynamically simulate the entire progression of the conversation. As each new utterance is introduced, the previously tuned modules for event summarization, persona extraction, and memory retrieval are utilized to collect the necessary context, retrieved memories, and both user and agent personas related to the utterance. This comprehensive data is then integrated into a response generation prompt (see Appendix. E.3). The original responses from the MSC and CC datasets are used as ground truth sentences.

3 Experiments

We aim to answer the following research questions:

- **RQ1:** How does LD-Agent perform in long-term dialogue tasks?
- **RQ2:** How is the generality and practicality of LD-Agent?

3.1 Evaluation Settings

In this subsection, we briefly introduce the experimental dataset, evaluation metrics, and baseline models in our study. Detailed evaluation settings are elaborated in Appendix. C.

Datasets. Extensive experiments are conducted on two illustrative multi-session datasets, **MSC** (Xu et al., 2022a) and **CC** (Jang et al., 2023), each comprising 5 sessions with approximately 50 conversational turns per sample, to investigate the effectiveness of LD-Agent on long-term dialogue scenarios. The experiments cover model independence assessment, module ablation, persona extractor analysis, and cross-domain evaluation.

Additionally, to evaluate the transferability of the LD-Agent, we apply our method to the **Ubuntu IRC benchmark** (Hu et al., 2019), a dataset known for its multiparty interaction tasks.

Metrics. Our evaluation combines both automatic and human assessments to thoroughly investigate the effectiveness of LD-Agent. For automatic evaluation, we use three widely used standard metrics: BLEU-N (BL-N) (Papineni et al., 2002), ROUGE-L (R-L) (Lin, 2004), and METEOR (MET) (Banerjee and Lavie, 2005) to measure the quality of response generation. Additionally, accuracy (ACC) is employed to evaluate the classification performance of the persona extractor. In human evaluation, we measure topic coherence across sessions, interaction fluency, and user engagement using the metrics of coherence, fluency, and engagingness, respectively.

Baselines. To demonstrate the effectiveness and model independence of LD-Agent, we deploy LD-Agent on multiple platforms and models. Specifically, the LLM-based models (online model: ChatGPT; offline model: ChatGLM (Zeng et al., 2023)) and traditional language models (BlenderBot (Roller et al., 2021), and BART (Lewis et al., 2020)) are employed as our baselines. In our experiments, The notation "**Model**$_{\text{LDA}}$" denotes models that incorporate the LD-Agent framework, while "**Model**" refers to the original baseline models without LD-Agent. Additionally, we also utilize HAHT (Zhang et al., 2022), the previous state-of-the-art model in long-term dialogue task, as a contrast. See the above baselines stand and their role in rich literature in Appendix. A.

3.2 Results of Multi-Session Dialogue

We adopt two multi-session dialogue dataset to evaluate our method in long-term dialogue scenarios. The first session is used to initialize conversation and the subsequent four sessions are used to evaluate the performance of long-term dialogue. In these experiments, LD-Agent is applied to both

	Model	Session 2			Session 3			Session 4			Session 5		
		BL-2	BL-3	R-L	BL-2	BL-3	R-L	BL-2	BL-3	R-L	BL-2	BL-3	R-L
MSC													
Zero-shot	ChatGLM	5.44	1.49	16.76	5.18	1.55	15.51	5.63	1.33	16.35	5.92	1.45	16.63
	ChatGLM$_{\text{LDA}}$	5.74	1.73	17.21	6.05	1.73	16.97	6.09	1.59	16.76	6.60	1.94	17.18
	ChatGPT	5.22	1.45	16.04	5.18	1.55	15.51	4.64	1.32	15.19	5.38	1.58	15.48
	ChatGPT$_{\text{LDA}}$	8.67	4.63	19.86	7.92	3.55	18.54	7.08	2.97	17.90	7.37	3.03	17.86
Tuning	HAHT	5.06	1.68	16.82	4.96	1.50	16.48	4.75	1.45	15.82	4.99	1.51	16.24
	BlenderBot	5.71	1.62	16.15	8.10	2.50	18.23	7.55	1.96	17.45	8.02	2.36	17.65
	BlenderBot$_{\text{LDA}}$	8.45	3.27	19.07	8.68	3.06	18.87	8.16	2.77	18.06	8.31	2.69	18.19
	ChatGLM	5.48	1.59	17.65	6.12	1.78	17.91	6.14	1.63	17.78	6.16	1.69	17.65
	ChatGLM$_{\text{LDA}}$	**10.70**	**5.63**	**23.31**	**10.03**	**5.12**	**21.55**	**9.07**	**4.06**	**20.19**	**8.96**	**4.01**	**19.94**
CC													
Zero-shot	ChatGLM	8.94	4.44	21.54	8.34	4.03	21.00	8.28	3.82	20.67	8.12	3.81	20.54
	ChatGLM$_{\text{LDA}}$	9.53	4.82	22.76	9.22	4.43	22.18	9.15	4.48	22.18	8.99	4.43	22.10
	ChatGPT	10.57	5.50	22.10	10.58	5.59	22.04	10.61	5.58	21.92	10.17	5.22	21.45
	ChatGPT$_{\text{LDA}}$	15.89	11.01	26.96	12.92	8.27	24.31	12.20	7.35	23.69	11.54	6.74	22.87
Tuning	BlenderBot	8.99	4.86	21.58	9.44	5.19	22.13	9.46	5.21	22.08	8.99	4.75	21.73
	BlenderBot$_{\text{LDA}}$	14.47	10.16	27.91	15.66	11.33	29.10	15.13	10.80	28.38	14.08	9.72	27.37
	ChatGLM	15.89	9.90	30.59	15.97	10.06	30.27	16.10	10.31	30.54	15.10	9.34	29.43
	ChatGLM$_{\text{LDA}}$	**25.69**	**19.53**	**39.67**	**25.93**	**19.72**	**39.15**	**25.82**	**19.40**	**39.05**	**24.26**	**18.16**	**37.61**

Table 1: Experimental results of the automatic evaluation for response generation on MSC and CC.

zero-shot models, including ChatGLM and ChatGPT, and to tuned models like BlenderBot and ChatGLM with the results reported in Table 1.

Impressive performance on long-term dialogue tasks. On both datasets, all models employing LD-Agent consistently achieve significant improvements across all sessions and metrics, showcasing the powerful ability of LD-Agent on supporting long-term dialogue. Most notably, comparing with previous state-of-the-art model HAHT, BlenderBot employing LD-Agent, which has similar parameter scale to HAHT, outperforms it with a large performance gap of 3.39%, 3.72%, 3.41%, and 3.32% on BLEU-2 ranging from session 2 to 5. This further highlighting the effectiveness of LD-Agent on long-term dialogue tasks.

Remarkable generality of LD-Agent. The generality of LD-Agent are proved from two aspects: data transferability and model transferability. The consistently improvements brought by LD-Agent on both benchmarks demonstrate the generality of our framework on various long-term dialogue scenarios. In parallel, we observe that LD-Agent also plays positive roles in the zero-shot setting, employing to the online model of ChatGPT and the offline model of ChatGLM. In the tuning setting, LD-Agent achieves significant enhancements on both LLM of ChatGLM and traditional model of BlenderBot, fully proving the remarkable model transferability of LD-Agent. These results comprehensive demonstrate the generality of LD-Agent.

3.3 Ablation Studies

To further analyze the effectiveness of each components, we conduct ablation studies for memory module and personas module. We adopt ChatGLM as our backbone, which is tuned solely using the context of the current session, referred to here as "Baseline". Afterward, we separately add "Event Memory", "Agent personas", and "User personas" modules for additional tuning on top of the baseline. The results are presented in Table 2.

The results clearly demonstrate that all modules positively influence long-term dialogue capabilities, with the event memory module contributing the most significant improvements. It is worth noting that although all modules experience a performance decline as the number of sessions increased, the addition of the event memory module results in more stable performance compared to the use of user or agent personas. This highlights the critical role of event memory in maintaining coherence across multiple sessions.

3.4 Persona Extraction Analysis

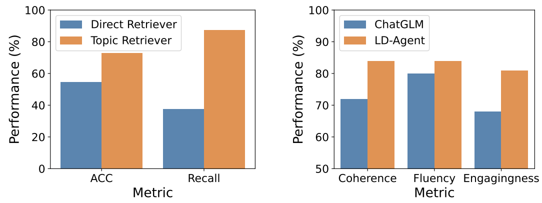
To explore the effect of different persona extractor, including zero-shot ChatGLM with Chain-of-Thought (Wei et al., 2022b) and ChatGLM tuned on the persona extraction dataset collected from MSC training set, we carry out comparison experiments on two perspectives: Persona Extraction

Model	Session 2			Session 3			Session 4			Session 5		
	BL-2	**BL-3**	**R-L**	**BL-2**	**BL-3**	**R-L**	**BL-2**	**BL-3**	**R-L**	**BL-2**	**BL-3**	**R-L**
Baseline	5.48	1.59	17.65	6.12	1.78	17.91	6.14	1.63	17.78	6.16	1.69	17.65
+ Mem	7.57	2.49	19.50	7.70	2.48	19.46	7.53	2.31	19.26	7.56	2.33	19.03
+ Persona $_{user}$	7.54	2.57	19.68	7.51	2.38	19.39	7.30	2.09	18.80	7.08	2.27	18.79
+ Persona $_{agent}$	7.00	2.27	18.70	7.23	2.33	18.75	7.32	2.18	18.47	7.13	2.36	18.48
Full	**10.70**	**5.63**	**23.31**	**10.03**	**5.12**	**21.55**	**8.96**	**4.01**	**19.94**	**9.07**	**4.06**	**20.19**

Table 2: Ablation study results of LD-Agent on MSC. The experiments are conducted on tuned ChatGLM. Baseline denotes the model tuned with context of current session. "+ module name" indicates the model tuned solely with context and corresponding module. "Full" indicates the model tuned with all modules.

Extractor	Extraction				Generation		
	BL-2	**BL-3**	**R-L**	**ACC**	**BL-2**	**BL-3**	**R-L**
CoT	5.05	2.69	25.54	61.6	5.82	1.69	16.95
Tuning	**8.31**	**5.65**	**43.70**	**77.8**	**6.12**	**1.75**	**17.03**

Table 3: The effect of different extractors on persona extraction and response generation on MSC.



(a) Retrieval Mechanism (b) Response Generation

Figure 3: The results of human evaluation on retrieval mechanism and response generation.

Accuracy and Impact to Response Generation. The results are shown in Table 3.

Extraction Accuracy. We evaluate the extraction accuracy on the persona extraction dataset collected from MSC testing set, through BLEU-2/3, R-L, and ACC. ACC is to assess the classification accuracy of dividing utterance into "with personas" or "without personas". The results of extraction in Table 3 show that the extractor after tuning performs better than CoT on all metrics. The higher BLEU and R-L indicates the tuned extractor performs better capability to extract personas, while higher ACC indicates a stronger capability to distinguish whether personas are contained in an utterance.

Impact to Response Generation. In addition, to explore the effect of different persona extractor to the final response generation, we conduct experiments on MSC by comparing the results of zero-shot ChatGLM$_{\text{LDA}}$ with personas extracted by CoT and tuned extractor, respectively. The Generation results in Table 3 indicate the tuned extractor performs better in most sessions. As the number of sessions increases, the gap is also constantly expanding, demonstrating tuned extractor is more suitable for long-term dialogue.

3.5 Human Evaluation

To further explore the performance of LD-Agent in real-life conversation, we adopt human evaluation to evaluate the ability of memory recall and response generation with the results on Figure 3.

Retrieval Mechanism Analysis. Retrieval mechanism plays a crucial role for event memory accurately utilized in long-term dialogue. To evaluate the superiority of topic-based retrieval approach than direct semantic retrieval commonly used in previous methods, we conduct an event memory human evaluation. In the beginning, we initialize a conversation using first four sessions and store event memories for each session into long-term memory bank. In the last session, we let evaluators select relevant memories from long-term memory bank for each utterance as the ground truths. Consequently, we separately utilize direct semantic retrieval and topic-based retrieval to search relevant memories for each utterance, and calculate the accuracy and recall based on human annotations. The results are shown in Figure 3(a). The topic-based retrieval outperforms direct semantic retrieval with significant gap on both ACC and Recall, proving that our retrieval method accurately retrieves relevant memories.

Response Generation Analysis. To further validate the superiority of LD-Agent in long-term open-domain dialogue tasks, we organize multiple multi-session human-bot conversations on ChatGLM with LD-Agent and w/o LD-Agent. We first initialize a predefined dialogue as the first session for all chatbots. Subsequently, we employ some human evaluators to chat with each chatbot with a

Model	Session 2			Session 3			Session 4			Session 5		
	BL-2	BL-3	R-L	BL-2	BL-3	R-L	BL-2	BL-3	R-L	BL-2	BL-3	R-L
MSC												
Zero-shot	5.44	1.49	16.76	5.59	1.49	16.47	5.63	1.33	16.35	5.92	1.45	16.63
Zero-shot$_{\text{LDA}}$	5.74	1.73	17.21	6.05	1.73	16.97	6.09	1.59	16.76	6.60	1.94	17.18
CC-tuning	5.81	1.74	18.79	6.08	1.83	18.58	5.96	1.74	18.31	5.95	1.68	18.23
CC-tuning$_{\text{LDA}}$	7.86	3.63	21.00	7.46	3.16	20.00	7.15	2.87	19.53	7.12	2.64	19.30
MSC-tuning	5.48	1.59	17.65	6.12	1.78	17.91	6.14	1.63	17.78	6.16	1.69	17.65
MSC-tuning$_{\text{LDA}}$	**10.70**	**5.63**	**23.31**	**10.03**	**5.12**	**21.55**	**9.07**	**4.06**	**20.19**	**8.96**	**4.01**	**19.94**
CC												
Zero-shot	9.53	4.82	22.76	9.22	4.43	22.18	9.15	4.48	22.18	8.99	4.43	22.10
Zero-shot$_{\text{LDA}}$	8.94	4.44	21.54	8.34	4.03	21.00	8.28	3.82	20.67	8.12	3.81	20.54
MSC-tuning	8.37	3.88	22.93	8.49	3.99	22.96	7.97	3.75	22.15	7.60	3.70	21.87
MSC-tuning$_{\text{LDA}}$	21.71	15.42	34.97	20.87	14.74	34.01	19.57	13.51	32.72	18.59	12.80	31.68
CC-tuning	15.89	9.90	30.59	15.97	10.06	30.27	16.10	10.31	30.54	15.10	9.34	29.43
CC-tuning$_{\text{LDA}}$	**25.69**	**19.53**	**39.67**	**25.93**	**19.72**	**39.15**	**25.82**	**19.40**	**39.05**	**24.26**	**18.16**	**37.61**

Table 4: The results of cross-domain evaluation on MSC and CC. "Zero-shot" indicates the ChatGLM without tuning. "CC-tuning" indicates the ChatGLM tuned on CC. "MSC-tuning" indicates the ChatGLM tuned on MSC.

time interval from first session. The interactions are evaluated on three aspects: coherence, fluency and engagingness. The results in Figure 3(b) demonstrate the advantages of LD-Agent in long-term real-life dialogue scenarios.

3.6 Generality Analysis

We further explore its generality from two perspectives: cross-domain and cross-task capability.

Cross-domain Results. The cross-domain capability is crucial for open-domain dialogue task. Poor cross-domain performance, common in models tuned with specific datasets, limits their real-world practicality. To assess our tuned model's real-world potential, we conduct cross-evaluation on MSC and CC, two multi-session datasets with significant domain gaps due to different collection methods, including manual annotation and LLM generation. We first tune ChatGLM on CC and test it on MSC, then reverse the process. The results, shown in Table 4, indicate that models tuned on one dataset still performs well on the other dataset, only with a slight performance decrease than the models tuned on the same dataset. Besides, cross-domain tuned models consistently outperform zero-shot models by a significant margin. These experiments highlight the strong cross-domain capability and practical potential of LD-Agent.

Cross-task Results. The other capability worth exploring is the transferability of LD-Agent to different dialogue tasks. We explore the effectiveness of our method on multiparty dialogue, a task requires playing multiple roles simultaneously. We conduct our experiments on Ubuntu IRC dataset (Hu et al., 2019), a commonly used multiparty dialogue dataset. where our backbone adopts BART (Lewis et al., 2020). We compare our method with some previous multiparty dialogue methods, including GPT-2 (Radford et al.), GSN (Hu et al., 2019), HeterMPC$_{\text{BART}}$ (Gu et al., 2022), and BART tuned without prompt. The results are reported at Table 5 in Appendix. D. It can be seen that BART tuned with LD-Agent obtained the state-of-the-art performance in most metrics, outperforming previous multiparty dialogue approach HeterMPC$_{\text{BART}}$, which also employs BART as backbone. This well proves the powerful task transferability of LD-Agent.

4 Conclusion

In this work, we delved into the long-term open-domain dialogue agent to meet real-life chatbot demands for long-term companionship and personalized interactions. We introduced a model-agnostic long-term dialogue agent framework, LD-Agent, which comprehensively considers both historical events and user-agent personas to support coherent and consistent conversation. Our framework, decomposed into three learnable modules, significantly enhances adaptability and transferability. Extensive experiments demonstrated LD-Agent's strong capability in handling long-term dialogue tasks, showcasing its practicality across multiple benchmarks, models, and tasks.

Limitations

Though LD-Agent exhibits impressive effectiveness and generality on long-term dialogue, we believe that the research on long-term open-domain dialogue still has a long way to go. For instance, there are some remained limitations of this work from the following perspectives:

Lacking Real-World Datasets . Current long dialogue datasets are typically synthetic, created manually (Xu et al., 2022a) or generated by large language models (Jang et al., 2023; Maharana et al., 2024), which introduces a gap from real-world data. Due to the challenges in gathering authentic long-term dialogue data, our work is currently confined to these synthetic datasets. We aim to validate our approach on real long-term dialogue data in the future.

Sophisticated Module Design . In this paper, LD-Agent provides a general framework for long-term dialogue that allows for modular optimization. However, the module implementations only employ some basic methods without more sophisticated design, which can be further explored in the future.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*, pages 65–72.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *ACL*, pages 7984–8002.

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

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023b. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *CoRR*, abs/2305.13614.

Xiuyi Chen, Jiaming Xu, and Bo Xu. 2019. A working memory model for task-oriented dialog response generation. In *ACL*, pages 2687–2693.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023c. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *EMNLP Findings*, pages 1170–1183.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *ACL-IJCNLP Findings*, volume ACL/IJCNLP 2021, pages 5062–5074.

Junyan Cheng and Peter Chin. 2024. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *ICLR*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samual Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. In *NeurIPS*.

Yang Deng, Yaliang Li, Wenxuan Zhang, Bolin Ding, and Wai Lam. 2022. Toward personalized answer generation in e-commerce via multi-perspective preference modeling. *ACM Trans. Inf. Syst.*, 40(4):87:1–87:28.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S^3: Social-network simulation system with large language model-empowered agents. *CoRR*, abs/2307.14984.

Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *EMNLP-IJCNLP*, pages 1845–1854.

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022. Hetermpc: A heterogeneous graph neural network for response generation in multi-party conversations. In *ACL*, pages 5086–5097.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. GSN: A graph-structured network for multi-party dialogues. In *IJCAI*, pages 5010–5016.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *CoRR*.

Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender AI agent: Integrating large language models for interactive recommendations. *CoRR*, abs/2308.16505.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *EMNLP*, pages 13584–13606.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, pages 986–995.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*.

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 *ACL*, pages 1417–1427.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. *CoRR*, abs/2402.17753.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST*, pages 2:1–2:22.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *CoRR*, abs/2307.07924.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, pages 5370–5381.

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 *EACL*, pages 300–325.

Omar Shaikh, Valentino Chai, Michele J. Gelfand, Diyi Yang, and Michael S. Bernstein. 2023. Rehearsal: Simulating conflict to teach conflict resolution. *CoRR*, abs/2309.12309.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *EMNLP*, pages 13153–13187.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *CoRR*, abs/2305.16291.

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

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, and Weiping Wang. 2023b. Enhancing empathetic and emotion support dialogue generation with prophetic commonsense inference. *CoRR*, abs/2311.15316.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2023c. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *CoRR*, abs/2310.00746.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *ICLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Bowen Wu, Mengyuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. In *ACL*, pages 53–65.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *ACL*, pages 5180–5197.

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. AI for social science and social science of AI: A survey. *CoRR*, abs/2401.11839.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *ACL Findings*, pages 2639–2650.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *ICLR*.

An Zhang, Leheng Sheng, Yuxin Chen, Hao Li, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2023a. On generative agents in recommendation. *CoRR*, abs/2310.10108.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023b. Mind the gap between conversations for improved long-term dialogue generation. In *EMNLP Findings*, pages 10735–10762.

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? In *ACL*, pages 2204–2213.

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 *EMNLP Findings*, pages 3395–3407.

Zheng Zhang, Minlie Huang, Zhongzhou Zhao, Feng Ji, Haiqing Chen, and Xiaoyan Zhu. 2019. Memory-augmented dialogue management for task-oriented dialogue systems. *ACM Trans. Inf. Syst.*, 37(3):34:1–34:30.

Kang Zhao, Wei Liu, Jian Luan, Minglei Gao, Li Qian, Hanlin Teng, and Bin Wang. 2023. Unimc: A unified framework for long-term memory conversation via relevance representation learning. *CoRR*, abs/2306.10543.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *AAAI*, pages 19724–19731.

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. *CoRR*, abs/2311.16832.

Appendix

In this Appendix, we discuss the following topics: **(1)**: We elaborate on some related work about long-term open-domain dialogue and LLM-based autonomous agent in Appendix. A. **(2)**: We visualize responses of original ChatGLM and LD-Agent to further demonstrate the ability of LD-Agent in long-term dialogue (see Appendix. B). **(3)**: More detailed experimental settings are introduced in Appendix. C. **(4)**: More experimental results are shown in Appendix. D. **(5)**: In the Appendix. E, the prompts utilized in LD-Agent is illustrated.

A Related Work

A.1 Long-term Open-domain Dialogue

Open-domain dialogue aims to develop a human-like chatbot that can emulate human conversation, facilitating free-flowing dialogue on a wide range of topics. However, the dialogue's extent in earlier studies is often limited by conversation length, focusing primarily on brief conversations of about 2-15 turns within a single session (Li et al., 2017; Zhang et al., 2018; Rashkin et al., 2019). To support more realistic and extended conversations, a series of studies have explored the role of both external (Wang et al., 2023b, 2024) and internal knowledge (Zhang et al., 2022; Xu et al., 2022b) on maintaining the feasibility of long-term dialogue. Commonly referenced external knowledge, such as commonsense (Wang et al., 2024), medical (Chen et al., 2023b), and psychological (Chen et al., 2023c) knowledge, serves as supplementary guidance for the reasoning process, ensuring logical coherence in extended contexts. In parallel, internal knowledge captured dynamically during long conversations generally contains historical events (Xu et al., 2022a; Zhang et al., 2023b, 2022; Jang et al., 2023) and personas (Gu et al., 2019; Xu et al., 2022b; Cao et al., 2022; Deng et al., 2022). Historical events are typically summarized and stored into a memory bank to maintain dialogue coherence across sessions, while interlocutors' personas are maintained via a dynamic persona memory bank, which ensures character consistency in long-term conversations. In this study, we focus on the internal knowledge to integrate dynamically updated historical events and personas to conduct long-term personalized conversations.

A.2 LLM-based Autonomous Agents

AI Agent conception is geared towards autonomous environmental perception, decision-making, and problem-solving capabilities. With the large language models (LLMs) underlining their impressive generalization potential, leading to their widespread adoption as substitutes for human operators in various research fields (Deng et al., 2023; Qian et al., 2023; Dillion et al., 2023; Zhang et al., 2023a). Generally, these agents can be categorized into task-oriented agents (Deng et al., 2023; Wang et al., 2023a; Qian et al., 2023; Zhong et al., 2024) and simulation-oriented agents (Dillion et al., 2023; Shaikh et al., 2023; Gao et al., 2023; Zhang et al., 2023a; Huang et al., 2023). Task-oriented agents are designed to accurately perform and achieve predefined tasks, as seen in applications for web assistance (Deng et al., 2023), game-playing (Wang et al., 2023a), and software development (Qian et al., 2023). On the other hand, simulation-oriented agents are devised to emulate human emotive and cognitive behaviors, having played roles in psychological studies (Dillion et al., 2023), social networking platforms (Gao et al., 2023), conflict resolution scenarios (Shaikh et al., 2023), and recommendation systems (Zhang et al., 2023a; Huang et al., 2023). In addition, recent developments have seen the advent of individual-level agents that are utilized to simulate specific character behaviors, enhancing the realism and personalization of user-agent interactions (Shao et al., 2023; Zhou et al., 2023; Wang et al., 2023c). This paper falls into the simulation-oriented agent to build a human-like open-domain dialogue agent with memory retrieved and character analysis modules.

B Response Visualization

To further analyze the ability of LD-Agent in long-term dialogue, we illustrate an example in Figure 4. It can be seen that the response generated by LD-Agent successfully captures the information about "General Nathan Bedford Forrest" they talked about in the history session.

C Detailed Evaluation Settings

In this section, we introduce the detailed experimental dataset, evaluation metrics, baseline models, and our implementation details.

C.1 Datasets

Multi-session Dataset. Our experiments are conducted on two illustrative multi-session datasets: **MSC** (Xu et al., 2022a) and **CC** (Jang et al., 2023). Both datasets feature 5 sessions, with approximately 50 conversational turns per sample. MSC extends the PersonaChat dataset (Zhang et al., 2018), utilizing PersonaChat for the initial session and employing human-human crowd workers to simulate the dialogues in subsequent sessions. The time intervals between sessions can span several days, and the dataset includes records of the participants' personas. We follow the split of (Zhang et al., 2022) with 4,000 conversations for training, 500 conversations for validation, and 501 conversations for testing. CC is complied by ChatGPT, which guides interactions according to a predefined event graph and participant relationships, with time intervals between sessions extending over several years. We employ the same data scale as MSC, with 4,000 conversations for training, 500 conversations for validation, and 501 conversations for testing.

Multi-party Dataset. To explore the transferability of LD-Agent on other dialogue tasks. We apply our method to the **Ubuntu IRC benchmark** (Hu et al., 2019), a dataset of multiparty tasks. We follow the split of previous works (Hu et al., 2019; Gu et al., 2022) with 311,725 dialogues for training, 5,000 dialogues for validation, and 5,000 dialogues for testing.

C.2 Metrics

Automatic Evaluation Metrics. BLEU-N (Papineni et al., 2002) (BL-N) and ROUGE-L (Lin, 2004) (R-L) metrics are commonly used automatic evaluation metrics in dialogue generation tasks. BLEU-N measures N-gram overlaps between the generated text and the reference text, while ROUGE-L focuses on sequential coherence. We employ the METEOR (MET) (Banerjee and Lavie, 2005) metric in multi-party tasks as a complement to the BLEU metric, enhancing it with synonym calculation capabilities. In addition, accuracy (ACC) is calculated to measure the classification accuracy of different persona extractors.

Human Evaluation Metrics. In human evaluation, we evaluate LD-Agent on three aspects: coherence, fluency, and engagingness. Coherence measures the chatbot's capabilities to maintain the coherence of topic and logic across sessions. Fluency reflects the natural and fluent degree of interactions, making the interaction similar to human-human interactions. Engagingness measures a user's interest in interacting with the target chatbot.

C.3 Baselines

To validate the effectiveness of our method on various baselines, we employ LD-Agent on both online and offline models, tuned and zero-shot models, LLMs, and non-LLMs.

- **HAHT** (Zhang et al., 2022): This is the state-of-the-art model crafted for multi-session, open-domain dialogue. It encodes all historical information and utilizes an attention mechanism to capture the relevant information to an ongoing conversation.

Figure 4: Example of separately chatting with original ChatGLM and ChatGLM with LD-Agent. A more relevant response to history conversation is generated.

- **BlenderBot** (Roller et al., 2021): This is a commonly used large-scale open-domain dialogue model pre-trained on online social discussion data.
- **ChatGLM3** (Zeng et al., 2023): This is an offline large language model 6B parameters. The model is pre-trained on 1T corpus, performing remarkable zero-shot reasoning capabilities.
- **ChatGPT**: This is an online large language model based on the GPT architecture with excellent human-like cognitive and reasoning abilities. In this paper, we use the API service with the model of "gpt-3.5-turbo-1106".
- **BART** (Lewis et al., 2020): This is a denoising autoencoder with transformer architecture, trained to reconstruct original text from corrupting text.

C.4 Implementation Details.

For the event summarizer, persona extractor, and response generator modules, we employ the LoRA mechanism across all configurations. All training and evaluation operated on a single NVIDIA A100 GPU. For the ChatGLM3-6B, it is optimized by an Adam (Kingma and Ba, 2015) optimizer with the learning rate of 5e-5. We configure this model with a batch size of 4 and train it over 3 epochs. For BlenderBot, the initial learning rate is set to 2e-5, with the batch size and the number of training epochs set at 4 and 5, respectively.

D Additional Experimental Results

In this section, we introduce some additional experimental results.

D.1 Cross-task Results

In Section 3.6, we present experiments designed to evaluate the cross-task capabilities of LD-Agent. Specifically, we apply LD-Agent to a multiparty dialogue task using the Ubuntu IRC dataset (Hu et al., 2019) as the benchmark. The results, summarized in Table 5, reveal that BART, when fine-tuned with LD-Agent, consistently outperforms other methods across most metrics. This underscores the robustness of LD-Agent in facilitating effective task transferability.

E Prompt

In this section, we separately provide the illustrations of the prompts used in the Event Module, Persona Module, and Response Module.

Table 5: Multi-party performance on the Ubuntu IRC benchmark

Model	BL-1	BL-2	BL-3	BL-4	MET	R-L
GPT-2 (Radford et al.)	10.37	3.60	1.66	0.93	4.01	9.53
GSN (Hu et al., 2019)	10.23	3.57	1.70	0.97	4.10	9.91
HeterMPC$_{\text{BART}}$ (Gu et al., 2022)	12.26	4.80	**2.42**	**1.49**	4.94	11.20
BART (Lewis et al., 2020)	11.25	4.02	1.78	0.95	4.46	9.90
BART$_{\text{LDA}}$	**14.40**	**4.92**	2.07	1.00	**5.30**	**12.28**

E.1 Prompt of Event Summary

Prompt 1: Event Summary Prompt

SYS PROMPT:

You are good at extracting events and summarizing them in brief sentences. You will be shown a conversation between {*user name*} and {*agent name*}.

USER PROMPT:

Conversation: {*context*}.
Based on the Conversation, please summarize the main points of the conversation with brief sentences in English, within 20 words.
SUMMARY:

E.2 Prompt of Persona Extraction

Prompt 2: Persona Extraction Prompt

SYS PROMPT:

You excel at extracting user personal traits from their words, a renowned local communication expert.

USER PROMPT:

If no traits can be extracted in the sentence, you should reply NO_TRAIT. Given you some format examples of traits extraction, such as:
1. No, I have no longer serve in the millitary, I had served up the full term that I signed up for, and now work outside of the millitary.
Extracted Traits: I now work elsewhere. I used to be in the military.
2. That must a been some kind of endeavor. Its great that people are aware of issues that arise in their homes, otherwise it can be very problematic in the future.
NO_TRAIT
Please extract the personal traits who said this sentence (no more than 20 words):{*sentence*}

E.3 Prompt of Response Generation

Prompt 3: Base Response Generation Prompt

SYS PROMPT:

As a communication expert with outstanding communication habits, you embody the role of {*agent name*} throughout the following dialogues.

USER PROMPT:

<CONTEXT>
Drawing from your recent conversation with {*user name*}:
{*context*}
Now, please role-play as {*agent name*} to continue the dialogue between {*agent name*} and {*user name*}.
{*user name*} just said: {*input*}
Please respond to {*user name*}'s statement using the following format (maximum **30** words, **must be in English**):
RESPONSE:

Prompt 4: Agent Response Generation Prompt

SYS PROMPT:

As a communication expert with outstanding communication habits, you embody the role of {*agent name*} throughout the following dialogues. Here are some of your distinctive personal traits: {*agent traits*}.

USER PROMPT:

<CONTEXT>
Drawing from your recent conversation with {*user name*}:
{*context*}
<MEMORY>
The memories linked to the ongoing conversation are:
{*memories*}
<USER TRAITS> During the conversation process between you and {*user name*} in the past, you found that the {*user name*} has the following characteristics:
{*user traits*}
Now, please role-play as {*agent name*} to continue the dialogue between {*agent name*} and {*user name*}.
{*user name*} just said: {*input*}
Please respond to {*user name*}'s statement using the following format (maximum **30** words, **must be in English**):
RESPONSE: