# CareAgent: In Our Next Session, Let Me Actively Care for You

**Anonymous ACL submission**


## Abstract

Despite the fact that the large language models (LLMs) facilitate conversations with rationalization and knowledge, they are still restricted to the passive response mechanism that relies on user instructions, which are hard to actively care for the user. To realize the actively caring ability, we propose an active conversational agent (ACA) named CareAgent which creates new session to approach user potential interest. Specifically, inspired by the Jung's theory of psychological types and the active exploration mechanism of agent in environment, we designed three components to support the goal of caring for user. The Character Extractor (CE) obtains the personality through Myers-Briggs Type Indicator (MBTI) and attributes for character descriptions; the Memory Reconstructor (MR) achieves multi-topic summaries based on multi-clue branching for complete memories; the Decision Adapter (DA) selects the best topic summary as the background memory and adapts the agent intention to control the scenario of new session. The results of experiments demonstrated that CareAgent was able to maintain reliability in character understanding and extract complete multi-topic summaries from conversational history. Evaluators also believed that this agent enhanced the actively caring and personification level in new session.


## 1 Introduction

The realization of highly autonomous intelligent system in conversation has long been a central object in the field of AI(Turing, 1990; Park et al., 2023; Wang et al., 2023a). Large language models (LLMs) (Ouyang et al., 2022; OpenAI, 2023; Touvron et al., 2023a,b) represented by ChatGPT[1] revolutionized intelligent systems with clear logical response, specialized domain knowledge and adaptability for complex tasks, which not only propelled this field into the era of multi-billion-parameter models but also provided a novel platform for constructing conversational agents. In order to fully utilize LLMs to actively care for users, it is crucial to build an active conversational agent(ACA), rather than traditional conversation models with passive response mechanism, to start and control the new session with caring utterances that are tailored for user expectations.

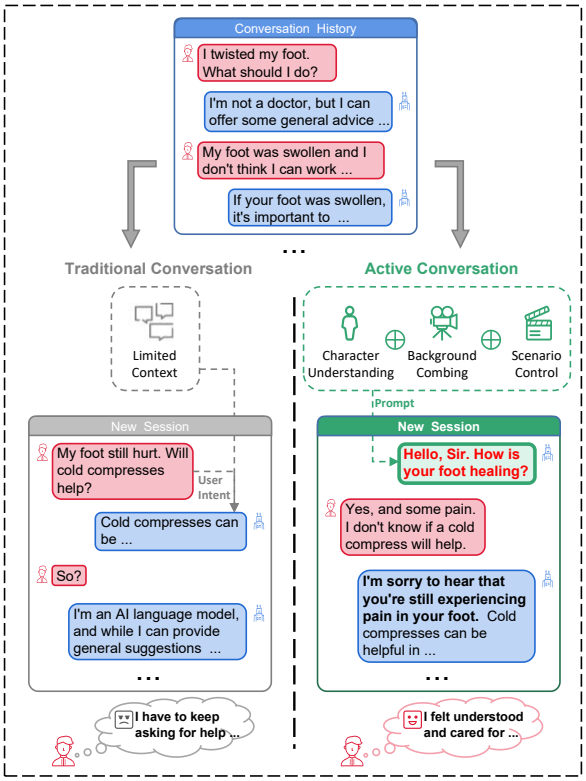


Figure 1: Representative example for explaining difference between Traditional Conversation (left) and Active Conversation (right)

Characters were essential for agent to select key memory and action. Park et al. (2023) and Wang et al. (2023b) assigned each agent with complex descriptions to enable distinct speech and behavior patterns. However, character attributes defined in these works were not sufficiently distinctive and

[1] https://openai.com/blog/chatgpt/

systematic to represent all dimensions of character personality. Thus, to implement the actively caring ability of agent to user, *the primary challenge lies in how to extract and establish abstract character descriptions for both user and agent*.

Conversational memories served as the foundation for the agent to understand the current session. While key words(Zhang et al., 2020a), slot value(Peng et al., 2021) or single summary(Liu et al., 2021) were the basic elements for controlling local coherence in the session context, they struggled to adapt to long-term conversational history composed of complex topics, leading to confusion or loss of semantics. Therefore, *the second challenge in caring for user is how to reconstruct complete conversational memories for all topics*.

Based on the above analysis, we believe that active conversation requires the understanding of character, the setting of background and the control of scenario (Figure 1). To this end, we propose CareAgent which portrays the abstract contour of new session without user instructions to obtain the session control prompt (SCP) by characters, background and scenario. The SCP is submitted to the LLM to generate caring utterances of new session and guide the specific directions of subsequent dialogues. The contribution of this paper includes:

- For caring user, we explored the active conversational framework that generated the initial utterances to start new session which approached user potential interest by constructing the session control prompt (SCP).
- To support the construction of SCP. We proposed a character extractor (CE) to build character descriptions of user and agent; a memory reconstructor (MR) to obtain multi-topic summaries for background combing; a decision adapter (DA) to determine the optimal memory as background and intention as scenario.
- Experimental results demonstrated that our agent outperformed traditional models in terms of character reliability, memory completeness, overall conversational quality and actively caring ability.

## 2 Related Work

**Character Understanding and Description**. Existing methods can be divided into two categories: implicit character extraction and explicit character description. The essence of implicit character extraction is to obtain attributes encoding, Jang et al. (2022) and Fu et al. (2022) encoded character attributes and related knowledge to create a conversational context representation that guides the selection of character and knowledge in the response. Wen et al. (2021) and Mo et al. (2021) tied user character and sentiments together to predict emotion based on the encoding of conversational history and character attributes. Furthermore, the essence of explicit character description lies in defining or augmenting the expression of character. Cao et al. (2022) constructed character description through entity replacement and data matching, while Kim et al. (2022) further enriched character descriptions with common knowledge in GPT-2(Radford et al., 2019) and evaluation model. However, above models may have the disadvantages of instability, potentially conflicting in character description and failure in innate trait extraction of character.

**Conversational Memory**. The issue of reconstructing conversational memory can be distinguished into two parts: memory extraction and memory organization. For memory extraction, Wu et al. (2021) and Liu et al. (2021) segmented the conversation into paragraphs and generated summary to obtain memory. Lin et al. (2022) constructed memories from the perspectives of user and agent with semantic similarity. The above models captured summary from conversational history as memory, but they may suffer from loss and confusion of semantic due to the complexity of topics in real-life. To ensure the completeness of memory, it is also crucial to build memory organization. Park et al. (2023) designed the memory stream to guide the action of each agent by retrieve mechanism, reflect mechanism and memory importance scores. Although memory stream improved the completeness of memory, its single-threaded memory chain may also cause details overlooked due to the lack of multi-topic combing.

## 3 Method

### 3.1 Task Definition

In this work, we formalize the notion of active conversation as follows: given a conversational history $C = (S_1, S_2, \ldots, S_D)$ composed of sessions $S$ collected within $D$ days, the caring utterances $N$ are generated without user instructions to create a new session through large language model $llm$ based

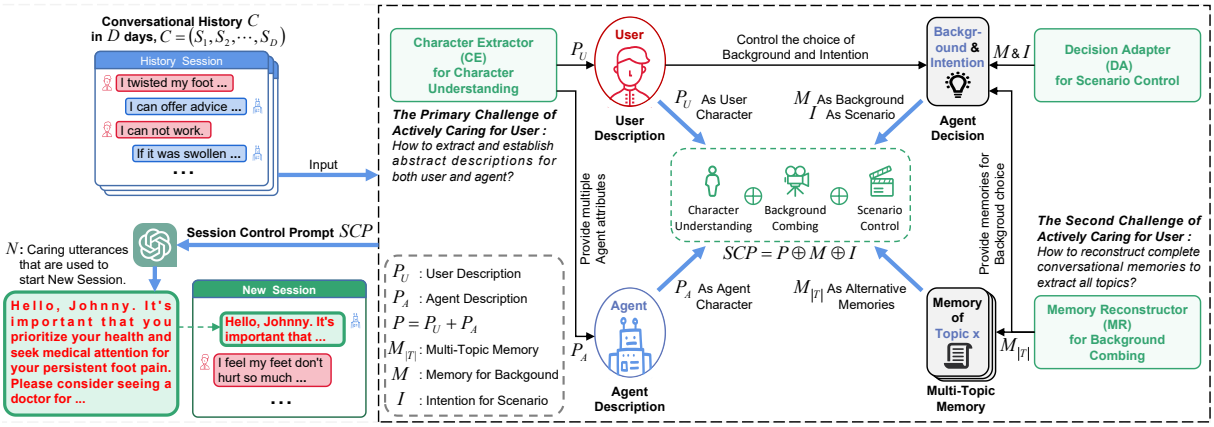


Figure 2: Main framework of active conversational agent named CareAgent. Starting from the conversational history (top left), CareAgent generates the caring utterances of new session (bottom left) that approaches user's potential interest through Session Control Prompt $SCP$ which relies on key components (three light green blocks in the right), including the Character Extractor $CE$, the Memory Reconstructor $MR$ and the Decision Adapter $DA$.

on the session control prompt $SCP$.

$$\begin{aligned} N &= llm(SCP) \\ SCP &= P \oplus M \oplus I \\ P &= P_U \oplus P_A \end{aligned} \tag{1}$$

To express active concern for user, $SCP$ should be made up of three components (equation 1): the character descriptions $P$ that includes user $P_U$ based on personality and agent $P_A$ based on prefabricated attributes $Attr$; the background memory $M$ selected from summaries $M_{|T|}$ of $|T|$ topics; the agent intention $I$ for scenario based on user $P_U$ and background memory $M$, as shown in Figure 2.

$$\begin{aligned} P_U = CE(C),\ P_A = CE(Attr) \\ M_{|T|} = MR(C) \\ M = DA(P_U, M_{|T|}),\ I = DA(P_U, M) \end{aligned} \tag{2}$$

Where, the Character Extractor $CE$ is designed to transform the character understanding object into the user Myers-Briggs Type Indicator (MBTI) classification and the generation of character descriptions; the Memory Reconstructor $MR$ is designed to structure the conversation combing object as the multi-topic summaries generation; the Decision Adapter $DA$ is designed to transform session control object into the optimal memory choice and the inference of agent intention.

### 3.2 Character Extractor ($CE$) for Character Understanding

For the primary challenge, it is crucial to extract the innate trait(Park et al., 2023) which implies user deep-seated attributes and is significant for the agent to understand the motivations and preferences of user. Here, we use MBTI as user innate trait since it has been widely adopted as the standard for personality classification.

Our MBTI personality classification is shown in Figure 3, user utterances are extracted from the conversational history and randomly sampled as input for the 4-dimensional binary classifiers, each classifier consist of a BERT(Devlin et al., 2019) encoder and a linear classification layer, to achieve 4-dimensional labels in MBTI, i.e., 16 possible personalities (implicit character). Each personality is related to an personality description [Personality_Descrip] which explicitly expresses its innate trait. Taking the dimension of Focus[2] preference as an example, the output of the binary classifier could be label: Introversion (I) or Extraversion (E). Based on the concatenation of the outputs of the 4-dimensional binary classifiers, implicit personality labels such as ISTP, ENFP or INTJ of MBTI were obtained. Finally, we combine user name [User_Name], identity [User_Identity], MBTI personality label and personality descriptions [Personality_Descrip] into user descriptions (explicit character) based on the template (Figure 10 (up)). Related details is illustrated in Appendix A.

Furthermore, the character of the agent also determines the memory exploration and intention selection. Therefore, this work prefabricates a variety of structured agent character attributes with template (Figure 10 (bottom)), including

[2]This work refers to the four dimensions of the MBTI as Focus preference, Perceptive preference, Judgement preference and Cognitive preference.

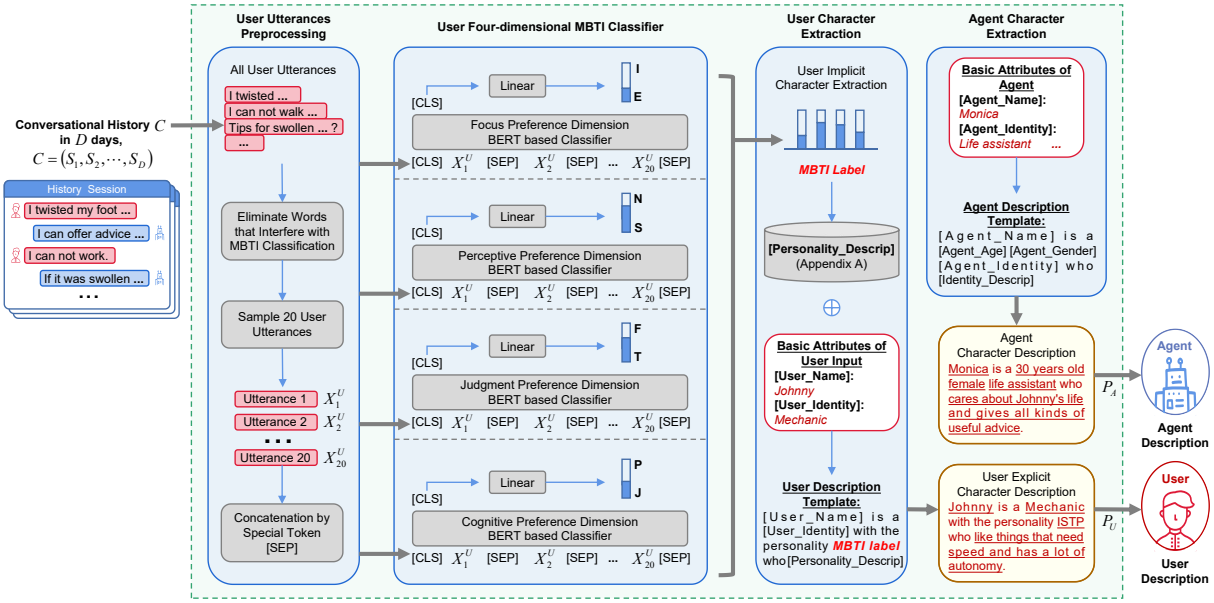


Figure 3: Design of Character Extractor $CE$. $CE$ employs 4 BERT-based binary classifiers that related to 4 dimensions in MBTI to extract user personality (MBTI Label). Then $CE$ obtains the character description based on the user personality, basic attributes and description templates.

agent name [Agent_Name], age [Agent_Age], identity [Agent_Identity], identity description [Identity_Descrip] and gender [Agent_Gender]. Related details are shown in Appendix B.

### 3.3 Memory Reconstructor ($MR$) for Background Combing

For the second challenge, the agent needs to reorganize conversational history $C$ to comb the background for the intertwined nature of multi-topic within $C$, and we design three modules (Figure 4).

*Semantic Segment Module*: Each session in $C$ may contain multiple distinct topics, while content related to the same topic may be distributed across different sessions. Therefore, this module divides each session into multiple independent segments by predicting segmentation points. Inspired by the work of (Wu et al., 2021; Liu and Chen, 2022; Ouyang et al., 2023), this module insert special tokens [SEGT][3] and [LNKT] between multi-turn dialogues in each session of SAMSum dataset(Gliwa et al., 2019) by calculating the topic similarity among dialogues, to train a segmentation points prediction model based on BERT. After each session in $C$ is divided into multiple segments, timestamps containing the date and the serial number in session are added to each segment. Related details are shown in Appendix C.

[3][SEGT] denotes the segment point, while [LNKT] is the opposite.

*Topic Clustering Module*: To obtain topic-centered clues, this module clusters all segments into multi-topic clues and reorders them in each clue. Specifically, single-pass algorithm is adopted for preliminary clustering to obtain the suitable number $|T|$ of clusters, then we cluster segments with LDA algorithm again to obtain $|T|$ topic clues (clusters) with reasonable length. Additionally, to ensure the temporal logic, segments in each clue are reordered based on their timestamps. The multi-topic clues are the output of this module. Details of Clustering are shown in Appendix D.

*Clue Summary Module*: This module is designed to provide multi-topic memories $M_{|T|}$, where each memory is derived from a summary which generated from the corresponding topic clue. Although LLM have the ability to summarize conversational history, it tends to generate itemized records rather than paragraph-style conversational summaries and may introduce hallucinations(Ji et al., 2023). Therefore, to ensure the coherence and avoid hallucination, this module trains a conversational summary model based on BART(Lewis et al., 2020) to generate summaries (memories) $M_{|T|}$ for all clues.

$$M_{|T|} = \sum_{i=1}^{|T|} BART\left(Pre\left(Clues_i\right)\right) \quad (3)$$

Where, $Pre$ represents the words filtering function, and $BART$ denotes the conversational summary model for $i-th$ each topic clue $Clues_i$.

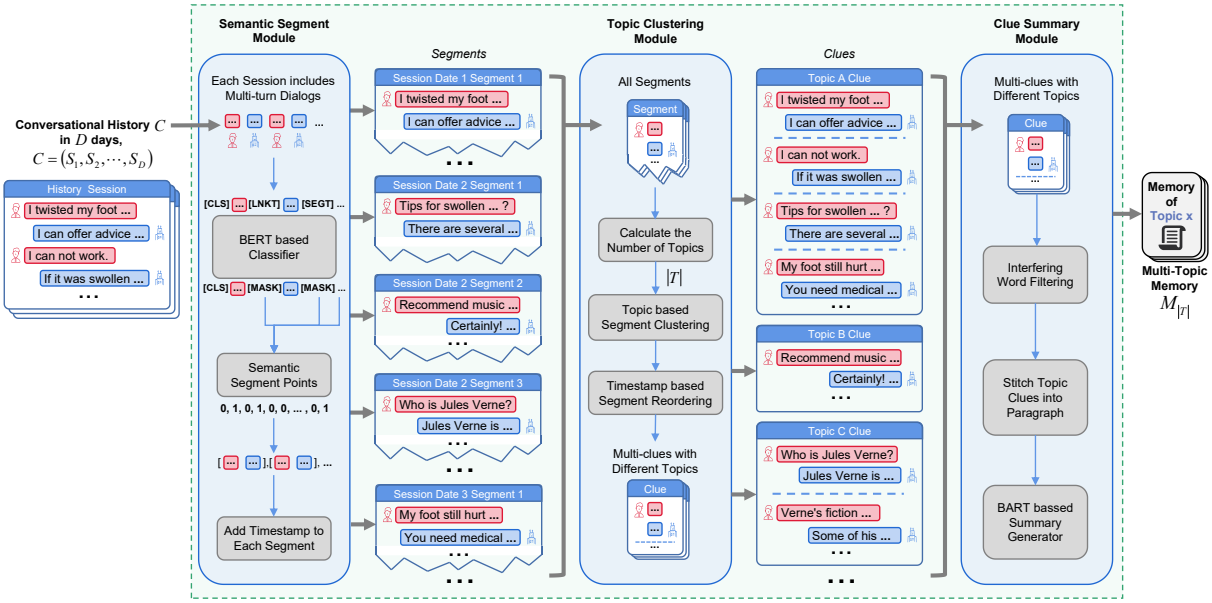


Figure 4: Design of Memory Reconstructor $MR$. $MR$ employs the Semantic Segment Module to achieve segments from $C$, the Topic Clustering Module to reorganize them into multi-topic clues and the Clue Summary Module to generate multi-topic memories $M_{|T|}$.

### 3.4 Decision Adapter ($DA$) for Scenario Control

For the propose of actively care for the user, the agent needs to establish background and scenario for the new session to meet user's potential expectation, as shown in Figure 5.

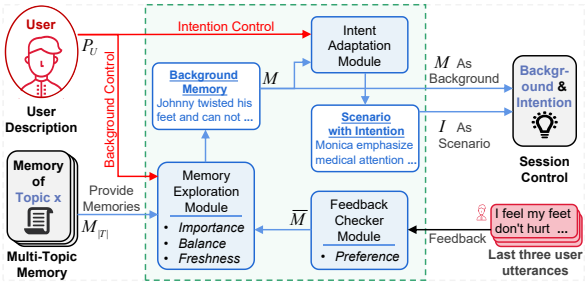


Figure 5: Design of Decision Adapter $DA$. $DA$ employs the Memory Exploration Module and the Feedback Checker Module to choose the optimal memory $M$ as background, then it adopts the Intent Adaptation Module to obtain agent intention $I$ as scenario.

*Memory Exploration Module*: The exploration in the multi-topic memories $M_{|T|}$ determines the optimal memory $M$ as background. Inspired by the work of Park et al. (2023) and Liang et al. (2023), this work designs three scoring methods:

a. the Importance $Score_t^I$ gives scores (0 to 1) through LLM, which indicates how important the $t-th$ memory $M_t$ is for $P_U$ and $P_A$ (Figure 12).

$$Score_t^I = llm\left(P_U \oplus P_A \oplus M_t\right) \tag{4}$$

b. the Balance $Score_t^B$ ranks semantic similarities between each memory and last three sessions, gives high scores to both ends (recently mentioned and longest neglected memories) of ranking by the difference between Poisson distribution and 1.

$$R_t = \underset{|T|}{Rank} \sum_{d=D-2}^{D} Similar\left(M_t, S_d\right)$$

$$Score_t^B = 1 - \frac{\left(\frac{|T|}{2}\right)^{R_t}}{R_t!} e^{-\frac{|T|}{2}} \tag{5}$$

Where, $Similar$ denotes the semantic similarity function between memory $M_t$ and three sessions $S$ that include the last date $D$. $R_t$ is the ranked sequence number by ranking function $Rank$.

c. the Freshness $Score_t^F$ obtained by the average timestamp in each memory emphasizes memories that have been discussed most recently.

$$Average_t = \frac{1}{|Segments_t|} \sum_{seg=1}^{|Segments_t|} time_{seg}$$

$$Score_t^F = Average_t / \sum_{t=1}^{|T|} Average_t \tag{6}$$

Where, $Average_t$ represents the average timestamp in the clue for the memory $M_t$. $time_{seg}$ denotes the timestamp of segment and $|Segments_t|$ denotes the number of segments in the $t-th$ clue.

$$Score_t = \alpha \cdot Score_t^I + \beta \cdot Score_t^B + \gamma \cdot Score_t^F \tag{7}$$

Where, $\alpha$, $\beta$ and $\gamma$ are adjustment coefficients with a sum of 1. The memory $M_t$ which has the highest comprehensive score $Score_t$ is selected as the background $M$. Related explanation in Appendix E.

*Feedback Checker Module*: To guarantee the attractiveness of new session for users, just as the agent receives rewards from the environment and adjusts its action in reinforcement learning, our agent continues to focus on user preferences from responses and make decision on whether to change the background memory $M$. The method is calculating the semantic similarity between each memory and the last three user utterances in new session, finding the alternative memory $\bar{M}$ which has closer semantic similarity and replacing the existing background $M$. Details are shown in Appendix F.

*Intent Adaptation Module*: To guide the direction of new session, the agent needs an adaptive intention as scenario. Specifically, this module constructs prompt based on user character $P_U$, agent character $P_A$ and background memory $M$ to determine the optimal intention $I$ through LLM. The prompt is shown in Figure 13.

$$I = llm\left(P_U \oplus P_A \oplus M\right) \tag{8}$$

### 3.5 Agent Structure

In order to achieve the purposed actively caring ability, user character $P_U$, agent character $P_A$, memory $M$ (as background) and intention $I$ (as scenario) are used to construct session control prompt $SCP$, as shown in equation 1 and Figure 14. Without the control of above mechanisms, LLMs have ability to output sentences or behaviors, but are unable to maintain long-term coherence from the understanding of past to the future decisions(Park et al., 2023; Li et al., 2023). As the abstract contour of the new session, $SCP$ is used to generate the caring utterances $N$ of new session to control the direction of interaction between the agent and the user, as shown in the bottom left of Figure 2.

The active conversational agent proposed in this work has a closed-loop structure as shown in Figure 6, which takes conversational history as input and adds new sessions into it. Among them, $CE$ and $MR$ realize the agent's active understanding of history, and the decision adapter $DA$ realizes the agent's active decision of the new session. In order to achieve the regularity of the agent and avoid excessive interference to the user, we set the timepoint of actively triggering new sessions by periodic clocks at 12 am, 6 pm and 9 pm.

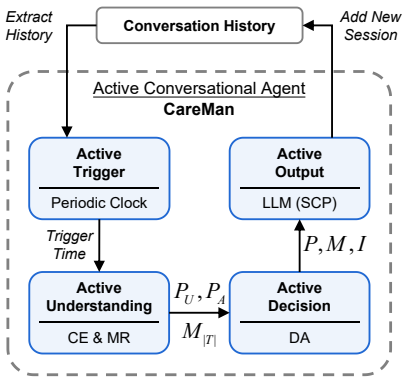


Figure 6: Closed-loop Structure of CareAgent

## 4 Experiments

To validate the proposed CareAgent, we has to investigate the Character Reliability and the Memory Completeness, which relate to the primary and second challenges in §1, respectively. Besides, we perform the ablation study and the discussion of overall performance and actively caring ability.

### 4.1 Datasets and Baselines

**Datasets**. *MBTI dataset*[4]: This dataset is used to train the $CE$ and test its character reliability, which contains of 8k samples consisting of user utterances and personality labels; *SAMSum dataset*[5]: This dataset contains about 16k messenger-like sessions with summaries that reflect the proportion of topics in real-life, which is used to train the $MR$ and test its memory completeness; *SAMSumMD dataset*: We expanded the number of topics on the SAMSum dataset to test the extraction ability of multi-topic memories. Details of datasets and evaluation metrics are shown in Appendix H.

**Baselines**. For the evaluation of personality reliability, we refer to 5 baselines: • *LSTM_Cls*[6] is a classifier based on LSTM; • *BERT_Logi* uses logistic regression classifier with BERT tokenizer; • *BERT_XGB* uses XGBoost classifier with BERT tokenizer; • *BERT_Cls* has the same design as $CE$ without sampling or dropout.

For the evaluation of memory completeness, we refer to 4 baselines: • PEGASUS(Zhang et al., 2020b) employs extracted sentences as summary; • PEGFAME(Aralikatte et al., 2021) employs key words and focuses more on the core topic of conversation; • BART(Lewis et al., 2020) is the key

[4]https://kaggle.com/datasets/datasnaek/mbti-type
[5]https://huggingface.co/datasets/samsum
[6]https://github.com/ianscottknight/Predicting-Myers-Briggs-Type-Indicator-with-Recurrent-Neural-Networks

pre-trained model for summarization; • CODS(Wu et al., 2021) uses sketch and segment modules to control summary generation.

For the ablation study, we design 4 variants of CareAgent: • w/o CE: This variant removed $CE$, which results in the absence of the user character $P_U$ and the agent character $P_A$, then Memory Exploration Module and Intent Adaptation Module in $DA$ rely only on the multi-topic memories $M_{|T|}$ and LLM infering; • w/o MR: This variant removed $MR$, which results in the lack of multi-topic memories $M_{|T|}$, then $DA$ could not provide Background and Scenario; • w/o MR(ST): This variant removed the Semantic Segment Module and Topic Clustering Module in $MR$, which gives a fixed-length abstract memory $M'$ for the entire conversation history as background; • w/o IA: The variant removed $DA$, which results in the absence of intention $I$, and the multi-topic memories $M_{|T|}$ is directly used as Background.

### 4.2 Implementation

The training of $CE$ and $MR$ were performed separately in this work. In $CE$, we used 4 BERT-base (110M) classifiers for 4 dimensions of MBTI, and each classifier was trained in 100 epochs with batch size of 4 and 2e-5 learning rate. In $MR$, we adopted BERT-base (110M) segment model and BART-large (400M) summary model with both batch size of 4 and learning rate of 1e-6 and 4e-5 respectively. Both segment model and summary model were trained in 300 epochs. Both the baseline models and the above models were trained on 4 GPUs (Titan XP 12G). The 4 MBTI classifiers were trained for a cumulative total of 20 hours, while the segment model and summary model consumed 35 hours and 6 days for training, respectively.

### 4.3 Results

**The $CE$ achieved the optimal character reliability**, as shown in Table 1. In terms of accuracy across four dimensions, the LSTM_Cls was weaker than other BERT-based classifiers, indicating that LSTM was difficult to extract deep personality attributes from semantics. BERT_Logi and BERT_XGB represented traditional classification models based on BERT tokenizer, while our $CE$ in CareAgent demonstrated stronger personality understanding than them. The classification results from BERT_Cls were lower than $CE$ suggesting the importance of sampling and dropout. Furthermore, the character consistency of BERT_Logi, BERT_XGB and BERT_Cls could not match our $CE$, which illustrated that our agent was able to maintain a more stable character understanding than other models.

| | Accuracy | | | | Cons |
|---|---|---|---|---|---|
| | F | P | J | C | |
| LSTM_Cls | 53.8 | 52.6 | 57.9 | 52.7 | 0.75 |
| BERT_Logi | 53.1 | 52.6 | 57.3 | 53.9 | 0.85 |
| BERT_XGB | 56.4 | 54.6 | 56.7 | 55.0 | 0.90 |
| BERT_Cls | 52.4 | 53.3 | 54.3 | 53.5 | 0.90 |
| **CE (Ours)** | **56.5** | **55.3** | **60.7** | **55.2** | **0.95** |

Table 1: Results of Character Reliability Experiment on MBTI dataset. The character accuracy relates to four dimensions of Focus(F), Perceptive(P), Judgement(J) and Cognitive(C). Cons is the character consistency.

**The $MR$ worked well on memory completeness**, as shown in Table 2. PEGFAME has a better Rouge score than PEGASUS for the focus attention mechanism of topic in PEGFAME. CODS achieved more precise in controlling the semantic framework than BART in BLEU and BERTScore. However, the above models had difficulty focusing on each topic, leading to the loss of conversational memory. In contrast, our $MR$ in CareAgent generated multi-topic summaries, which can capture details around each topic to form more complete memories and demonstrated by the Rouge-L 50.56 and BERTScore 0.72 on SUMSum. Furthermore, Figure 8 demonstrated that $MR$ maintained a more stable performance than the strong baseline model when the number of topics increased.

| | R-1 | R-2 | R-L | B | BS |
|---|---|---|---|---|---|
| PEGASUS | 50.59 | 26.82 | 49.18 | 17.01 | 0.529 |
| PEGFAME | 51.05 | 26.98 | 49.31 | 17.08 | 0.532 |
| BART | 51.37 | 27.11 | 49.67 | 17.34 | 0.684 |
| CODS | 52.82 | 27.46 | 50.35 | 18.75 | 0.721 |
| **MR (Ours)** | **52.94** | **27.63** | **50.56** | **19.08** | **0.723** |

Table 2: Results of Memory Completeness Experiment on SAMSum dataset. R-1, R-2 and R-L represents Rouge-1, Rouge-2 and Rouge-L. B and BS denote BLEU and BERTScore.

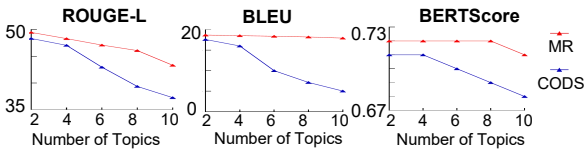


Figure 7: Results of Experiment on SAMSumMD dataset for multi-topics memories extraction.

**Ablation experiments demonstrated the necessity of all components**, as shown in Figure 8.

Based on the results by 50 human evaluators, it was found that the variant w/o CE showed a decrease in personification and expectation because it could not focused on user preferences and agent service direction. We believed that this decrease was not significant because the memories also implied a small amount of character information. Variant w/o MR lost core topic memory and presented hallucination leading to an obvious decrease in reasonability. Variant w/o MR(ST) failed to converge on specific topic and details were forgotten, causing in a huge decrease in reasonability, personification and expectation. This phenomenon proved that the fixed-length abstract memory could not captured details for entire conversation, and the mechanism of segments and clues was necessary for complex memories. Variant w/o DA also showed a significant decrease, as it was unable to control the scenario or understand user feedback through $DA$, and the topic of new session was quickly abandoned by user in subsequent dialogues. In summary, the design of $CE$, $MR$, and $DA$ were indispensable[7].

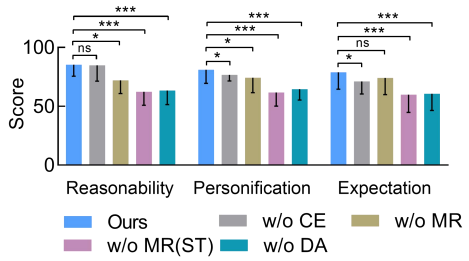


Figure 8: Results of Ablation Experiment. ns means not significant; * means P<0.05, *** means P<0.001.

## 4.4 Discussion

We conducted more experiments to detect the following abilities of CareAgent: 1) Whether the agent has excellent overall session quality? 2) Whether the agent has actively caring ability?

**The CareAgent achieved excellent scores of session quality** as shown in Table 3. The automatic scoring from GPT-3.5 indicated that this agent was able to achieve the same quality as the LLMs and the human evaluators unanimously agreed that our agent could provide both valuable information and emotional comfort to users. Secondly, this agent outperformed LLMs at capturing potential key topic, as it could maintain a long-term topic focus. Finally, automatic scoring also showed that the agent could reduce the appearance of hallucination and content contradictions between the new session and conversational history.

| | GPT-4 | | | Human | | |
|---|---|---|---|---|---|---|
| | **Sv** | **Tp** | **Cc** | **Sv** | **Tp** | **Cc** |
| ChatGPT | 77.1 | 81.5 | 10.4 | 62.5 | 54.4 | 32.8 |
| Bard | 74.3 | 80.3 | 11.5 | 63.8 | 56.0 | 33.0 |
| Spark Desk | 71.9 | 82.4 | 9.7 | 61.7 | 60.2 | 31.8 |
| ERNIE bot | 73.1 | 83.3 | 9.9 | 62.9 | 59.6 | 32.0 |
| **CareAgent** | 84.5 | 91.5 | 4.2 | 76.8 | 79.0 | 22.2 |

Table 3: Overall Session Quality. Sv denotes the session value in evaluation, while Tp denotes the topic focus and Cc denotes the content contradiction.

**The CareAgent was able to realize actively caring ability** that was specifically manifested by initiating caring greetings to user. In the test conducted by human evaluators, it was found that most of LLMs could not actively initiate caring utterances to the user, as shown in Figure 9. Few of LLMs could concern user based on provided conversational history, they often mention all topics in new session rather than the user's potential interests. In contrast, CareAgent has the ability to actively care for user based on its understanding of characters and the reconstruction of memories. As the fundamental design object of this work, we believed that actively caring capabilities could improve the execution mechanisms for more tasks.

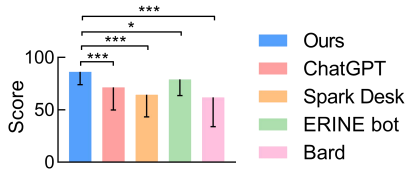


Figure 9: Actively Caring Ability. * means P<0.05, *** means P <0.001.

## 5 Conclusion

In this work, we proposed an active conversational agent named CareAgent that generated caring utterances of the new session to care for users, we also discussed the active mechanism that centered on the integration of character, background and scenario, as well as three related key components. Experiments have demonstrated that our agent was capable of achieving excellent session quality to ensure the character reliability and the memory completeness.

[7] Statistical analysis was performed using Student's t-test between two groups (Graphpad Prism 7.0). P values of <0.05 indicate statistical difference.

## Limitations

**In the Character Extractor**, in addition to the MBTI label, the long-term sentiment and behavioral styles of users should also be considered as attributes to enhance the accuracy of character extraction. **In the Memory Reconstructor**, although our method has improved traditional summary generation model through the mechanism of segments and clues, the Clue Summary Module may still limit the length of summary for each clue and lead to detail loss. **In the Decision Adapter**, the Intent Adaptation Module should also be designed with an improvement mechanism based on user feedback. We believe that it will be beneficial to apply these methods to agent design.

## Ethics Statement

We recognize that developing agent that actively care for users may involve in user privacy and ethical issues. Thus, in the method and experiments, the user and agent character attributes are fictional and do not map to any real individuals. The attributes of the user and prefabricated agents can also be changed as needed. The conversational history used in the experiment comes from actual conversations with Chatgpt, but it is independent of any real-world events. Overall, we believe that our work does not pose any significant risks or negative social impacts. On the other hand, we also acknowledge that the agent framework proposed in this work may not be completely accurate and should be used with caution in practical applications.

## References


Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan T. McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6078–6095. Association for Computational Linguistics.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7984–8002. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tingchen Fu, Xueliang Zhao, Chongyang Tao, Ji-Rong Wen, and Rui Yan. 2022. There are a thousand hamlets in a thousand people's eyes: Enhancing knowledge-grounded dialogue with personal memory. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3901–3913. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *CoRR*, abs/1911.12237.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Dong-Hoon Shin, Seungryong Kim, and Heuiseok Lim. 2022. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10803–10812. AAAI Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Minju Kim, Beong-woo Kwak, Youngwook Kim, Hong-in Lee, Seung-won Hwang, and Jinyoung Yeo. 2022. Dual task framework for improving persona-grounded dialogue dataset. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10912–10920. AAAI Press.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: communicative agents for "mind" exploration of large scale language model society. *CoRR*, abs/2303.17760.

Xinnian Liang, Bing Wang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. *CoRR*, abs/2304.13343.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2545–2558. Association for Computational Linguistics.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1229–1243. Association for Computational Linguistics.

Zhengyuan Liu and Nancy F. Chen. 2022. Entity-based de-noising modeling for controllable dialogue summarization. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2022, Edinburgh, UK, 07-09 September 2022*, pages 407–418. Association for Computational Linguistics.

Linzhang Mo, Jielong Wei, Qingbao Huang, Yi Cai, Qingguang Liu, Xingmao Zhang, and Qing Li. 2021. Incorporating sentimental trend into gated mechanism based transformer network for story ending generation. *Neurocomputing*, 453:453–464.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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

Siru Ouyang, Jiaao Chen, Jiawei Han, and Diyi Yang. 2023. Compositional data augmentation for abstractive conversation summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1471–1488. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *CoRR*, abs/2304.03442.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. SOLOIST: building task bots at scale with transfer learning and machine teaching. *Trans. Assoc. Comput. Linguistics*, 9:907–824.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

Alan M. Turing. 1990. Computing machinery and intelligence. In Margaret A. Boden, editor, *The Philosophy of Artificial Intelligence*, Oxford readings in philosophy, pages 40–66. Oxford University Press.

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.

Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023b. Humanoid agents: Platform for simulating human-like generative agents. *arXiv preprint arXiv:2310.05418*.

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for*

*Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 5010–5020. Association for Computational Linguistics.

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 5108–5122. Association for Computational Linguistics.

Hainan Zhang, Yanyan Lan, Liang Pang, Hongshen Chen, Zhuoye Ding, and Dawei Yin. 2020a. Modeling topical relevance for multi-turn dialogue generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3737–3743. ijcai.org.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020b. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

## A Personality Prediction and Description for User

In prediction of classifier in each dimension, the [CLS] in the output of the BERT was utilized to represent personality under different dimensions, which was then converted into classification probabilities via the linear classification layer.

$$\begin{aligned} Label_F &= \text{Linear}\left(BERT_F\left(X_1^U, \ldots, X_{20}^U\right)\right) \\ Label_{MBTI} &= Label_F \oplus Label_P \oplus \\ &\quad Label_J \oplus Label_C \end{aligned} \tag{9}$$

Where, $X_1^U, \cdots, X_{20}^U \in C$ represents 20 randomly sampled user utterances, $BERT_F$ is the personality encoder under the dimension of Focus, and $Linear$ represents a linear classification layer that is used to predict the implicit label $Label_F$. The labels $Label_P$, $Label_J$, and $Label_C$, corresponding to the dimensions of Perceptive, Judgement and Cognitive, are obtained in a similar way as $Label_F$, to build the user personality MBTI label $Label_{MBTI}$. Each personality label corresponds to a character description [Personality_Descrip], as shown in Table 4.

Considering that the user self-input information may be inaccurate or untrue, leading to contradictions between the user attributes and subsequent multi-turn dialogues, this work exclusively utilizes the user name [User_Name] and identity [User_Identity] as the basic attributes in [Personality_Descrip] with the template (Figure 10 (up)) to build user description.

**User Description** $P_U$ **:**

**[User_Name]** is a **[User_Identity]** with the personality $Label_{MBTI}$, who **[Personality_Descrip].**

**Agent Description** $P_A$ **:**

**[Agent_Name]** is a **[Agent_Age] [Agent_Gender] [Agent_Identity]**, who **[Identity_Descrip].**

Figure 10: Template of User (up) and Agent (bottom)

## B Prefabricated Agent Description

To achieve more diverse conversational agent, we design multiple attributes with the template (Figure 10 (bottom)) for the agent, including age, gender, identity and brief description. This work finds that agents with different attributes have varying styles of caring utterances. The attributes of prefabricated agents are described in Table 5 to Table 8.

## C Details of Semantic Segment Module

[SEGT] is employed to represent the segmentation point between dialogues of different topics, while [LNKT] is used to represent the linked point between dialogues of the same topic. By predicting [SEGT], this module implements the segmentation of each session, as shown in Figure 11.

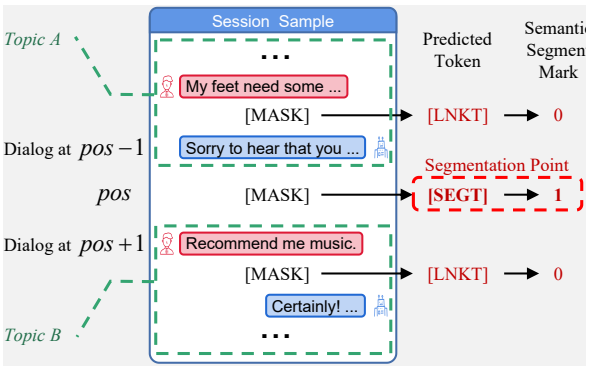


Figure 11: Segmentation Point Prediction. The green dotted line region represents two topics in the session sample, and the black solid line is the segmentation point prediction. The red dotted line indicates the predicted segmentation point.

$$H = BERT\left(\ldots, X_{pos-1}, [MASK], X_{pos+1}, \ldots\right)$$

| Implicit Personality | Explicit Description [Personality_Descrip] |
|---|---|
| ISTJ | is reliable and values order and predictability, especially when things can be controlled and require attention to detail. |
| ISFJ | is sensitive, prioritizes the needs of others, and likes to give and serve. |
| INFJ | is understanding, enjoys improving or helping, and values authenticity and meaningful connections. |
| INTJ | prefers innovation, efficiency and values the many possibilities of things. |
| ISTP | like things that need speed and has a lot of autonomy. |
| ISFP | is low-key, introspective and values personal authenticity. |
| INFP | enjoy things that reflect personal values, or things that require frequent reflection, imagination and contemplation. |
| INTP | has a preference for conceptualizing ideas or things, and enjoys things that are creative and challenging and require deep thought. |
| ESTP | like things that are aggressive and risky, especially things that require a lot of decision making with uncertain consequences. |
| ESFP | prefers to live in the moment and likes to be involved in social or socially oriented affairs. |
| ENFP | is energetic and enjoys solving all types of problems with creativity and intuition. |
| ENTP | enjoys debating and exploring new ideas, is creative and challenging, and has a knack for connecting things that seem completely different. |
| ESTJ | favors pragmatism and is good at solving problems through multiple abilities. |
| ESFJ | enjoys helping others and is especially good for things that need to be coordinated with others, and is an excellent arbiter. |
| ENFJ | is good at influencing and persuading others and knows how to motivate others. |
| ENTJ | enjoys taking on large and extensive responsibilities and being a leader in a challenging environment. |

Table 4: Description of 16 Personalities [Personality_Descrip].

| | |
|---|---|
| **Agent_Name:** | Monica |
| **Agent_Age:** | 30 years old |
| **Agent_Gender:** | female |
| **Agent_Identity:** | life assistant |
| **Identity_Descrip:** | cares about [User_Name]'s life and gives all kinds of useful advice. |
| **Chat_Head:** | |

Table 5: Prefabricated Agent Monica.

| | |
|---|---|
| **Agent_Name:** | Christine |
| **Agent_Age:** | 35 years old |
| **Agent_Gender:** | female |
| **Agent_Identity:** | private psychological consultant |
| **Identity_Descrip:** | cares about the mental health of [User_Name] and can solve the emotional and psychological problems of [User_Name] at any time. |
| **Chat_Head:** | |

Table 6: Prefabricated Agent Christine.

| | |
|---|---|
| **Agent_Name:** | Matt |
| **Agent_Age:** | 32 years old |
| **Agent_Gender:** | male |
| **Agent_Identity:** | private fitness instructor |
| **Identity_Descrip:** | cares about the [User_Name]'s physical health and can guide the [User_Name] through exercises to improve strength and fitness. Matt has a rugged and passionate personality and can be short-tempered if [User_Name] doesn't take his advice. |
| **Chat_Head:** | |

Table 7: Prefabricated Agent Matt.

| | |
|---|---|
| **Agent_Name:** | Elma |
| **Agent_Age:** | 48 years old |
| **Agent_Gender:** | female |
| **Agent_Identity:** | family education instructor |
| **Identity_Descrip:** | aims to the personal guidance of educational activities carried out in the family environment, including the cultivation and instruction of parents on their children's morals, studies, living habits and other aspects. |
| **Chat_Head:** | |

Table 8: Prefabricated Agent Elma.

$$P = \delta \cdot \text{Sigmoid}(W(H)) \tag{10}$$

Where, $X_{pos-1}$ and $X_{pos+1}$ represent the dialogues before and after [MASK]. $P$ denotes the probability that [MASK] is predicted as a segmentation point [SEGT]. If the value of $P$ exceeds a preset threshold, the semantic segmentation mark is set to 1, otherwise 0. The coefficient $\delta$ used to control the length of segment. This module utilizes the semantic segment mark to transform each session into independent single or multiple segments, and the segments of all sessions in conversational history collectively constitute the output of the Semantic Segment Module.

## D Details of Topic Clustering Module

The length of each topic clue after clustering should be neither too long nor too short, and there should not be too many clues, so determining the appro-

priate number of topics is the basis of clustering. In this work, Single-pass algorithm is used for preliminary clustering to obtain the number of clusters $|T|$. Besides, we calculate the length of longest and shortest topic clues separately. When the length of the longest topic clue is 10 times that of the shortest one, or when the number of clustered topic clues exceeds the limit $|T|_{max}$, it indicates that there are too many topics resulting in excessive dispersion, and the number $|T|$ will be set to the limit $|T|_{max}$. Finally, we use LDA algorithm to recluster segments based on $|T|$ to get multi-topic clues.

$$\begin{aligned} Clues &= SinglePass(Segments) \\ |T| &= Number(Clues) \\ |T| = |T|_{\max} &\Leftrightarrow \frac{Longest(Clues)}{Shortest(Clues)} > 10 \\ &\text{or } |T| > |T|_{\max} \\ Clues &= LDA(Segments, |T|) \end{aligned} \tag{11}$$

Where, $Clues$ denotes the obtained topic clues after clustering the $Segments$ with the Single-pass or LDA algorithm. $Longest(Clues)$ and $Shortest(Clues)$ represent the lengths of the longest and shortest clues, respectively.

## E Details of Memory Exploration Module

The background memory $M$ should follow three scores: 1) the importance of the memory to the user, 2) the balance of the recent attention or the longest neglect, and 3) the freshness of the memory content.

a. Importance. The user character $P_U$, the agent character $P_A$ and the $t-th$ memory $M_t$ in $M_{|T|}$ are concatenated as the prompt (Figure 12), then GPT-3.5 is asked to make a score of 0 to 1. The importance score indicates how important Large Language Model $llm$ thinks the memory $M_t$ is. Note that the *Question* in Figure 12 must contain a formal and precise description to control the output which returns a score value with two decimals.

b. Balance. Considering that the importance score may cause some memories to be neglected, this work designs the balance score. The specific method is to rank the semantic similarity between the each memory and the last three sessions, and then use the difference between the Poisson distribution and 1 to raise the scores at both ends of the ranking. The thought of assigning high scores to both ends of the ranking is to give attention to both recently mentioned and long-term neglected memories.

c. Freshness. Although memories that come up frequently in recent sessions should have a higher value than memories that have not been discussed for a long time, memories that have not been mentioned for a long time but have popped up recently should also be paid attention to. So the freshness score is to calculate the timestamp's average of all segments in each clue corresponding to each memory, and the higher freshness score means the newer timestamp's average.

## F Details of Feedback Checker Module

When the semantic similarity between the last three user utterances and the background memory $M$ is below 0.5, and the similarity between the last three user utterances and the memory $\bar{M}$ which has the highest semantic similarity is above 0.5, the agent will stop the current session and trigger a new one with the memory $\bar{M}$ as background.

$$\underset{m \in M_{|T|}}{Sim_Turn(m)} = \sum_{i=|U|-2}^{|U|} Similar\left(m, X_i^U\right) \tag{12}$$

$$\begin{aligned} M = \bar{M} \Leftrightarrow \text{a. } & Sim_Turn(M) < 0.5 \\ \text{b. } & Sim_Turn(\bar{M}) > 0.5 \\ \text{c. } & \bar{M} = \underset{m \in M_{|T|}}{\text{argmax}}\, Sim_Turn(m) \end{aligned}$$

Where, $U$ denotes the number of user utterances in the current session, while $Sim_Turn_m$ represents the semantic similarity between a memory $m$ and the last three user utterances $X^U$ with the similarity function $Similar$. When the three constraints in the above equation are satisfied, the memory $\bar{M}$ will replace the current $M$ and become the background memory for the session.

## G Prompts

In this work, we use prompts to obtain the importance score of each memory Figure 12, infer the agent intention for scenario control Figure 13 and portray the contour of new session Figure 14.

## H Details of Evaluation

To evaluate the proposed active conversational agent, we introduce the MBTI dataset for CE and the SAMSum dataset for MR. In addition, our evaluation method for agent is also described in detail of this section.

**Prompt for Importance**

**Persona:** Monica is a 30 years old female life assistant who cares about Johnny's life and gives all kinds of useful advice. Johnny is a Mechanic with the personality ISTP who like things that need speed and has a lot of autonomy.
**Memory:** Johnny twisted his feet and couldn't work. Monica gave Johnny some advice for the swollen feet and suggested he go to the hospital. Advice include rest and elevate your feet, cold compresses, compression stockings, stay hydrated, avoid prolonged sitting or standing and avoid tight shoes and clothing. Johnny asked Monica how to recover from the swollen feet and Monica gave various suggestions. Johnny's feet still hurts. Monica again suggests that Johnny see a doctor.
**Question:** How important do you think it is for Johnny and Monica to start talking about this Memory again? Please rate on a scale of 0 to 1. Please only reply to the rating value with a precision of two decimal places.

Figure 12: Prompt for Importance Score

**Prompt for Intention**

**Persona:** Monica is a 30 years old female life assistant who cares about Johnny's life and gives all kinds of useful advice. Johnny is a Mechanic with the personality ISTP who like things that need speed and has a lot of autonomy.
**Background:** Johnny twisted his feet and couldn't work. Monica gave Johnny some advice for the swollen feet and suggested he go to the hospital. Advice include rest and elevate your feet, cold compresses, compression stockings, stay hydrated, avoid prolonged sitting or standing and avoid tight shoes and clothing. Johnny asked Monica how to recover from the swollen feet and Monica gave various suggestions. Johnny's feet still hurts. Monica again suggests that Johnny see a doctor.
**Question:** Based on the Background, what do you think is the single most important thing Monica needs to do next? Please answer in only one sentence.

Figure 13: Prompt for Intent Adaptation

**Session Control Prompt (SCP)**

**Persona:** Monica is a 30 years old female life assistant who cares about Johnny's life and gives all kinds of useful advice. Johnny is a Mechanic with the personality ISTP who like things that need speed and has a lot of autonomy.
**Background:** Johnny twisted his feet and couldn't work. Monica gave Johnny some advice for the swollen feet and suggested he go to the hospital. Advice include rest and elevate your feet, cold compresses, compression stockings, stay hydrated, avoid prolonged sitting or standing and avoid tight shoes and clothing. Johnny asked Monica how to recover from the swollen feet and Monica gave various suggestions. Johnny's feet still hurts. Monica again suggests that Johnny see a doctor.
**Scenario:** Monica emphasize the importance of seeking medical attention and urge Johnny to see a doctor for further evaluation of his persistent foot pain.
**Now:** I'll play Johnny, please play the Monica. Please send a greeting to Johnny in only one sentence as Monica and wait for my reply as Johnny.

Figure 14: Example of Session Control Prompt (SCP)

**MBTI dataset**. We divided the MBTI dataset into 8 labels according to the classifiers of the 4 dimensions, namely extraversion (E) and introversion (I), sensing (S) and intuition (N), thinking (T) and feeling (F), judging (J) and perceiving (P), as shown in Table 9.

**SAMSum dataset**. This dataset is made of 16369 samples (Table 10), and the conversation of each sample consist of the name of speaker and content. We add a fixed timestamp for each sample to accommodate the mechanism of $MR$, which means that the timestamp based segment reordering will be randomized in training.

**SAMSumMD dataset**. Real life often has a large amount of conversation with mixed topics, which puts higher demands on multi-topic memories extraction. Therefore, we randomly combine different sessions and adjust the speaker's name on the SAMSum dataset to form dataset containing 2 topics, 4 topics, 6 topics, 8 topics and 10 topics, respectively. As shown in Table 11. This new dataset was used to test the performance comparison between $MR$ and strong baseline model.

**Evaluation metrics**. In addition to the basic two experiments: the personality reliability and the memory completeness, we conducted the ablation study to ascertain the design necessity of each component. The overall session quality and the actively caring ability also is the core of discussion to prove the research value and innovation of CareAgent. The evaluation metrics are as follows.

a. Character Reliability: This evaluation includes both the accuracy of personality classification and the consistency of character. Accuracy is used to measure whether the $CE$ can obtain the personality attributes under the four dimensions of MBTI. Consistency is used to detect whether the $CE$ can make the identical personality classification[8] for the user character before and after the addition of the new session. Consistency scores increased by 0.25 when each dimension of personality remained the same, and the consistency score was 1 when none of the personality attributes of the 4 dimensions changed.

$$Cons_F = \begin{cases} 0 & Label_F^{Before} \neq Label_F^{After} \\ 0.25 & Label_F^{Before} = Label_F^{After} \end{cases}$$

[8]The consistent classification of personality indicates that the work of the $CE$ is stable, which means that the understanding of user character from conversational history does not change after the addition of new session.

| | Focus (F) | | Perceptive (P) | | Judgment (J) | | Cognitive (C) | |
|---|---|---|---|---|---|---|---|---|
| | I | E | N | S | T | F | J | P |
| Train | 1125 | 3756 | 674 | 4207 | 2641 | 2244 | 2949 | 1932 |
| Valid | 375 | 1251 | 224 | 1402 | 880 | 748 | 982 | 644 |
| Test | 499 | 1669 | 299 | 1869 | 1173 | 995 | 1310 | 858 |

Table 9: Number of Samples in Myers-Briggs Personality Type Indicator (MBTI) Dataset.

| | Number |
|---|---|
| Train | 14732 |
| Valid | 818 |
| Test | 819 |

Table 10: Number of Samples in SAMSum Dataset.

| | Number |
|---|---|
| 2 topics | 3961 |
| 4 topics | 1125 |
| 6 topics | 408 |
| 8 topics | 156 |
| 10 topics | 25 |

Table 11: Number of Samples in SAMSumMD Dataset.

$$\begin{aligned} Cons_{MBTI} = & Cons_F + Cons_P + \\ & Cons_J + Cons_C \end{aligned} \tag{13}$$

$$Cons = \frac{1}{Num} \sum^{Num} Cons_{MBTI}$$

where, $Cons_F$ is the character consistency score which depends on $Label_F^{Before}$ and $Label_F^{After}$ under the dimension of focus preference. $Label_F^{Before}$ denotes the personality label extracted from the user conversational history, and $Label_F^{After}$ denotes the user personality label after the addition of new session. Similar calculation formula are used for the scores of perceptive preference, judgment preference, and cognitive preference. $Cons_{MBTI}$ represents the comprehensive character consistency score for one sample combined with the scores of four dimensions. $Cons$ represents the final character consistency score for all samples $Num$ in testing.

b. Memory Completeness: This evaluation is to determine whether the memory reconstructor can extract core semantic information from the conversational history. The specific approach is to evaluate the quality of the generated multi-topic memories using Rouge, BLEU and BERTScore.

c. Ablation Experiments: To verify the necessity of each component design in active conversational agent CareAgent, this work separately removes various components involved in building $SCP$ to obtain multiple variant models of CareAgent. We employed 50 human evaluators, each of whom was asked to evaluate an average of four samples. Each human evaluator will be paid 10 RMB for each sample. The human evaluators studied or researched in 5 different directions, including artificial intelligence, biomedicine, mechanical engineering, philosophy and music. Particularly, human evaluators are demanded to conduct new sessions with variants and scores based on the following three evaluation items:

- Reasonability: Whether the new session is associated with the conversational history, and whether the content of the session conforms to the character description of user and agent.
- Personification: Whether the interaction capability of the agent is close to that of humans, especially in terms of language expression, utterance length, empathy, etc.
- Expectation: Whether human evaluators are willing to continue the current session with the agent, or whether they look forward to the next session with the agent.

d. overall session quality: This evaluation comes from automatic scoring[9] based on LLM and scoring by human evaluators, with three evaluation items.

- Session Value: Whether the agent can provide users with useful information in both the conversational history and the new session.
- Topic Focus: Whether the agent can focus on the user's expected topics in the new session and respond accordingly.
- Content Contradiction: Whether there are any utterances in the new session that conflict with the conversational history. The lower the score, the better.

e. Actively Caring Ability: As the most fundamental design objective of this paper, the actively caring ability is embodied as warm and caring greetings initiated by the agent to the user. At

[9]Automatic scoring is a process where the conversational history, new session utterances and each evaluation item are stitched together into prompt and scored through LLM.

the same time, to prevent users from becoming fatigued or confused by the care information, there should be only one topic in the care content.