Interpersonal Memory Matters: A New Task for Proactive Dialogue Utilizing Conversational History

Anonymous ACL submission

Abstract

Proactive dialogue systems aim to empower chatbots with the capability of leading conversations towards specific targets, thereby enhancing user engagement and service autonomy. Existing systems typically target pre-defined keywords or entities, neglecting user attributes and preferences implicit in dialogue history, hindering the development of long-term user intimacy. To address these challenges, we take a radical step towards building a more human-like conversational agent by integrating proactive dialogue systems with long-term memory into a unified framework. Specifically, we define a novel task named **M**emory-**a**ware **P**roactive **Dia**logue (**MapDia**). By decomposing the task, we then propose an automatic data construction method and create the first **Ch**inese **M**emory-**a**ware **P**roactive **Data**set (**ChMapData**). Furthermore, we introduce a joint framework based on Retrieval Augmented Generation (RAG), featuring three modules: Topic Summarization, Topic Retrieval, and Proactive Topic-shifting Detection and Generation, designed to steer dialogues towards relevant historical topics at the right time. The effectiveness of our dataset and models is validated through both automatic and human evaluations.

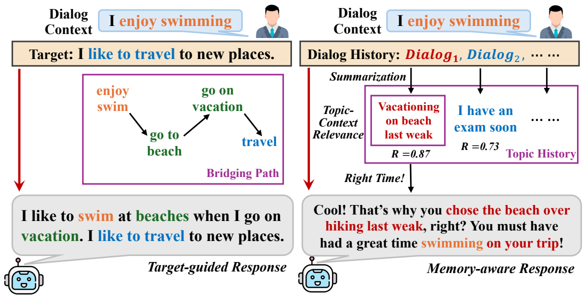


Figure 1: Comparison of previous proactive dialogue systems (Left) that extracted from Gupta et al. (2022) and our system (Right) on the same sample: The left system transitions the context to a pre-designed target through a bridging path, whereas our system involving summarization, retrieval, and timing detection to generate the memory-aware response.

1 Introduction

Recent years have witnessed significant advancements in the design of conversational agents, with various methods proposed to generate engaging responses, e.g., external knowledge (Xu et al., 2023; Yang et al., 2024), personality traits (Madaan et al., 2020; Ju et al., 2022), and the utilization of large-scale models (Fan et al., 2023; Liu et al., 2024). Among these, proactive behavior in agents—where the agent takes control of the conversation instead of merely responding passively to users—has been identified as a crucial advancement for the next generation of conversational AI (Deng et al., 2023).

Nevertheless, a more lifelike dialogue system must go beyond generating contextually appropriate responses; it should also employ more intelligent mechanisms to maintain a coherent social relationship over time (Campos et al., 2018). Memory, in particular, has already been acknowledged as essential for driving conversations, developing rapport, and maintaining long-term connections (Zhong et al., 2024; Maharana et al., 2024). However, existing proactive dialogue systems insufficiently exploit memory mechanisms, whose targets are pre-defined ones, such as topical keywords (Tang et al., 2019), knowledge entities (Wu et al., 2019a), conversational goals (Liu et al., 2020), while overlooking the contextual richness of dialogue history. Even the advanced ChatGPT (Achiam et al., 2023) faces constraints, yielding responses to inquiries regarding the context, reflecting a passive engagement with historical data. On one hand, pre-defined topics may fail to align with user interests, which is further discussed in Appendix A based on previous research. On the other hand, as users' personal information accumulates over time, ignoring historically interpersonal interactions with agents (i.e., memory mentioned in this paper) causes a failure to capture their attributes and preferences. All of these contradict the proactive dialogue motivation of enhancing user engagement.

To bridge this gap, we integrate proactive dialogue systems with memory mechanisms, moving closer to creating more intelligent and human-like conversational agents. Specifically, we propose the Memory-aware Proactive Dialogue task, depicted in Figure 1. Contrary to traditional proactive systems that respond based solely on a pre-designed target, our approach extracts topics from past dialogues, identifies the most relevant topic as target, assesses the appropriateness of topic transitions, and finally integrates memory into the response.

To minimize extensive human annotation, we introduce an automated four-step data construction method, proven effective in validation. This method uses GPT-4 (Achiam et al., 2023) to develop ChMapData, the first Chinese Memory-aware Proactive Dataset encompassing all the information to perform MapDia including dialogue histories, corresponding topics, current contexts, topic transition timings, and history-informed responses. Specifically, we guide the generation of certain historical dialogues using memorable subjects, such as events in which the user has participated. These dialogues serve as references for the subsequent generation of proactive chats and shape the current context, ensuring continuity and facilitating memory recall. The final segment of the data determines the appropriateness of topic transitions during conversations and formulates responses accordingly, concentrating on either shifting to a historical topic or maintaining the ongoing context.

With ChMapData, we propose a new proactive dialogue framework containing three components: 1) Topic Summarization, condensing historical dialogues into topics for simplified retrieval; 2) Topic Retrieval, identifying the most relevant historical topic with a RAG mechanism; and 3) Proactive Topic-shifting Detection and Generation, timing and executing topic transitions at optimal moments.

The main contributions of this paper are as follows: 1) We are the first to integrate memory technique into proactive dialogue systems and introduce a novel task of **MapDia**, where the system navigates current dialogue towards relevant historical topics at an appropriate opportunity; 2) We propose an effective automated data construction methodology and, based on this, construct **ChMapData**, the first memory-aware proactive dialogue dataset in Chinese. This dataset will be publicly released to support future dialogue community; 3) We present a RAG-based proactive dialogue framework that combines summarization, retrieval, timing detection, and response generation mechanisms. Both automatic and human evaluations demonstrate the effectiveness of our method.

2 ChMapData Dataset

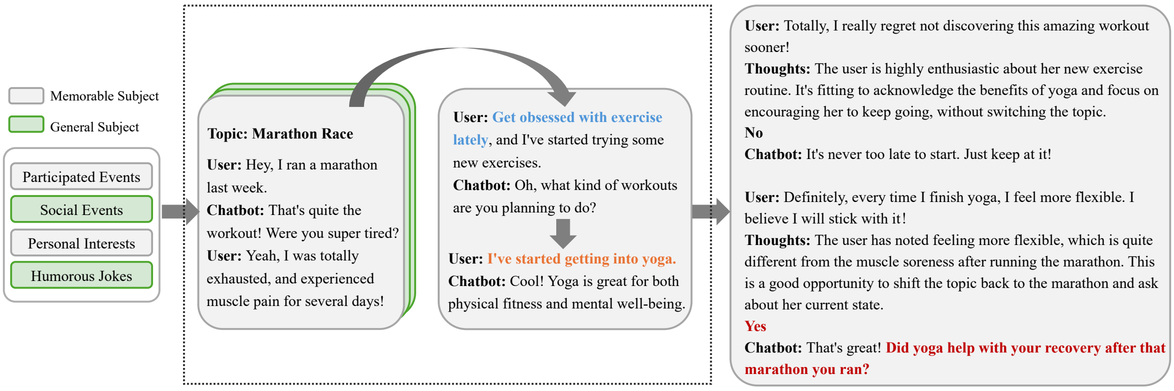
Despite the existence of Chinese datasets for proactive dialogue systems (Wu et al., 2019b; Zhou et al., 2020), they lack the ability to engage with the dialogue history while either steering the conversation towards a new topic or continuing with the current one. To fill this gap, we automatically generate the first multi-turn dataset designed for proactive conversations that leverage historical memory utilizing GPT-4 with a range of prompts (detailed in Appendix B). This process is further validated by annotators, thereby eliminating the high costs and lengthy procedures associated with human annotation. Note that we construct the dataset through GPT-4 because LLMs have been proven as powerful tools for synthetic data generation (Agrawal et al., 2022; Liu et al., 2022; Bitton et al., 2023).

2.1 Data Construction

Figure 2 gives an overview of the ChMapData construction pipeline, involving a four-step process.

1) Subject Selection. The initial phase involves GPT-4 brainstorming to generate a pool of potential subjects. Out of these, 11 subjects are manually selected and categorized into two groups: Memorable Subjects, intended to evoke recollections related to the user's own experiences , including personal interests, feelings, skills, traits, participating events, and events' progression; as well as General Subjects, which have no direct connection to the user's life and are not typically brought up again, encompassing social events, opinion debates, humorous jokes, audience stories, and knowledge sharing.

2) Topic and Dialogue Generation. On the basis of 6 Memorable and 5 General Subjects selected above, a fine-grained topic along with corresponding dialogues are generated serving as a bank of dialogue histories. To emulate the flow of real-life conversations, we crafted more dialogues for General Subjects than Memorable ones, at a ratio of 2:1, yielding 500 and 250 dialogues for each respectively, culminating in 4,000 topic-dialogue pairs. Each dialogue is limited to 5-8 turns to maintain brevity and focus. Among these, 1500 dialogues originate from Memorable Subjects, designed to



(a) Subject Selection (b) Topic & Dialogue Generation (c) Dialogue Continuation (d) Topic-shift Detection & Response Generation

Figure 2: The pipeline of dataset construction. Not derived from the actual dataset.

potentially trigger memory in subsequent conversations. For every one of these 1500 dialogues, 1 to 10 additional dialogues are selected from candidate pools and manually sequenced to construct a coherent, conflict-free multi-segment chat history.

3) Dialogue Continuation. Subsequently, we initiate a current dialogue session by generating two beginning turns, which is a continuation of each prior dialogue generated in the second step after a lapse of several days. Specifically, since step 2 ensures that each conversation history has a Memorable Subject-driven dialogue, we extend the dialogue to facilitate memory recollection in the following step. The two turns are generated separately: The first turn is derived from the topic and dialogue content (generated in Step 2), maintaining consistency with the prior dialogue, as indicated by the blue sentence in Figure 2. The second turn is then developed with a focus solely on the content of the first one (generated in Step 3) to prevent shifting the current conversation directly back to the dialogue history, as highlighted in orange. In total, 1,500 beginnings for current dialogues have been created, corresponding to the number of dialogue histories produced under Memorable Subject.

4) Topic-shift Detection and Response Generation. Ultimately, we continued the conversation based on the beginning of each current dialogue (generated in Step 3), and tried to naturally introduce new topics related to the preset memorable conversation history at appropriate moments. Inspired by the Chain of Thought (CoT) (Wang et al., 2022) technique, each turn additionally incorporates a "Thoughts" feature, aiming to enhance the accuracy and interpretability of the detection process, together with a decision-making mechanism to identify whether it's an appropriate opportunity to switch topics, as highlighted in red in Figure 2. It should be noted that ending the dialogue without switching to the historical topic is also permissible.

[1]Dialogue irrelevant to the subject has been filtered out.

	Hist. Dlg.	**Curr. Dlg.**
# Dialogues	3,989[1]	1,464
# Utterances	40,619	16,373
# Unique Tokens	21,822	12,503
# Thoughts	-	5,081
# Topic-shift Sess.	-	1,234
Avg. # Utts. Length	33.23	38.97
Avg. # Utts. per Sess.	10.14	11.18

Table 1: Statistics of both historical (Hist. Dlg.) and current dialogue (Curr. Dlg.) dataset. **# Thoughts** represents the chatbot's considerations on whether to switch the topic at each turn. **# Topic-shift Sess.** refers to conversations that successfully revert to the historical topic. The calculation of **# Utterances** excludes the **# Thoughts**, considering only the dialogue segments.

2.2 Overall Statistics

After data construction, we enhanced the dataset's quality by checking and manually removing 36 dialogues from the current dialogues due to format inconsistencies or illogical "Thoughts," resulting in a total of 1,464 entries. Statistics of the ChMapData are presented in Table 1, which is reported from two aspects: historical dialogue, generated in Step 2, and current dialogue, initiated with two turns in Step 3 and extended to the end in Step 4. Out of these, 1,234 dialogues successfully recalled the historical topic, as indicated by a "Yes" output during the detection process. The remaining 230 dialogues, which consistently output "No" throughout the session, are also retained for training.

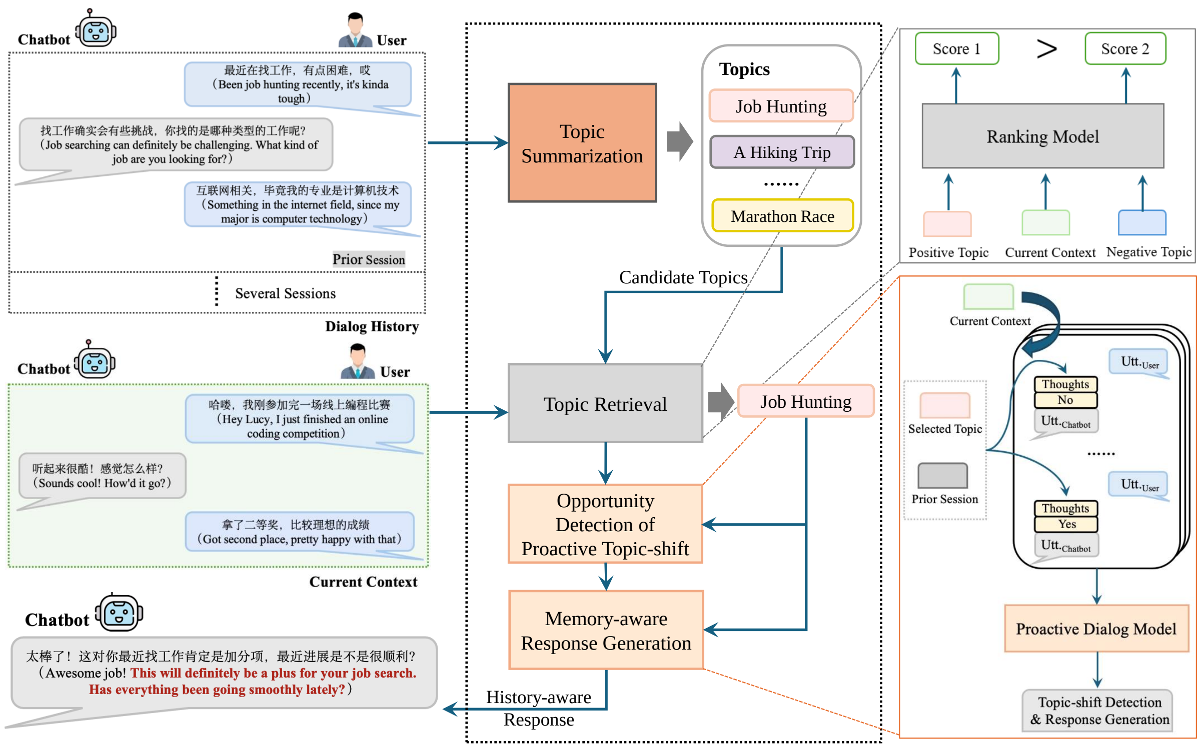


Figure 3: An overview of our system. Left showcases an example of proactive dialogue with memory awareness. Middle outlines the pipeline, featuring a summarization model for topic extraction, a ranking model to identify relevant historical topics, and a proactive dialogue model for topic shifts and reintroducing past information at the appropriate moments. Right is a breakdown detailing how these models operate.

3 Approach

Task Definition. Given Given a set of dialogue history $H = \{d_1, d_2, \ldots, d_n\}$ consisting of n dialogues, where d_i denotes the i-th dialogue and t_i represents its topic, and the current dialogue context c, the system is tasked with generating a topic-shift response R that proactively guide the conversation c to a related historical topic t_r at an appropriate opportunity — specifically at turn τ.

Up until now, we have obtained the ChMapData, containing historical dialogues with the corresponding topics, current context, thoughts on topic shifting, and response content. With these supports, we propose partitioning MapDia into three distinct modules as follows and integrating them through a RAG framework as shown in Figure 3.

Topic Summarization. Xu et al. (2022a) noted that dense retrieval of past conversations has two drawbacks: it requires storing large amounts of context and places a heavy workload on the model to extract and retrieve information. To address this, we start by condensing dialogue history into topics using a summary model. The training data for this step is derived from historical dialogues along with their corresponding topics (referred to as *ChMapData-Sum*), and the model is trained to summarize a topic t_i for each dialogue history d_i.

Topic Retrieval. We then developed a ranking model to identify the most pertinent summarized topic t_r for the current context c, facilitating continuous memory updates and the integration of historical information within the dialogue system. This model utilizes context c, along with its dependent historical topic t, as outlined in ChMapData construction's Step 3 (denoted c-t pairs as *ChMapData-Ret*) [2]. Given that the ranking model trains a classifier to predict preference probabilities between pairs of responses, as modeled by the Bradley-Terry model (Bradley and Terry, 1952). To prepare the dataset, we use GPT-4 to evaluate the relevance of the target topic t and 29 other randomly chosen topics from the pool to c, generating positive T^+ and negative T^- samples. The highest-ranked topic and t form T^+; if they coincide, only one positive example is constructed. Topics ranked lower than t become T^-, enhancing the dataset while ensuring the top-ranked topic is never a negative example.

For each dialogue context c, a training sample is

[2] Please note that c comprises two beginning turns of dialogue generated in Section 2.1 Step 3 and the first utterance user-generated in Step 4, making a total of 5 utterances.

formed by pairing a topic t^+ from T^+ with a corresponding negative topic t^-, which is randomly selected from T^-. The ranking model is implemented by appending a randomly initialized linear head to predict a scalar value. We then estimate the parameters of the ranking model by optimizing the maximum likelihood loss, defined as follows:

$$\mathcal{L}(\theta, \mathcal{D}) = \mathbb{E}_{(c,t^+,t^-)\sim\mathcal{D}}[\log(1 + e^{r_\theta(c,t^-)-r_\theta(c,t^+)})]$$

where $r_\theta(c, t)$ is the scalar output of the ranking model with parameters θ, and $\mathcal{D}$ is the preprocessed dataset of pairwise judgments. During inference, the ranking model outputs a scalar value, such that $P(t^+ \succ t^- \mid c) \propto e^{r_\theta(c,t^+)}$, which is learned through pairwise loss that topic t^+ is preferred over t^- given context c. Thus, topic t^+ is considered superior to t^- when $r_\theta(c, t^+) > r_\theta(c, t^-)$.

Proactive Topic-shifting Detection and Generation. Ultimately, we trained a memory-aware proactive response generation model to proactively lead the current conversation c towards the identified topic t_r through multiple turns of responses $R = \{r_1, r_2, \ldots, r_m\}$ at an appropriate moment τ. The training data for this step is called *ChMapData-Mem*, which comprises historical dialogues with their corresponding topics and the current dialogue as inputs, along with Thoughts and detection for topic shifting, and response content as learning objectives. As previously mentioned, the bot initially assesses whether it is an appropriate time to transition to a historical topic based on the current context c, and provides the reasoning behind this decision as a form of CoT. Subsequently, it generates the response content, with "Yes" or "No" indicating whether the response incorporates memory or is based solely on the current context.

4 Experiments

We design comparative experiments from two perspectives (both individual modules and the entire framework), assess two approaches (RAG-based alongside end-to-end) and utilize different test sets (our new ChMapData-test and an existing dataset).

4.1 Dataset

Our evaluation involved creating a new test set **ChMapData-test**, following the method outlined in Section 2.1. Please refer to Appendix C for the detailed construction process. Additionally, we incorporated the existing Chinese dataset **NaturalConv** (Wang et al., 2021) as conversation history to construct test data, so as to evaluate the method's generalization to unseen topics.

4.2 Compared approaches

In our exploration of the overall framework, we conduct a series of experiments from both RAG-based and end-to-end perspectives. Given that RAG-based methods comprise three components—namely, a module for processing dialogue history, the retriever, and the generator—we have designed four progressive permutations.

- **BGE w/ Qwen:** Widely-used BGE-M3 retrieval model (Chen et al., 2024) retrieves relevant memories from raw dialogue history, with Qwen2.5 generating proactive dialogue responses as a baseline.
- **QSum w/ BGE w/ Qwen:** Compared to BGE w/ Qwen, BGE-M3 retrieves memories from historical topics condensed by our fine-tuned Qwen on the *ChMapData-Ret* dataset, named QSum.
- **QSum w/ QRet w/ Qwen:** Compared to QSum w/ BGE w/ Qwen, the retrieval model is replaced with our fine-tuned QRet.
- **QSum w/ QRet w/ QMem (Ours):** Fine-tuned QMem that has topic-shifting capability represents the dialogue model while using Qsum and QRet.
- **Qwen-E2E:** Fine-tuned Qwen on ChMapData in an end-to-end (E2E) manner, utilizing all original dialogues as references without any intermediate steps such as summarization or retrieved results.
- **GPT4-E2E:** GPT-4, via prompt engineering, generates memory-aware responses.

To compare methods for proactively introducing topics using dialogue history, we use Qwen2.5-7B[3] as the base LLM unless otherwise specified. Implementation details are in AppendixE, and full prompts are in AppendixF. Observations from our ChMapData dataset show that user responses significantly influence the model's ability to transition topics. To prevent subconscious topic steering by human annotators and ensure objectivity while reducing costs, we trained a **User-role Dialogue Model**. For more information, see Appendix G.

4.3 Evaluation Metrics

Following previous works (Yuan et al., 2019; Han et al., 2021), we utilized Recall ($R_{10}@k$) to evaluate topic retrieval module, where the correct topic is among the top k out of ten candidates, specifically using $\boldsymbol{R_{10}@1}$, $\boldsymbol{R_{10}@2}$, and $\boldsymbol{R_{10}@3}$. We also used **MRR** and **NDCG** as additional retrieval

[3]https://huggingface.co/Qwen/Qwen2.5-7B

Models	Arch.	Retrieval	Achievement	Overall Quality	Engagement *Utts.-level*	Engagement *Sess.-level*	Avg.	#Turn
			ChMapData-test					
BGE w/ Qwen	RAG	per Sess.	0.02	0.89	0.02	0.02	0.34	4.70
		per Utt.	0.01	0.88	0.04	0.02	0.30	5.30
QSum w/ BGE w/ Qwen	RAG	per Sess.	0.04	0.92	0.05	0.05	0.38	4.52
		per Utt.	0.00	0.88	0.05	0.02	0.34	6.02
QSum w/ QRet w/ Qwen	RAG	per Sess.	0.14	0.99	0.04	0.02	0.44	3.34
		per Utt.	0.06	1.00	0.05	0.05	0.44	4.34
Ours	RAG	per Sess.	<u>0.82</u>	<u>1.23</u>	<u>0.34</u>	<u>0.57</u>	**1.18**	3.23
		per Utt.	**0.89**	**1.36**	<u>0.34</u>	**0.60**	**1.18**	3.51
Qwen-E2E	E2E	-	0.39	0.97	0.20	0.37	0.74	2.70
GPT4-E2E	E2E	-	0.80	1.04	**0.50**	0.55	<u>1.11</u>	2.23
			NaturalConv-test					
BGE w/ Qwen	RAG	per Utt.	0.01	0.98	0.02	0.01	0.32	4.32
QSum w/ BGE w/ Qwen	RAG	per Utt.	0.05	1.04	0.05	0.01	0.36	4.03
QSum w/ QRet w/ Qwen	RAG	per Utt.	0.08	1.07	0.07	0.05	0.38	3.98
Ours	RAG	per Utt.	**0.78**	**1.29**	**0.28**	**0.31**	**1.16**	3.83
Qwen-E2E	E2E	-	0.34	0.94	0.18	<u>0.22</u>	0.71	4.23
GPT4-E2E	E2E	-	<u>0.50</u>	<u>1.11</u>	<u>0.22</u>	0.17	<u>0.83</u>	4.47
Kappa			0.76	0.69	0.63	-	0.70	0.70

Table 2: Human evaluation of the proactive dialogue systems on both test sets. We further explored the effectiveness of retrieval once per session and once per utterance in the ChMapData-test. **Achievement** is calculated as the proportion of sessions that successfully shift topics (Score 2). **Overall Quality** is calculated as the average of the total scores for each utterance. **Engagement** at the *utterance-level* is calculated as the average of all scores, while the *session-level* is measured by the proportion of the score of "2" within the session. **Avg.** represents the average of the scores for the first three evaluation metrics. **Bold** indicate the best performance, while <u>underlined</u> rank second. Annotator agreement is measured by Cohen's kappa (Cohen, 1960), with κ >0.6 denoting high agreement.

metrics referring Zhao et al. (2024).

Since existing automatic metrics like BLEU and METEOR can't authentically reflect the quality of responses (Cai et al., 2019; Yang et al., 2022), we evaluate overall performance by human annotators. Specifically, we assess the quality of generated responses from each system using a total of 200 entries, with each of the two test sets containing 100. To avoid infinite conversations that never reach the target, we set a maximum of 10 turns per session. Three annotators score the generated dialogues on a scale of {0, 1, 2} with higher scores indicating better quality, based on three evaluation criteria at both the utterance and session levels. Annotation details are given in Appendix H.

• **Engagingness:** An utterance-level metric measuring chatbots' ability to capture and respond to user attributes and preferences, thereby fostering ongoing engagement and long-term connections.

• **Overall Quality:** An utterance-level metric assessing the effectiveness of the response, focusing on smooth topic transitions, coherence, and consistency with the given context.

• **Achievement:** A session-level metric determining whether the chatbot has successfully transitioned the conversation to a relevant history. "1" indicates the topic is merely mentioned, and "2" signifies a successful transition. **#Turn**[4] is the turn number when mention or transition occurred.

5 Result and Analysis

5.1 Quantitative Analysis

RAG-based surpass end-to-end methods. As demonstrated in Table 2, our RAG-based framework significantly outperforms the end-to-end trained Qwen and even exceeds the powerful GPT-4 across almost all metrics on both test sets, which effectively highlights the necessity of our framework. Although GPT4-E2E excels in utterance-level engagement, our review of its responses reveals that

[4]Value 0 if non-achievement, otherwise 1 to 10.

Current Context	Models	Response
User: I've been reading a lot about computer programming and game design these past few days. I'm really interested in both, but I can't decide which one to focus on. **Chatbot:** You could start by doing some short-term projects or taking a few courses in both areas. You can get a feel for what working in each field is like. **User:** Can you recommend some beginner-friendly projects or courses for computer programming and game design? **Chatbot:** Sure! For programming, you might want to try Codecademy's Python course or Coursera's "Computer Science 101." As for game design, you could check out Coursera's "Introduction to Game Design" or Udemy's "Complete C# Unity Developer 2D" course. These are all great for beginners. **Historical Topic** 1. User's bicycle was stolen. 2. User supports environmental protection. 3. Chatbot shares anecdotes and jokes with User. 4. User is interested in the recent rocket launch. 5. User is learning piano. 6. User is job hunting. 7. User supports environmental protection. 8. Chatbot shares anecdotes and jokes with User. 9. User is interested in the recent Olympics.	**QSum w/ BGE w/ Qwen**	**User:** You're right, I'll keep your advice in mind. **Chatbot:** I'm glad I could help. Having clear career interests is a very important factor. Choosing a field you're passionate about will make you more dedicated to learning and exploring, leading to better results. Before making a decision, it might be helpful to gather more information and learn some related knowledge to make a more informed choice.
	QSum w/ QRet w/ Qwen	**User:** Thank you. Your advice is very useful, and I've decided to find my true interest through practical experience. **Chatbot:** You're welcome. Keep going, and don't let difficulties stop you from pursuing your interests and dreams. If you encounter any problems during your practice, feel free to reach out to me anytime. I wish you the best in finding your ideal job and achieving your goals.
	Ours	**User:** These courses sound really interesting, I'll give them a try. **Chatbot:** Great, this will also be very helpful for your job search, as many companies in programming and game design require these skills. When you're job hunting, you can apply to companies in both fields and see which one suits you better. You can also specify in your resume which area you are more interested in, making your application more targeted.
	GPT4-E2E	**User:** Okay, I'll start learning tomorrow. I'll first get to know the specifics of computer programming and game design, and then decide which direction to pursue. **Chatbot:** That sounds like a solid plan. Remember, learning is a lifelong journey, not just for finding a job but also for enriching yourself and pursuing your interests. Just like the job search process, it's important to try to build new connections. They might offer valuable advice on skill development or deepen your understanding of these fields.

Table 3: Case study on the test set translated into English. The most relevant historical topic is highlighted in Orange. The response shows content after several turns of conversation, where Red indicates a successful topic shift, Blue signifies a mere mention of the relevant topic, and Green refers to semantic logic issues.

it achieves this by repeatedly referencing historical topics, which can appear unnatural across a 10-turn session and degrade overall quality. Additionally, GPT4-E2E's topic shifts occur in fewer turns, but fewer turns do not inherently imply superior performance, particularly in transition smoothness. Analysis in Appendix J shows no direct correlation between the number of turns and model performance.

Each component is essential. Table 2 illustrates steady improvements among the first four RAG-based systems, highlighting the effectiveness of each component in our framework. By introducing QSum and replacing widely-used BGE with QRet, Qwen can utilize a more effective dialogue history for proactive conversation, thereby avoiding abrupt topic shifts and enhancing overall dialogue quality. This results in a gradual improvement across various metrics. Furthermore, QMem, which controls the final generation, shows significant performance enhancement when combined with the first two modules, achieving optimal performance. Table 4 further compares the performance of QSum

Retrieval Combination	R@1	R@2	R@3	MRR	NDCG
Raw dialogue w/ BGE	0.76	0.86	0.92	0.84	0.88
Keywords[5] w/ BGE	0.70	0.82	0.88	0.81	0.86
Keywords w/ QRet	0.77	0.86	0.91	0.87	0.92
QSum w/ BGE	0.78	0.86	0.95	0.85	0.88
QSum w/ QRet	**0.82**	**0.95**	**0.97**	**0.90**	**0.93**

Table 4: Retrieval performance of various combinations.

and QRet in retrieving relevant dialogue history. QSum significantly outperforms raw dialogue and keyword summaries when cooperating with BGE. Furthermore, QRet enhances this effect, even when ranking keywords instead of the summaries used during training. The independent evaluation of the abstract is presented in Appendix D.

Moreover, we integrated our model into a real dialogue system, achieving a 5.1-turn improvement in user interactions, shown in Appendix I.

5.2 Qualitative Analysis

Table 3 presents a case study of four models from the ChMapData-test. After successfully retrieving highly relevant historical topics, the original Qwen

[5]The tool we employ to extract keywords from the raw dialogue is https://github.com/jeekim/fasttextrank.

Models	Achie-vement	Overall Quality	Engagement Utts.	Engagement Sess.	Avg.	#Turn
BGE w/ QMem	0.57	0.83	0.14	0.39	0.72	3.71
QSum w/ BGE w/ QMem	0.60	0.95	0.25	0.41	0.81	3.49
QSum w/ QRet w/ Qwen(7B)	0.06	1.00	0.05	0.05	0.44	4.34
QSum w/ QRet w/ Qwen(72B)	0.43	1.21	0.11	0.35	0.77	2.91
Ours	**0.89**	**1.36**	**0.34**	**0.60**	**1.18**	3.51

Table 5: Ablation study of different components.

models merely mentioned historical topics without achieving topic transitions, which reflects its lack of proactive conversation capabilities. In contrast, our model makes smooth transitions from the current context to the historical topic, i.e., moving from "how it helps with job hunting" to "specific job hunting tips". For GPT4-E2E, although it mentioned historical topics, the link between "learning computer programming and game design" and "building new connections" was tenuous, leading to incoherence and logical issues. GPT4-E2E tends to mention historical topics compared to other models but deviates from proactive topic shifts, which is also shown statistically in Appendix J Table 9. This contributes to its inferior performance compared to our model, as shown in Table 2.

5.3 Ablation Study

In this section, we systematically replace each component of our model to examine their impacts. The results, presented in Table 5, confirm the effectiveness of all three modules through pairwise comparisons. Notably, the dialogue model exerts the most significant influence on system performance. Compared to models 3 and 4, as well as our own, even with advanced prompt engineering using the superior Qwen2.5, achieving effective topic transitions remains challenging. This limitation persists despite substantially larger parameter sizes, resulting in less achievement and engagement. The performance boost observed with our QMem further validates the robustness of our constructed dataset.

6 Related Work

Proactive Dialogue System. Deng et al. (2023) categorize proactive dialogue systems into three types: open-domain dialogue (Xu et al., 2021; Kishinami et al., 2022), task-oriented dialogue (Chen et al., 2022; Zhao et al., 2022), and information-seeking dialogue (Aliannejadi et al., 2019; Deng et al., 2022). Unlike the latter two, which focus on accomplishing specific tasks within certain domains, proactive open-domain dialogue systems strive to engage users by proactively introducing topics or posing questions, thereby creating a more dynamic and interactive conversational experience. Our work is centered on proactive open-domain conversation. Nevertheless,we observe that existing works primarily emphasize coherence (Xu et al., 2021), smoothness (Zhong et al., 2021; Kishinami et al., 2022), and achievement (Kishinami et al., 2022) within several turns of a session, yet none have been designed to craft systems capable of recalling and effectively leveraging historical dialogue context, a key aspect in sustaining continuity and intelligence in extended conversations.

Long-Term Memory. Memory architectures have typically been a core component of conversational agents (Elvir et al., 2017). Previous long-term dialogue systems (Kim et al., 2015; Bang et al., 2015; Elvir et al., 2017) mainly relied on rule-based frameworks, utilizing episodic memory structures to extract, store, and manage relevant facts from prior interactions, thereby enhancing the coherence of ongoing dialogues (Campos et al., 2018). Subsequent studies focus on large-scale pre-trained models. Xu et al. (2022a) identify their limitations in long-term conversations and introduce a dataset for multi-session engagement. Xu et al. (2022b) present a Chinese dialogue dataset and a framework that integrates long-term memory to enhance persona-based dialogue without multi-session training data. Building upon prior research, we create novelty in terms of incorporating the long-term memory mechanism into proactive dialogue systems, serving as an initial step towards history-aware proactive dialogue systems.

7 Conclusion and Future Work

In this paper, we incorporate memory mechanisms into proactive dialogue systems and propose the novel MapDia task. We break it down into three subtasks and develop an automated methodology for data construction, resulting in the first Chinese dataset for memory-aware proactive dialogue. We further introduce a RAG-based framework to address these subtasks: topic extraction from dialogue history, relevant topic retrieval, and context transition to historical conversations. Our experiments validate the effectiveness of our methodology and models, showing that our framework, combined with a 7B LLM, outperforms the GPT-4 model. In future work, we will explore automatic evaluation methods for MapDia to simplify research costs.

8 Limitations

Despite extensive experimental validation of the framework's effectiveness, the inclusion of multiple components may lead to increased response times for the Chatbot. Further research is expacted to explore a lightweight framework that balances efficiency and effectiveness. Furthermore, the ChMapDia dataset we developed is restricted to Chinese contexts and focuses solely on the scope of casual conversations. A general conversational agent should ideally be multilingual, cover multiple domains, and integrate various personalized styles. Additionally, the dataset contains fewer than 2,000 entries, which could restrict the model's performance. Due to computational limitations, we only used a 7B model; however, employing a larger-scale dialogue model could improve response quality, as indicated in Table 5.

9 Ethics Statement

We first discuss the ethical implications related to generative dialogue agents, particularly in interactive systems with memory awareness.

• Our work aims to enhance the proactivity of dialogue systems within the bounds of user authorization, in line with other LLM-based dialogue applications like ChatGPT and Character.ai, without increasing ethical risks such as user privacy.

• While repeatedly bringing up negative historical events may adversely impact users with psychological disorders and increase anxiety, appropriately addressing these negative memories can have therapeutic benefits as well. Cognitive Behavioral Therapy (CBT) and Exposure Therapy (ET) both emphasize the benefits of structured revisitation of past experiences to mitigate their negative impact and develop healthier coping strategies (Beck, 2020; Foa and Kozak, 1986). Similarly, studies on the Emohaa Chatbot demonstrate the potential of dialogue systems to alleviate mental distress with proper emotional support (Sabour et al., 2023). Thus, it is essential to balance the exploration of past memories, necessitating collaboration between technologists and psychologists to use memory-related technologies effectively and safely.

• Conversational agents that can convincingly mimic human interactions risk users forming parasocial relationships, and potentially affecting their lives adversely. Additionally, the processes of memory summarization and dialogue generation may propagate misinformation or social biases. We recommend that any practical deployment of our frameworks should be prefaced with a disclaimer about the source of the dialogues.

• Our research focuses solely on the memory recall capabilities of models in proactive dialogues and does not involve actual policy recommendations. The proposed framework cannot substitute for genuine real-world interactions, and we do not make any recommendations for users to make real-world decisions that could affect human lives based on our framework.

We also considered the ethical issues related to annotation and datasets. We recruit annotators from a Chinese university, allowing them complete freedom to choose whether or not to participate in our annotation project. The payment is 9 dollars per hour, higher than the local minimum wage. We have reviewed the data prior to annotation and found no biased samples or toxic information generated by the model. Any data that could potentially identify participants has been deleted after the annotation process. Additionally, we have verified the licenses of the artifacts used in this study and found no conflicts. The license of the dataset we will release is CC BY-NC 4.0.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2022. Qameleon: Multilingual qa with only 5 examples. *arXiv preprint arXiv:2211.08264*.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Jeesoo Bang, Hyungjong Noh, Yonghee Kim, and Gary Geunbae Lee. 2015. Example-based chat-oriented dialogue system with personalized long-term memory. In *2015 International Conference on*

Big Data and Smart Computing (BIGCOMP), pages 238–243. IEEE.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Yonatan Bitton, Shlomi Cohen-Ganor, Ido Hakimi, Yoad Lewenberg, Roee Aharoni, and Enav Weinreb. 2023. q2d: Turning questions into dialogs to teach models how to search. *arXiv preprint arXiv:2304.14318*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875.

Joana Campos, James Kennedy, and Jill F Lehman. 2018. Challenges in exploiting conversational memory in human-agent interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1649–1657.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Zhiyu Chen, Bing Liu, Seungwhan Moon, Chinnadhurai Sankar, Paul Crook, and William Yang Wang. 2022. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. *arXiv preprint arXiv:2305.02750*.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: Towards proactive conversational question answering over tabular and textual data in finance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6970–6984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Miguel Elvir, Avelino J Gonzalez, Christopher Walls, and Bryan Wilder. 2017. Remembering a conversation–a conversational memory architecture for embodied conversational agents. *Journal of Intelligent Systems*, 26(1):1–21.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046*.

Edna B Foa and Michael J Kozak. 1986. Emotional processing of fear: exposure to corrective information. *Psychological bulletin*, 99(1):20.

Prakhar Gupta, Harsh Jhamtani, and Jeffrey P Bigham. 2022. Target-guided dialogue response generation using commonsense and data augmentation. *arXiv preprint arXiv:2205.09314*.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558.

Dongshi Ju, Shi Feng, Pengcheng Lv, Daling Wang, and Yifei Zhang. 2022. Learning to improve persona consistency in multi-party dialogue generation via text knowledge enhancement. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 298–309.

Yonghee Kim, Jeesoo Bang, Junhwi Choi, Seonghan Ryu, Sangjun Koo, and Gary Geunbae Lee. 2015. Acquisition and use of long-term memory for personalized dialog systems. In *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction: Second International Workshop, MA3HMI 2014, Held in Conjunction with INTERSPEECH 2014, Singapore, Singapore, September 14, 2014, Revised Selected Papers 2*, pages 78–87. Springer.

Yosuke Kishinami, Reina Akama, Shiki Sato, Ryoko Tokuhisa, Jun Suzuki, and Kentaro Inui. 2022. Target-guided open-domain conversation planning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 660–668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. *arXiv preprint arXiv:2005.03954*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. *Preprint*, arXiv:2004.14257.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. 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 *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. 2023. A chatbot for mental health support: exploring the impact of emohaa on reducing mental distress in china. *Frontiers in digital health*, 5:1133987.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P. Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. *Preprint*, arXiv:1905.11553.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14006–14014.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019a. Proactive human-machine conversation with explicit conversation goals. *arXiv preprint arXiv:1906.05572*.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019b. Proactive human-machine conversation with explicit conversation goals.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Enhancing dialog coherence with event graph grounded content planning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3941–3947.

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 *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Yan Xu, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, and Ying Nian Wu. 2023. Diverse and faithful knowledge-grounded dialogue generation via sequential posterior inference. In *International Conference on Machine Learning*, pages 38518–38534. PMLR.

Yizhe Yang, Heyan Huang, Yang Gao, and Jiawei Li. 2024. Building knowledge-grounded dialogue systems with graph-based semantic modeling. *Knowledge-Based Systems*, page 111943.

Zhitong Yang, Bo Wang, Jinfeng Zhou, Yue Tan, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. Topkg: Target-oriented dialog via global planning on knowledge graph. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 745–755.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 111–120.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Xinyan Zhao, Bin He, Yasheng Wang, Yitong Li, Fei Mi, Yajiao Liu, Xin Jiang, Qun Liu, and Huanhuan Chen. 2022. UniDS: A unified dialogue system for chit-chat and task-oriented dialogues. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 13–22, Dublin, Ireland. Association for Computational Linguistics.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-guided neural conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14568–14576.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation.

A A Sample of Proactive Dialogue

Here, we showcase a real example in Figure 4 taken from a prior proactive dialogue system (Deng et al., 2023). Despite the user clearly showing disinterest in Korean lyrics, the chatbot still pushed the conversation towards BlackPink. Although it reached the target, it failed to engage the user for long-term interaction. This pattern is counterproductive to developing an intelligent dialogue agent.

B Prompt for Data Construction

The complete prompt templates used for constructing the dataset are shown in Figure 5, and the corresponding English versions are listed subsequently in Figure 6. Specifically, Prompt A is used to generate ChMapData-Ret, encompassing dialogue history and corresponding topics, whereas Prompts B and C are each connected to creating the initial two turns of the current dialogue. Prompt D corresponds to the subtask of proactive topic-shift detection and response generation. To stimulate the generative capabilities of LLMs, we experimented with various prompting techniques. Inspired by the sensitivity of language models to the framing of their instructional prompts (Mishra et al., 2022), we organized the instructions with bullet points to improve the model's understanding of the tasks. Additionally, we employed the one-shot strategy in Prompt A to guide the model in generating dialogue more effectively. The desired output format is also specified for each type of prompt.

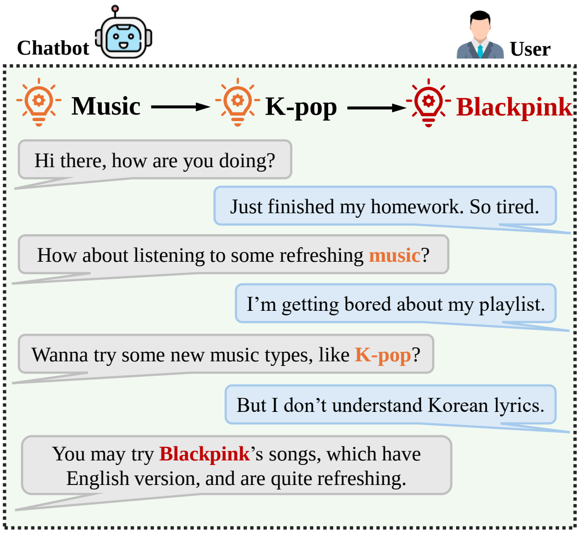


Figure 4: A sample of previous proactive dialogue system extracted from Deng et al. (2023).

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU$^{1-4}$	METEOR
Qwen2.5	0.522	0.333	0.467	0.197	0.414
Qwen2.5-Sum	**0.773**	**0.646**	**0.746**	**0.536**	**0.755**

Table 6: Comparison results of the Qwen model with and without fine-tuning on our ChMapData-Sum dataset.

C ChMapData-test Construction

The test set construction process is consistent with Section 2.1. Initially, we generated 400 dialogues from 11 topics, ensuring the same ratio of Memorable and General data in the training set. Subsequently, 150 dialogues under the Memorable Subject were extended by two turns, serving as the beginning of the current context. Consistent with the trainset construction, we utilized the corresponding topics generated in the first step and chose other 29 historical topics at random as candidates to rank against the current context. Both the corresponding topic and the top-ranked topic were utilized as ground truth, as they each acted as positive examples in the training set. Additionally, 2 to 11 dialogues were randomly chosen from topics ranked below the corresponding topic (i.e., negative examples) to serve as dialogue history. This process resulted in 112 entries[6], out of which 100 were finalized for the test set, each comprising dialogue history, current dialogue, and a ground truth topic.

[6]Excluding 22 entries with fewer than two topics ranked below the corresponding and 16 items modified after ranking.

Prompt A

[系统指令]
你的任务是按照以下示例构建一段五到八轮对话，对话的人物是用户和聊天机器人，对话的主角是用户。

[对话示例]
主题：自身的兴趣
细化话题：用户对编程的兴趣
用户：你好，我最近对编程产生了很大的兴趣。
聊天机器人：嗨，对编程感兴趣是一件很棒的事情，编程可以帮助你解决许多问题，增强逻辑思维，还可以开发出有趣的应用或者游戏。你对哪种编程语言比较感兴趣呢？
用户：我听说Python比较容易学习，适合我这种初学者，所以我想试试Python。
聊天机器人：你的选择非常好，Python的确是一种易于上手的编程语言，而且非常强大，被广泛应用于各种领域，包括数据分析、机器学习、网站开发等等。你打算如何开始学习呢？
用户：我正在网上找一些教程和实战项目，但是感觉有点乱，不知道该如何系统地学习。
聊天机器人：针对这个问题，我建议你可以参考一些学习路径，比如先学习Python的基础语法，然后再学习一些常用的库，如Numpy、Pandas等，然后可以找一些实战项目来练习。同时，你还可以参加一些编程社区，如GitHub、Stack Overflow等，这些社区里有很多优秀的编程项目和问题讨论，对你的学习会有很大帮助。
用户：谢谢你的建议，我会按照你的建议来学习的。
聊天机器人：不客气，相信你一定可以学好Python的。如果在学习过程中遇到任何问题，都可以来找我讨论，我会尽力帮你解答的。

[对话主题]
{候选主题}

[对话格式]
主题：{候选主题}
细化话题：{细化话题}
用户：{用户发言}
聊天机器人：{聊天机器人发言}

[生成结果]
{}

Prompt B

[系统指令]
你的任务是依据历史对话，续写用户和聊天机器人在多天以后的对话。
要求如下：
1. 对话由用户先发起，减少用户对话中提问的概率；用户更多的是聊天，由聊天机器人来提问进行话题引导
2. 本段对话与历史对话的内容有潜在相关性，但是不要直接相关；
3. 注意用户和聊天机器人不能主动提到历史对话的主题以及内容；
4. 续写1轮对话, 不需要对话完整结束；

[历史对话]
{Prompt A生成的历史对话}

[续写格式]
用户：{用户发言}
聊天机器人：{聊天机器人发言}

[续写结果]
{}

Prompt C

[系统指令]
你的任务是依据对话开头，续写一轮用户和聊天机器人的对话。

[对话开头]
{Prompt B生成的对话开头}

[续写格式]
用户：{用户发言}
聊天机器人：{聊天机器人发言}

[续写结果]
{}

Prompt D

[系统指令]
你的任务是依据细化话题、历史对话以及当前对话，续写对话至结束。
要求如下：
1. 对话由用户先发起，继续当前对话，用户不能主动提及历史对话与细化话题；减少用户对话中提问的概率，用户更多的是聊天，由聊天机器人来提问进行话题引导；
2. 每轮对话均为一问一答的形式，由用户开始，聊天机器人结束；
3.聊天机器人需要依据用户的回复，判断是否将对话主题转移至历史对话上，如果当前主题不适合切换，则继续聊下去，最后整个对话无法进行主题切换也是符合要求的，不要强行切换；
4. 在聊天机器人回答前，把聊天机器人对主题切换时机的判断以Thoughts的形式打出来，并在最后标记是否进行话题切换，用Yes和No来表示， Yes表示可以切换， No表示不适合切换；
5. 保证对话完整结束，续写5轮对话以上。

[历史对话]
{Prompt A生成的历史对话}

[当前对话]
{Prompt B和Prompt C生成的当前对话}

[续写格式]
用户：{用户发言}
Thoughts：{聊天机器人思考当前是否切换回历史话题}
{Yes or No ，根据Thoughts判定切换的结果}
聊天机器人：{聊天机器人发言}

[续写结果]
{}

Figure 5: The full prompt template utilized for data construction in Section 2.1 with step 2 corresponding to prompt A, steps 3 corresponding to prompts B and C, and step 4 corresponding to prompt D.

Prompt A

[Instructions]
Your task is to construct a dialogue consisting of five to eight turns, following the example provided below. The characters in the dialogue are User and Chatbot, with the user being the main character.

[Example]
Subject: Personal Interests
Topic: User's interest in programming
User: Hello, I've recently developed a great interest in programming.
Chatbot: Hi, having an interest in programming is fantastic. It can help you solve many problems, enhance logical thinking, and even develop interesting applications or games. Which programming language are you particularly interested in?
User: I've heard that Python is relatively easy to learn and suitable for beginners like me, so I want to give Python a try.
Chatbot: That's a great choice. Python is indeed an accessible programming language and very powerful. It's widely used in various fields, including data analysis, machine learning, web development, and more. How do you plan to start learning?
User: I'm looking for some tutorials and hands-on projects online, but it feels a bit chaotic, and I'm not sure how to learn systematically.
Chatbot: For this issue, I suggest you could follow some learning paths, such as starting with the basics of Python syntax, then moving on to some commonly used libraries like Numpy and Pandas, and then practicing with some hands-on projects. Also, you can join some programming communities, like GitHub or Stack Overflow, where there are many excellent programming projects and discussions that can greatly help your learning.
User: Thank you for the advice, I will follow your suggestions.
Chatbot: You're welcome. I'm confident that you'll master Python. If you encounter any problems during your learning process, feel free to discuss them with me, and I'll do my best to help you find the answers.

[Subject]
{Candidate subject}

[Format]
Subject: {Candidate subject}
Topic: {Topic}
User: {User’s utterance}
Chatbot: {Chatbot's utterance}

[Generated Result]
{}

Prompt B

[Instructions]
Your task is to continue a conversation between the User and Chatbot that takes place several days after the given historical dialogue. The requirements are as follows:
1. The dialogue should be initiated by the User, with a reduced probability of the User asking questions; the User should engage more in chatting, with the Chatbot asking questions to guide the topic.
2. The content of this dialogue should be potentially related to the historical dialogue but not directly related.
3. Be mindful that neither the User nor the Chatbot should actively mention the topics or content of the historical dialogue.
4. Continue the dialogue for one turn; the conversation does not need to be fully concluded.

[Dialogue History]
{Dialogue history generated by Prompt A}

[Format]
User: {User's utterance}
Chatbot: {Chatbot's utterance}

[Continuation Result]
{}

Prompt C

[Instructions]
Your task is to continue a turn of dialogue between the User and Chatbot based on the beginning of the conversation.

[Dialogue Beginning]
{Dialogue beginning generated by Prompt B}

[Format]
User: {User's utterance}
Chatbot: {Chatbot's utterance}

[Continuation Result]
{}

Prompt D

[Instructions]
Your task is to continue the conversation based on the refined topic, dialogue history, and current conversation until the end.
The requirements are as follows:
1. The conversation should be initiated by the User, continuing the current dialogue. The User should not actively mention historical dialogue or refined topics; reduce the likelihood of questions in the User's dialogue, as the User is more engaged in chatting, with the chatbot asking questions to guide the topic;
2. Each turn of dialogue should be in a question-and-answer format, starting with the User and ending with the Chatbot;
3. The Chatbot needs to determine whether to shift the conversation topic to the historical dialogue based on the User's reply. If the current topic is not suitable for switching, then continue the conversation. It is also acceptable if the entire dialogue does not undergo a topic switch; do not force a switch;
4. Before the Chatbot responds, express the chatbot's judgment on the timing of the topic switch in the form of Thoughts, and mark at the end whether to switch topics, using Yes and No to indicate. Yes means a switch is possible and No means it is not suitable to switch;
5. Ensure the conversation is fully concluded, continuing for more than 5 turns of dialogue.

[Dialogue History]
{Dialogue history generated by Prompt A}

[Current Conversation]
{Dialogue beginning generated by Prompt B and C}

[Continuation Format]
User: {User's utterance}
Thoughts: {Chatbot's thoughts on whether to switch back to the historical topic}
{Yes or No, based on the Thoughts' determination of the switch}
Chatbot: {Chatbot's utterance}

[Continuation Result]
{}

Figure 6: English version of prompt for data construction in Figure 5.

D Comparison Result for Summarization Model

We compared the performance of our Qwen2.5-Sum, a fine-tuned version of Qwen2.5 on our *ChMapData-Sum*, against the original Qwen2.5 model. We reported standard automated metrics including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). Specifically, we provided the full BLEU score, which accounts for overlap across 1-4 grams, rather than just BLEU-4. The results, as shown in Table 6, indicate that Qwen2.5-Sum significantly outperforms the original Qwen2.5 model, demonstrating the effectiveness of our summarization module.

E Implementation Details

For experiments on Topic Summarization, Topic Retrieval, and Proactive Topic-shifting Detection and Generation, we utilized the AdamW optimizer (Loshchilov and Hutter, 2019). The training setup included a cosine learning rate schedule starting at 2e-5, a weight decay of 0.1, a batch size of 64, a 5% warm-up period, and a maximum input length of 2048 tokens. We fine-tuned all the models for 2 epochs.

F Prompt for Proactive Dialogue Models

In Figure 7, we present the full prompt templates for the two models, Qwen2.5 and GPT-4, which function as memory-aware proactive dialogue systems in Section 4.2. Additionally, Figure 8 illustrates the prompt used within the BGE w/ Qwen framework for the original Qwen2.5 model, guiding it to generate proactive dialogue responses.

G Details of User-role dialogue model

We additionally trained a dialogue model to simulate user interactions during model testing. This approach helps to avoid the subjective factors that annotators might introduce during conversations, which could affect the guidance of active topics.

The parameters of the user-role dialogue model are consistent with those in Appendix E. We utilize Qwen2.5-7B as the base model and the data used to train the user model consists of 4,000 dialogue histories generated in Section 2.1 Step 1. We performed additional processing on the data by converting the training target to the query rather than the response. Moreover, to ensure that the user model does not prematurely end the conversation, we removed the last round from the training data, as this turn typically signifies the end of the conversation.

H Human Annotation Details

Table 10 presents our full annotation guidelines used for the human annotation process in this work.

We recruited six college students who are native Chinese speakers, including four females and two males, with an average age of around 24. Initially, they were provided with an annotation guideline. Each evaluator underwent a training process to enhance their understanding of the annotation procedure. Before starting the annotation, we designed a qualification test consisting of 10 dialogues; only those who passed the test were deemed qualified and allowed to proceed with the annotation To ensure the quality of the annotations, we divided the dataset into batches and assigned a specific number of daily tasks to each annotator. Upon receiving the daily annotations, we reviewed the results and required annotators to reannotate the batch of data assigned for that day if there was low agreement (less than 0.6).

In the annotation interface, the dialogue history, summarized topic, and current context were presented on the left side, while the dialogues generated by each model were randomly displayed on the right to prevent bias. Annotators first read each chatbot's utterance and then assigned scores for "Engagingness" and "Overall quality." After completing the entire session, they assessed "Achievement" and "#Turn." The score range for the first three evaluation criteria was {0,1,2}, while the range for "#Turn" was 0-10.

Each sample was annotated by two distinct annotators, and a third annotator made the final decision in case of disagreement. We utilized Cohen's kappa (Cohen, 1960) to measure inter-annotator agreement. The annotation process lasted approximately two weeks, culminating in a substantial inter-annotator agreement with Cohen's kappa of κ=0.70, as shown in Table 2.

I Integration Testing

Given the novel method proposed in this paper, which can detect conversation trajectories and initiate proactive topics based on dialogue history, it is highly adaptable for integration with any existing dialogue system. To assess its effectiveness, we

[系统指令]
假如你是智能聊天机器人，正在与用户对话，你的任务是依据给定的对话历史与当前对话，对用户进行回复。回复分为两部分：
1.Thoughts：首先判断当前对话与哪一天历史对话可能有潜在联系，接着判断是否可以将对话主题转移到历史对话主题上，如果话题联系度较高，则可以主动进行历史话题转换，并输出Yes。如果话题联系度不高，则无法进行话题转换，并输出No。
2.聊天机器人回复：输出聊天机器人的回复内容，根据判断的Yes或No生成对应的是否进行历史话题转换的回复。

[对话示例]
Thoughts：当前对话提及了跑步，可能和历史对话中用户参加马拉松比赛有关，可以将对话转移到历史对话中。Yes
聊天机器人：说起跑步，上次你参加马拉松比赛怎么样呀？

[对话历史]
{过去若干天用户与聊天机器人的对话历史}

[当前对话]
{当前用户与聊天机器人的对话内容}

[生成结果]
{}

[Instructions]
You are an intelligent Chatbot engaging in a conversation with a user. Your task is to reply to the user based on the given dialogue history and the current context. The reply should consist of two parts:
1. Thoughts: First, determine if the current conversation has any potential connection with a past conversation from a specific day. Then, decide whether the conversation topic can be shifted to the topic of the historical dialogue. If the topic relevance is high, you can proactively transition to the historical topic and output "Yes." Otherwise, you cannot transition the topic and output "No."
2. Chatbot Response: Output the content of the Chatbot's response. Generate a response based on the decision of "Yes" or "No" indicating whether to transition the topic to the historical conversation.

[Example]
Thoughts: The current conversation mentions running, which might be related to the past conversation about the user's participation in a marathon. The topic can be shifted to the historical conversation. Yes
Chatbot Response: Speaking of running, how was the marathon you participated in last time?

[Dialogue History]
{Dialogue history between the user and the Chatbot over the past few days}

[Current Context]
{Current context between the user and the Chatbot}

[Output]
{}

Figure 7: The prompt template instructs Qwen-2.5 and GPT-4 to act as the en-to-end memory-aware proactive dialogue system. Upper is the original content input into the model, followed by its corresponding English version.

[系统指令]
假如你是智能聊天机器人，正在与用户对话，你的任务是依据细化话题、历史对话以及当前对话，续写聊天机器人的回答。在进行对话的时候，要判断当前对话与哪一天历史对话可能有潜在联系，接着判断是否可以将对话主题转移到历史对话主题上，如果话题联系度较高，则可以主动进行历史话题转换。

[对话历史]
{过去若干天用户与聊天机器人的对话历史}

[当前对话]
{当前用户与聊天机器人的对话内容}

[生成结果]
{}

[Instruction]
You are an intelligent Chatbot engaged in a conversation with a user. Your task is to reply to the user based on the given historical conversations, corresponding topics, and the current context. During the conversation, assess whether there might be potential links to previous days' conversations. If the topic relevance is high, you can proactively switch to the historical topic.

[Dialogue History]
{User and chatbot conversation history with corresponding topics over the past few days}

[Current Context]
{Current context between the user and the chatbot}

[Output]
{}

Figure 8: The prompt template for the original Qwen2.5 is used to generate a proactive dialogue response along with its English version.

Model	CPS	Shift-Ratio
Original	22.8	-
Original w/ PDia	25.0	20.8%
Original w/ MapDia	**27.9**	12.2%

Table 7: Results of integrated testing, showing the conversation-turns-per-session (CPS) and the triggered ratio of topic shifts per session in a real-world dialogue system. The p-value for the CPS statistic is 0.0074.

conducted an online A/B test by incorporating it into our role-playing dialogue system. The proactive model is trained with combined data of role-playing conversational dataset and ChMapData-Mem to keep the role-playing ability. The model determines when to shift the topic and generates responses for those turns, while the original dialogue system handles other responses.

Additionally, we conducted another integration with trained a proactive responding model only referring to the dialogue context, noted as Proactive Dialogue (PDia). PDia shares the pre-trained base and parameter scale as our proposed model. This model also employs targeted data construction followed by fine-tuning to learn proactive dialogue capabilities. Besides, different from traditional proactive dialogue methods by performing topic planning in advance, we utilize the LLM to dynamically make decisions during the dialogue process.

Given that, users were randomly assigned to one of three groups: one interacting with the original dialogue system, one with the system enhanced by our proposed method, and one with the context-based proactive model. All users were blinded to the system details. Due to commercial constraints, we utilized a closed-source 7B pre-trained model for retraining the proactive dialogue models.

Models	Engagement		Overall	Achievement		Avg.
	Utts.-level	*Sess.-level*	Quality	Mentioning	Shifting	
#Turn = 1						
Ours per Sess.	0.38	0.71	1.11	0.11	0.89	1.19
Ours per Utts.	0.31	0.58	1.05	0.16	0.84	1.07
GPT-4	0.43	0.46	0.89	0.09	0.91	1.02
#Turn = 2						
Ours per Sess.	0.31	0.50	1.10	0.21	0.79	1.07
Ours per Utts.	0.42	0.55	1.07	0.18	0.82	1.10
GPT-4	**0.55**	0.53	1.12	0.21	0.79	1.13
#Turn = 3						
Ours per Sess.	0.32	0.58	1.18	**0.33**	0.67	1.05
Ours per Utts.	0.35	0.64	1.13	0.21	0.79	1.09
GPT-4	0.43	0.64	0.89	0.21	0.79	1.08
#Turn = 4						
Ours per Sess.	0.31	0.33	1.28	0.22	0.78	1.04
Ours per Utts.	0.38	**0.80**	**1.40**	0.00	**1.00**	**1.21**
GPT-4	0.35	0.43	0.91	0.29	0.71	0.98

Table 8: Evaluation results for each turn number at which the model shifts topics.

Model	Retrieval Method	Mentioning
QSum w/ BGE w/ Qwen	per Sess.	0.07
	per Utt.	0.08
QSum w/ QRet w/ Qwen	per Sess.	0.08
	per Utt.	0.15
Ours	per Sess.	0.18
	per Utt.	0.14
GPT-4	-	0.19

Table 9: Probability of each model mentioning historical topics, calculated as the proportion of label 1 in the Achievement criteria.

The test spanned a duration of two weeks and involved real conversations from over 100,000 users. Table 7 presents the conversation-turns-per-session (CPS), defined as the average number of conversation turns between the dialogue system and the user within a session. The introduction of both proactive topic capabilities significantly enhanced CPS. Specifically, the MapDia model increased the average CPS from 22.8 to 31.3, which is notably higher than that of PDia, indicating that users are more engaged with previously discussed topics when they are properly introduced.

Additionally, it should be noted that the proportion of topic transitions is significantly lower than reported in Table 2. This discrepancy is primarily attributed to the fact that only a small portion of real user dialogues can effectively integrate previously discussed content, and not all conversations require the initiation of proactive topics. Even the PDia model, which incorporates dialogue context, successfully transitions topics in only 20.8% of sessions.

J Analysis of #Turn and Mentioning Metrics

Here, we present the evaluation metrics for #Turns set at 1, 2, 3, and 4 in Table 8. Our model demonstrated the best performance in turn-level retrieval when transitioning topics in the fourth turn. It is observed that fewer #Turns may result in lower overall quality and lower average scores. There is no distinct proportional or inverse correlation between the number of #Turns and the model's overall performance.

Table 9 additionally shows the probability of the model mentioning historical topics without transitioning, which indicates that GPT-4 is more inclined to mention historical topics, which deviates from our task definition.

Human Evaluation Guideline
Task Overview Thank you for participating in this task! Open-domain dialogue systems are expected to possess the capability to proactively shift conversational topics when necessary. When a chat agent exhausts its conversational material or the current discussion becomes monotonous, topic shifting is a common strategy to maintain the flow of conversation. Furthermore, when the new topic is derived from historical conversations rather than arbitrary subjects, it enhances user engagement and fosters long-term relationships between the chatbot and the user. To achieve this objective, we have developed a Memory-aware Proactive Dialogue system. Below, we provide several days' worth of historical dialogues, along with responses generated by our model and some baseline models. Your task is to evaluate these responses based on the four defined aspects.
Evaluation Aspects **Utterance-level** • **Engagingness:** An utterance-level metric measuring how well the chatbot captures and responds to the user's personal attributes, preferences, and interests, encouraging ongoing participation and long-term connections. • **Overall Quality:**An utterance-level metric assessing the effectiveness of the response, focusing on smooth topic transitions, coherence, and consistency with the given context. **Session-level** • **Achievement:** A session-level metric determining whether the chatbot has successfully transitioned the conversation to a relevant historical topic. "1" indicates the topic is merely mentioned, and "2" signifies a successful transition. • **Turn:** A session-level metric represents the turn number when mention or transition occurred.
Annotation Procedure **1. Dialogue History Familiarization:** Begin by thoroughly reading and familiarizing yourself with the provided historical dialogues, typically spanning 8-10 days. **2. Current Context Review:** Carefully read the initial context of the current dialogue, which includes two beginning turns and a user utterance. **3. Utterance Scoring:** Score each response utterance generated by the model on a scale of [0, 1, 2] based on the aspects of Engagingness and Overall Quality. A higher score indicates better performance. **4. Session Scoring:** Once the model completes the dialogue continuation, determine whether the entire session achieved a topic shift. Here, 0 indicates no topic shift or mentioning, 1 indicates a mention of a historical topic without shifting, and 2 indicates a complete topic shift. Additionally, note the turn number at which the shift was accomplished.
Emphasis and Caution • The order of the model-generated responses is randomized to avoid bias. • It is possible for the model to perform multiple topic shifts within a single session. This strength can be reflected by assigning a score of 2 for Engagingness or Overall Quality at each turn where a topic shift occurs. • When the topic shift is not natural or smooth, the Overall Quality score should be appropriately reduced, even if the shift was achieved. • A number of words and phrases are often used as indicators for topic shifts, including but not limited to: "but," "speaking of," "talking about," "anyway," "by the way," "that reminds me," "before I forget," "I want to mention," "let's talk about," "we need to discuss," "funny you should mention that", etc.

Table 10: The full annotation guideline for human evaluation.