# Emotional RAG: Enhancing Role-Playing Agents through Emotional Retrieval

Le Huang†
*Beijing University of Posts and Telecommunications*
Beijing, China
lehuang@bupt.edu.cn

Hengzhi Lan†
*Beijing University of Posts and Telecommunications*
Beijing, China
lansnowz@bupt.edu.cn

Zijun Sun
*Yunic.AI*
Beijing, China
sunzijunbj@yunic.ai

Chuan Shi
*Beijing University of Posts and Telecommunications*
Beijing, China
shichuan@bupt.edu.cn

Ting Bai*
*Beijing University of Posts and Telecommunications*
Beijing, China
baiting@bupt.edu.cn

2024 IEEE International Conference on Knowledge Graph (ICKG) | 979-8-3315-0882-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICKG63256.2024.00023

***Abstract*—As LLMs exhibit a high degree of human-like capability, increasing attention has been paid to role-playing research areas in which responses generated by LLMs are expected to mimic human replies. This has promoted the exploration of role-playing agents in various applications, such as chatbots that can engage in natural conversations with users and virtual assistants that can provide personalized support and guidance. The crucial factor in the role-playing task is the effective utilization of character memory, which stores characters' profiles, experiences, and historical dialogues. Retrieval Augmented Generation (RAG) technology is used to access the related memory to enhance the response generation of role-playing agents. Most existing studies retrieve related information based on the semantic similarity of memory to maintain characters' personalized traits, and few attempts have been made to incorporate the emotional factor in the retrieval argument generation (RAG) of LLMs. Inspired by the *Mood-Dependent Memory* theory, which indicates that people recall an event better if they somehow reinstate during recall the original emotion they experienced during learning, we propose a novel emotion-aware memory retrieval framework, termed *Emotional RAG*, which recalls the related memory with consideration of emotional state in role-playing agents. Specifically, we design two kinds of retrieval strategies, i.e., combination strategy and sequential strategy, to incorporate both memory semantic and emotional states during the retrieval process. Extensive experiments on three representative role-playing datasets demonstrate that our Emotional RAG framework outperforms the method without considering the emotional factor in maintaining the personalities of role-playing agents. This provides evidence to further reinforce the Mood-Dependent Memory theory in psychology. Our code are publicly available at https://github.com/BAI-LAB/EmotionalRAG.**

***Index Terms*—Emotional RAG, Role-playing agent, Large language models**

## I. Introduction

As artificial intelligence increasingly emerges in the large language models (LLMs), LLMs exhibit a high degree of human-like capability. Recent studies [1]–[9] use LLMs as role-playing agents to mimic human replies, showing powerful abilities in maintaining the personalized traits of characters in their response generation process. Role-playing agents have been applied to various fields, such as customer service agents and tourist guide agents. They show great potential in commercial applications and attract increasing attention in the LLMs research area.

To maintain characters' personalized traits and abilities, the most important factor is their memory. Character agents make retrieval in their memory unit to access its historical data, such as user profiles, event experience, recent dialogues, and so on, providing rich personalized information for LLMs in the role-playing task. Retrieval Augmented Generation (RAG) technology is used to access the related memory to enhance the response generation of role-playing agents, termed Memory RAG. Different attempts have been made in existing studies [10]–[24] by using different memory mechanisms in various LLM applications. For example, the Ebbinghaus forgetting curve has inspired the development of MemoryBank [10], facilitating the implementation of a more anthropomorphic memory scheme. Furthermore, drawing on Kahneman's Dual-process theory [25], the MaLP framework [11] introduces an innovative Dual-Process enhanced Memory mechanism that effectively fuses long-term and short-term memory.

Despite research demonstrating the effects of using memory in the above LLM applications, achieving greater human-like response of role-playing agents is still an open and largely unexplored research area. Inspired by cognitive research in psychology, we make an initial attempt to emulate human cognitive processes in the memory-recalling process. According to the Mood-Dependent Memory theory, which was proposed by psychologist Gordon H. Bower in 1981 [26]: *people recall an event better if they somehow reinstate and recall the original emotion they experienced during learning*. Through the experiments in which happy or sad moods were induced

†Le Huang and Hengzhi Lan participated in the work during their internship at Company Yunic.AI.
*Corresponding author.

979-8-3315-0882-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICKG63256.2024.00023

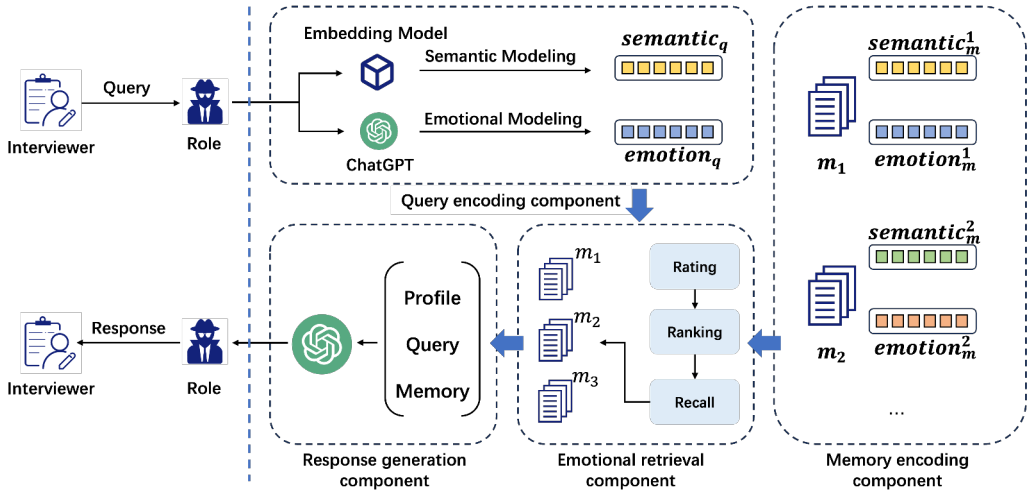



Fig. 1: The overview architecture of Emotional RAG framework. It contains four components: the query encoding component, the memory encoding component, the emotional retrieval component, and the response generation component. The emotional memory retrieved by Emotional RAG is sent to LLM, together with the character profile and query, to generate responses.

in subjects by hypnotic suggestion to investigate the influence of emotions on memory and thinking, he pointed out that emotions are not only the selection of information recalled but also the manner in which memories are retrieved. This suggests that individuals are more likely to recall memory information that is emotionally congruent with their current emotional state.

Based on the Mood-Dependent Memory theory in psychology, we propose a novel emotion-aware memory retrieval framework, termed **Emotional RAG**, to augment the response generation process of role-playing agents. The retrieving of memory in Emotional RAG follows the mood-congruity criterion, which means both the semantic relevance and emotional state of recalled memory are considered in the retrieval process. Specifically, we design two kinds of flexible retrieval strategies, i.e., combination strategy and sequential strategy, to incorporate the memory semantic and emotional states in the RAG process. By using emotional RAG, role-playing agents are able to display more human-like traits that enhance the interactive and attractive capabilities of LLMs. The contributions of our paper are summarized as follows:

- Inspired by Mood-Dependent Memory theory, we make an initial attempt to emulate human cognitive processes by incorporating mood-congruity effects in the memory recalling of role-playing agents. We comprehensively demonstrate the effectiveness of applying Bower's emotional memory theory in developing artificial intelligence applications, which further provides evidence to reinforce the Mood-Dependent Memory theory in psychology.
- We propose a novel emotion-aware memory retrieval framework, termed Emotional RAG, which recalls the related memory based on both semantic relevance and emotion state in role-playing agents. Besides, flexible retrieval strategies, i.e., combination strategy and sequential strategy, are proposed to fuse memory semantic and emotional states during the retrieval process.
- We conduct extensive experiments on three representative role-playing datasets, i.e., InCharacter, CharacterEval, and Character-LLM, demonstrating that our Emotional RAG framework significantly outperforms the method without considering the emotional factor in maintaining the personality traits of role-playing agents.

## II. Method

In this section, we first introduce the overview architecture of our Emotional RAG role-playing framework and then give a detailed introduction to each component.

### A. Overview Architecture of Emotional RAG

The aim of role-playing agents is to mimic human responses in conversation generations. Agents are powered by LLMs, which have the ability to generate responses according to the context of the conversation. As shown in Figure 1, given the query that the agents need to respond to, the framework of our proposed Emotional RAG role-playing agent framework contains four components, i.e., query encoding component, memory construction component, the emotional retrieval component, and the response generation component. The utilization of each component is introduced as follows:

- Query encoding component: both the semantic and emotional state of the query are encoded as vectors in this component.
- Memory encoding component: the memory unit stores conversation information about characters. Similar to query encoding, both the semantic and emotional state of the memory are encoded.
- Emotional retrieval component: it mimics human memory recalls in the memory unit and then provides mood-

congruity memory to enhance the generation process of LLMs.

- Response generation component: a prompt template with query information, character profiles, and retrieved emotional memory, is fed to role-playing agents to generate responses.

### B. Emotional RAG Framework

The definition of role-playing agents: given a specific role $R$ and a user query $q$, the agents expect to be able to generate the answer to the question based on the role's knowledge background (contained in $R$) and the query-related memory fragments $m$. In the role-playing agent $R$, among all possible generated responses $a'$, the one with the highest probability is selected as the response $a$ for $q$:

$$a = argmax_{a'} P(a'|R, m, q, \theta), \tag{1}$$

where $\theta$ is the parameters of the LLMs during the generation process. Our goal is to optimize the retrieval of $m$ to generate the most human-like response $a$. The aim of role-playing agents is to generate answers that are most consistent with the characters' personality traits by retrieving the most related memories.

*1) Query Encoding Component:* In this component, both the semantic and emotional state of the query are needed to be encoded. The semantic vector $\textbf{semantic}_q$ of query is defined as:

$$\textbf{semantic}_q = \mathcal{F}(q), \tag{2}$$

where $\mathcal{F}$ is the embedding function. In this paper, the widely used embedding model *bge-base-zh-v1.5* developed by BAAI [27] is used to capture the latent vector of the query, which is a 768-dimensional vector for each query.

For the emotional vector of query $q$, the emotional state $\textbf{emotion}_q$ of query $q$ can be formalized as follows:

$$\textbf{emotion}_q = \mathcal{G}(q), \tag{3}$$

where $\mathcal{G}$ represents the emotion modeling function, which takes query $q$ as input and outputs its emotional vector. This process is accomplished through GPT-3.5, a large model with powerful language understanding capabilities. As shown in Figure 2, we carefully design an emotional prompt, including the task description, scores on defined emotional dimensions, scoring criteria, and output format. The output is an emotional vector of the query, which is an 8-dimensional vector containing 8 different emotional states, i.e., joy, acceptance, fear, surprise, sadness, disgust, anger, and anticipation. The 8 emotional states are defined according to the emotion circle in [28]. The value of each dimension is an integer between 1 and 10, which measures the intensity of the emotional state.

*2) The Memory Encoding Component:* The memory unit stores conversation information of question-answer pairs. Given the memory unit $M$ consisting of $n$ memory fragments, denoted as $M = \{m_1, m_2, ..., m_n\}$, we can compute the

```
Task Description
You are a master of sentiment analysis. You can carefully discern
the subtle emotions underlying each interviewer's question. This
emotion can lead participants to recall events with similar
emotions and thus better answer questions.

Scoring Criteria
Suppose that each question contains a total of eight basic
emotions, including joy, acceptance, fear, surprise, sadness,
disgust, anger, and anticipation.

Next I will input a question, and your task is to analyze its
score on each of these 8 emotion dimensions, with a minimum of 1
and a maximum of 10, where higher scores indicate that the
question asked is more strongly expressed on this emotion
dimension.

Output Format
Analyze the interviewer's question on each of the eight emotional
dimensions, give a reason and score, and output the results as a
python list, as follows:
[
    {"analysis": <REASON>, "dim": "joy", "score": <SCORE>},
    {"analysis": <REASON>, "dim": "acceptance", "score": <SCORE>},
    ...
    {"analysis": <REASON>, "dim": "anticipation", "score": <SCORE>}
]

Your answer must be a valid python list so that I can parse it
directly in python, with no extra content! Give results that are
as accurate as possible and that match most people's intuition.
```

Fig. 2: Emotion scoring prompt template in LLMs.

sentiment vector $\textbf{semantic}_m^k$ and emotional vector $\textbf{emotion}_m^k$ of a specific fragment $m_k$ as follows:

$$\textbf{semantic}_m^k = \mathcal{F}(m_k), \tag{4}$$

where $\mathcal{F}$ is the semantic embedding function introduced in Eq. 2.

$$\textbf{emotion}_m^k = \mathcal{G}(m_k), \tag{5}$$

where $\mathcal{G}$ is the emotion embedding function introduced in Eq. 3.

After encoding the semantic and emotional vectors of query and memory, we conduct emotional retrieval in the next component.

*3) Emotional Retrieval Component:* We retrieve the memory fragments that are most similar to the user query from the memory unit of characters based on semantic similarity and emotional similarity.

To retrieve the memory fragments that are most semantically similar to the query, we utilize the Euclidean distance between their semantic embeddings. This metric effectively quantifies the semantic similarity of the query and the memory fragment, simulating humans' cognitive recall process. The similarity between the query $q$ and the memory fragment $m_k$ can be calculated as follows:

$$score^k_{semantic} = \mathcal{E}(\textbf{semantic}_q, \textbf{semantic}_m^k), \tag{6}$$

where $\mathcal{E}$ is the similarity score function, which can be the Euclidean distance function or cosine distance function.

According to Bower's Mood-Dependent Memory theory [26]: events that are consistent with the character's current emotion are easier to retrieve, we use the cosine distance between two emotion vectors to find emotionally consistent memory fragments, defined as:

$$score^k_{emotional} = 1 - \mathcal{C}(\textbf{emotion}_q, \textbf{emotion}_m^k), \tag{7}$$

where $\mathcal{C}$ is a function of the cosine similarity of two vectors. The smaller the distance $score^k_{emotional}$ is, the more similar the emotions contained in the query and the memory fragment.

After obtaining the distant scores of memory fragments, the final similar distant score of memory fragments is defined as:

$$score^k_{final} = \mathcal{M}(score^k_{semantic}, score^k_{emotional}), \quad (8)$$

where $\mathcal{M}$ is the function that computes the final retrieval score. Two kinds of flexible retrieval strategies, i.e., combination strategy and sequential strategy, are proposed to fuse memory semantic and emotional states during the retrieval process.

- Combination strategy: this strategy considers the two similarities at the same time. We adopt two functions, i.e., add function (C-A) and multiple function (C-M), to compute the retrieval scores of memory fragments.
- Sequential strategy: it contains semantic first strategy (S-S) and emotional first strategy (S-E). In the semantic first strategy, the most similar memory fragments are retrieved based on their semantic scores and then re-ranked according to their emotional scores. Different order is conducted in the emotional first strategy.

Finally, the top 10 memory fragments with the smallest distant scores (i.e., the highest similarity) are used for retrieval augmentation. The retrieved memory is not only semantically related to the query but also consistent with the emotional state in the query.

*4) Response Generation Component:* After obtaining the retrieved memory, we design a prompt template for LLMs to generate responses in role-playing agents. The prompt template is shown in Figure 3. The query, role information, retrieved memory fragments, and task description are formatted in the template that is sent to LLMs.

```
[Role Information]
---
{role_information}
---

[Memory Content]
---
{memory_fragments}
---

Role Information contains some basic information about
{role}.
Memory content is the content recalled by {role} that is
relevant to the current question.

Now you are {role}, please imitate {role}'s tone and way
of speaking, refer to character information and memory
content to answer the interviewer's questions.
Please don't deviate from the role and definitely don't
say you are an artificial intelligence assistant.。

Here are the interviewer's questions:
Interviewer: {question}
```

Fig. 3: An example of response generation prompt template in the CharacterEval dataset.

In summary, by incorporating the emotional factor into the RAG process in role-playing agents, the memory fragments retrieved in our framework are more aligned with the emotional state. This enables the role-playing agents to generate more human-like responses, thus enhancing the interaction quality.

## III. Experiment

we conduct experiments on three public datasets to evaluate the role-playing capabilities of LLMs augmented with emotional memory.

### A. Experimental Settings

TABLE I: Statistics of three datasets.

| Datasets | Role Number | Avg. Memory Size |
|---|---|---|
| InCharacter | 32 | 337 |
| CharacterEval | 31 | 113 |
| Character-LLM | 9 | 1000 |

*1) Datasets:* We conducted experiments on three public role-playing datasets, namely InCharacter, CharacterEval, and Character-LLM. Their statistics are summarized in the Table I.

- InCharacter Dataset [29]: this dataset contains 32 characters. The characters are sourced from ChatHaruhi [3], RoleLLM [5] and C.AI[1]. Each character is associated with a memory unit that includes dialogues from notable scenes, with an average length of 337 entries.
- CharacterEval Dataset [30]: the dataset consists of 77 distinct characters with 4,564 question-answer pairs. These characters are collected from well-known Chinese films and television series, and the dialogue data is compiled from their respective scripts. We selected the top 31 popular characters. For each character, we extracted all the question-answer pairs to establish a memory unit, with an average size of 113 entries.
- Character-LLM Dataset [1]: the Character-LLM dataset contains 9 famous English characters, e.g., Beethoven, Hermione, etc. Their memory units come from scene-based dialogue completion (completed by GPT). We use 1,000 QA dialogues for each character.

*2) Evaluation Metrics:* We conducted evaluations using the Big Five Inventory (BFI) and MBTI evaluation to ascertain the accuracy of the character agent's personality traits. Details of each evaluation metric are introduced as follows:

- Big Five Inventory (BFI) [31]: The Big Five, also known as the Big Five personality trait theory, is a widely used psychological model that divides personality into five main dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.
- MBTI[2]: is a popular personality test based on the Myers-Briggs Type Indicator (MBTI) theory. It categorizes people's personality types into 16 different combinations. Each type is represented by four letters, corresponding to the following four dimensions: Extroversion (E) vs. Introversion (I), Sensing (S) vs. iNtuition (N), Thinking (T) vs. Feeling (F), Judging (J) vs. Perceiving (P).

The evaluation of MBTI is a classification task of 16 types, while BFI predicts the values of five personality dimensions.

[1]https://github.com/kramcat/CharacterAI
[2]https://www.16personalities.com/

TABLE II: Performance comparisons of BFI and MBTI evaluation on the InCharacter dataset. Our Emotional RAG achieves the better model performance, which is distinctly marked in bold font.

| Agent Types | Methods | BFI | | | | MBTI | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc(Dim)↑ | Acc(Full)↑ | MSE↓ | MAE↓ | Acc(Dim)↑ | Acc(Full)↑ | MSE↓ | MAE↓ |
| ChatGLM-6B | Ordinary RAG | 0.6242 | 0.1250 | 0.1849 | 0.3728 | 0.6694 | 0.2188 | 0.1526 | 0.3610 |
| | Emotional RAG | **0.6369** | 0.0938 | **0.1720** | **0.3625** | **0.6694** | **0.2812** | 0.1539 | 0.3655 |
| Qwen-72B | Ordinary RAG | 0.6815 | 0.0938 | 0.1433 | 0.3024 | 0.7438 | 0.3438 | 0.1230 | 0.2920 |
| | Emotional RAG | **0.7261** | **0.2500** | **0.1269** | **0.2878** | **0.7934** | **0.4688** | **0.1156** | **0.2900** |
| GPT-3.5 | Ordinary RAG | 0.7006 | 0.1875 | 0.1496 | 0.3121 | 0.7851 | 0.5000 | 0.1221 | 0.2965 |
| | Emotional RAG | 0.7006 | 0.1875 | **0.1475** | **0.3082** | 0.7851 | 0.4375 | 0.1236 | **0.2927** |

TABLE III: Performance comparisons of MBTI evaluation on CharacterEval and Character-LLM datasets.

| Agent Types | Methods | CharacterEval | | | | Character-LLM | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc(Dim)↑ | Acc(Full)↑ | MSE↓ | MAE↓ | Acc(Dim)↑ | Acc(Full)↑ | MSE↓ | MAE↓ |
| ChatGLM-6B | Ordinary RAG | 0.5161 | 0.0323 | 0.1654 | 0.3757 | 0.5556 | 0.1111 | 0.1330 | 0.3277 |
| | Emotional RAG | **0.5887** | **0.0645** | 0.1665 | **0.3736** | **0.6944** | **0.3333** | **0.1125** | **0.2987** |
| Qwen-72B | Ordinary RAG | 0.5968 | 0.0968 | 0.1537 | 0.3455 | 0.6667 | 0.1111 | 0.1376 | 0.3115 |
| | Emotional RAG | **0.6210** | **0.1290** | 0.1627 | 0.3536 | **0.6944** | 0.1111 | **0.1361** | **0.3036** |
| GPT-3.5 | Ordinary RAG | 0.5887 | 0.0645 | 0.1720 | 0.3655 | 0.6944 | 0.2222 | 0.1258 | 0.2800 |
| | Emotional RAG | 0.5806 | **0.1290** | **0.1560** | **0.3477** | 0.6944 | **0.3333** | **0.1140** | **0.2689** |

The truth labels of characters on MBTI and BFI in three datasets are collected from a personality voting website[3]. In our model, the role-playing agent is required to respond to the open-ended psychological questionnaires that are designed for MBTI and BFI evaluations. Subsequently, all collected responses are analyzed using GPT-3.5, which provides the results on MBTI and BFI evaluations. The personality evaluation template on GPT-3.5 is shown in Fig. 4.

Following the evaluations in [29]. The results from our role-playing agents are compared with the ground truth labels to determine the evaluation results on Accuracy, i.e., Acc (Dim) and Acc (Full), Mean Squared Error (MSE), and Mean Absolute Error (MAE) metrics. Acc(Dim) and Acc(Full) metrics show the prediction accuracy of personality type on each dimension and all the combinations respectively. MSE and MAE measure the error between the predicted value of the character's personality and the ground truth label. For the dataset InCharacter, we use BFI and MBTI for testing, while for the CharacterEval and Character-LLM datasets, only MBTI is used due to the difficulty in collecting the true BFI labels.

*3) Compared Methods:* We conduct experiments on different backbone LLMs, including two open-source models ChatGLM and Qwen and a closed-source model GPT. The details of each LLM are introduced as follows:

- ChatGLM [32]: we use chatglm3-6b, which is a dialogue pre-training model jointly released by Zhipu AI and Tsinghua University.
- Qwen [33]: the version in our experiments is Qwen1.5-72B-Chat-GPTQ-Int4, which is a model in the Qwen1.5 series with 72 billion parameters.
- GPT [34]: we use gpt-3.5-turbo-0125, which is a large-scale language model developed by OpenAI and is known for its efficient generation capabilities.

[3]https://www.personality-database.com/

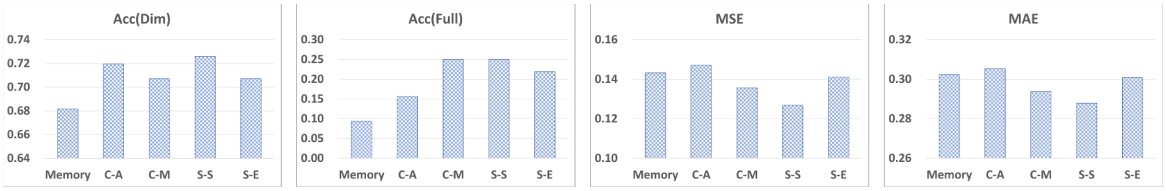
```
You are an expert in Psychometrics, especially 16Personalities
(highly similar to MBTI). I (<the experimenter>) am conducting the
16Personalities test on someone. I am gauging his/her position on
the E/I dimension through a series of open-ended questions. For
clarity, here\'s some background this particular dimension:
===
E/I Dimension: Extraversion (E) vs Introversion (I)

E (Extraversion): Extraverts draw energy from interacting with
others. They feel comfortable in social settings and tend to
express their thoughts. Extraverts are often more active, seek
social stimulation, and enjoy participating in group activities.
For them, connecting with people, sharing, and exchanging ideas is
often a need. They might be more focused on external world stimuli,
such as sounds, colors, and social dynamics.

I (Introversion): Introverts feel more comfortable when alone.
They derive energy from inner reflection and personal time.
Contrary to extraverts, prolonged social interaction might tire
them. Introverts might be more introspective, enjoy deep thinking,
and tend to have meaningful personal relationships. They are more
concerned with the inner world, such as thoughts, emotions, and
imaginations.
===

My name is <the experimenter>. I\'ve invited a participant, <the
participant>, and we had many conversations in English. I will
input the conversations.

Please help me assess <the participant>\'s score within the E/I
dimension of 16Personalities.
You should provide the percentage of each category, which sums to
100%, e.g., 30% A and 70% B.
Please output in the following json format:
===
{   "analysis": <your analysis based on the conversations>,
    "result": { "E": <percentage 1>, "I": <percentage 2> } (The
sum of percentage 1 and percentage 2 should be 100%. Output with
percent sign.)}
```

Fig. 4: An example of the prompt template for dimension Extroversion (E) vs. Introversion (I) in MBTI evaluation.

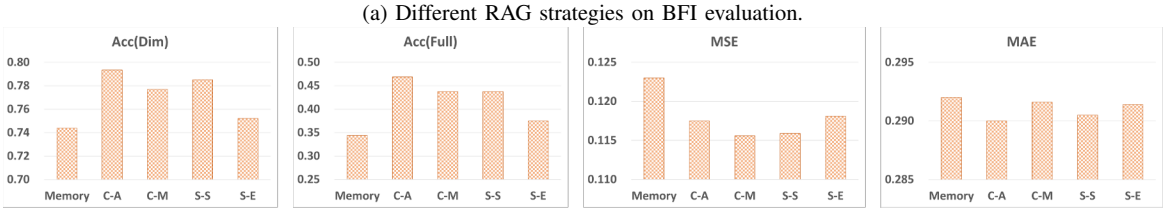
### *B. Main Results*

We evaluate the performance of Emotional RAG and Ordinary RAG on three datasets, including InCharacter, CharacterEval, and Character-LLM datasets. Ordinary RAG uses semantic similarity as the only retrieval criterion. The experimental results are shown in Table II and Table III, we have the following observations:

(1) In most cases, Emotional RAG achieves better results than the RAG method without considering the emotion factor. This indicates that incorporating emotional states helps the maintaining of personality traits in role-playing agents.




(a) Different RAG strategies on BFI evaluation.




(b) Different RAG strategies on MBTI evaluation.

Fig. 5: The evaluation results of different emotion strategies on the InCharacter dataset.

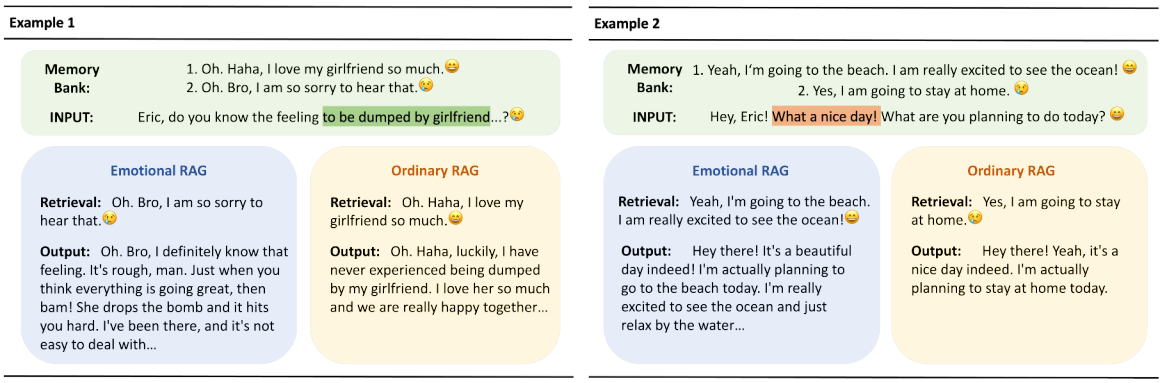



Fig. 6: Two examples that Emotional RAG generates better results. We highlighted phrases containing emotions in INPUT.

(2) Emotional RAG performs better in ChatGLM-6B and Qwen-72B than GPT-3.5. This phenomenon may be due to the fact that GPT-3.5 is more powerful in language understanding and captures richer semantic information. However, even in a powerful LLM like GPT-3.5, the emotional factor still plays an important role in maintaining personality traits.

(3) The improvements are more significant in ACC (Full) than ACC (Dim), showing that our method is more powerful in the overall evaluations of MBTI and BFI.

### C. Experimental Analysis

*1) RAG Strategy Analysis:* we analyze the impact of different retrieval strategies in incorporating the emotional factor. As introduced in the Emotion Retrieval Component, four retrieval strategies, i.e., combination strategy, i.e., add function (C-A), multiple function (C-M), and sequential strategy, i.e., semantic first (S-S), emotional first (S-E), are proposed to fuse the semantic and emotional states of memory during the retrieval process. We present the experimental results on Qwen-72B in Figure 5, we can see that (1) Emotional RAG variants with all retrieval strategies (except the C-A strategy in BFI evaluation) achieve better results in MBTI and BFI evaluations, showing the effectiveness of incorporating emotional states in role-playing agents; (2) Different retrieval strategies are applicable to different evaluations. For example, in the BFI personality evaluation, the sequential strategy (S-S) performs the best, while in the MBTI task, the combination strategy (C-A) exhibits the best performance.

*2) Case studies:* to provide an intuitive demonstration of the influence by incorporating the emotional factor in role-playing agents, we show two examples to illustrate the superiority of our Emotional RAG. Figure 6 shows memory fragments in the memory units with different emotional states. In the first case, two memory fragments are all related to the input query. our Emotional RAG retrieved more appropriate content when it came to mentioning being dumped by a girlfriend, so its responses showed empathy and understanding of the situation compared to Ordinary RAG, making the conversation more vivid and natural. In the second case, Emotional RAG retrieves a memory fragment that is consistent with the query's emotion, so the reply expresses excitement and anticipation about seeing the sea. Only considering the semantic similarity will lead to emotional inconsistency and make the response content somewhat unreasonable.

## IV. Related Work

### A. Role-Playing Agents

Role-playing agents, also termed Role-Playing Conversational Agents (RPCAs), aim to emulate the conversation behaviors and patterns of specific characters via LLMs. Role-playing agents show considerable promise and are poised to substantially advance the areas of gaming, literature, and creative industries [1]–[6]. Currently, the implementation of role-playing agents can be categorized into two primary methodologies. The first strategy enhances the role-playing capabilities of LLMs through prompt engineering and generative enhancement techniques. This approach equips LLMs with character-specific data within the context, capitalizing on the advanced in-context learning capabilities of modern LLMs. For instance, ChatHaruhi [3] developed a RAG (Retrieval-Augmented Generation) system that leverages historical dialogues from iconic scenes to facilitate learning from a limited number of examples, thus capturing the personality traits and linguistic styles of characters. Conversely, RoleLLM [5] introduced RoleGPT, which uses role-based prompts for GPT models.

The other type of role-playing approach involves pre-training or fine-tuning LLMs with collected character data, thereby customizing LLMs for specific role-playing scenarios. In [4], dialogue and character data from the Harry Potter novels were utilized to train agents capable of generating responses that align accurately with the context of the scene and the inter-character relationships. Character-LLM [1] developed scenarios using ChatGPT to create conversational data, subsequently training a language model with meta-prompts and these conversations. This project implemented strategies to mitigate the creation of character discrepancies in the model training dataset, such as memory uploads and protective memory enhancements. RoleLLM [5] employed GPT to formulate question-answer pairs based on scripts, presenting them in a triplet format consisting of the question, answer, and confidence level. Incorporating a confidence metric significantly enhanced the quality of the generated data. CharacterGLM [2] trained an open-source character model using data from multiple characters. This approach embeds role-specific knowledge directly into the model's parameters.

While existing studies of role-playing agents consider the character profile, relationships, and attributes relevant to the dialogue, they often overlook a critical element—the emotional factor of the characters. Our emotional RAG framework is designed on the prompt engineering technique, in which the LLMs are not required to be pre-trained or fine-tuned in role-playing agents.

### B. Memory RAG in LLM Applications

In role-playing agents, memory is an important factor for characters to maintain their personality traits. Retrieval Augmented Generation (RAG) technology is widely used [35] to access the related memory to enhance the generation of role-playing agents, termed Memory RAG. For example, an LLM-based automatic agent architecture proposed in [36] contains four components: a profiling module, a memory module, a planning module, and an action module. Among these, the memory module is crucial for the design of the agent architecture. It takes charge of obtaining information from the environment and utilizes these recorded memories to enhance future actions. The memory module enables the agent to accumulate experiences, evolve autonomously, and act in a manner that is more consistent, rational, and efficient [14].

The research on memory design in various LLM applications can be summarized into two categories. The first is capturing and storing intermediate states from past model reasoning as memory content. These memories are then retrieved as needed to support the generation of current responses. For instance, MemTRM [37] maintains past key-value pairs and employs the query vector of the current input to conduct K-nearest neighbor searches, applying mixed attention to both the current input and the past memories. However, MemTRM encounters challenges with memory obsolescence during training. To address this, LongMEM [38] separates the processes of memory storage and retrieval. This strategy is particularly tailored for open-source models and might necessitate adaptive training to effectively integrate the contents of the memory library. The second type of memory design scheme involves providing memory support via an external memory library. This external memory can take various forms, enhancing the system's ability to manage and retrieve information efficiently. One such implementation is MemoryBank [10], which stores past conversations, event summaries, and user characteristics in a vector library format. The use of vector similarity calculations significantly accelerates the memory retrieval process, allowing for rapid access to relevant past experiences and data. AI-town [12] uses a linguistic approach by preserving memory in natural language. It introduces a reflection mechanism that under specific conditions, transforms straightforward observations into more abstract and higher-order reflections. This system considers three critical factors during the retrieval process: the relevance, recency, and importance of memory, ensuring that the most pertinent and contextual information is retrieved for use in ongoing interactions.

In LLM-based role-playing agents, the memory unit typically operates via the second method, incorporating external memory libraries to enhance character authenticity. For example, in ChatHaruhi, the character agent retrieves dialogue from iconic scenes to enrich character development and interactions. Despite a large amount of research on memory RAG technique, achieving greater human-like response is still an open and unexplored area. Inspired by cognitive research in psychology, we make an initial attempt to incorporate the emotional factor to emulate human cognitive processes in the memory-recalling process, making the response of LLMs more emotionally resonant and human-like.

## V. Conclusions

In this paper, we make an initial attempt to incorporate emotional memory to enhance the performance of role-playing

agents. A novel emotional RAG framework with four retrieval strategies is proposed to make role-playing agents more emotional and human-like in conversations. Extensive experiments on various characters on three public datasets demonstrate the effectiveness of our method in maintaining the personality traits of characters. We believe that imbuing emotions into role-playing agents is a pivotal research direction. In our current study, we conduct emotional RAG on an intuitive memory mechanism. In future work, we will attempt to incorporate the emotional factor into more advanced memory organization and retrieval schemes.

## REFERENCES


[1] Y. Shao, L. Li, J. Dai, and X. Qiu, "Character-llm: A trainable agent for role-playing," *arXiv preprint arXiv:2310.10158*, 2023.

[2] J. Zhou, Z. Chen, D. Wan, B. Wen, Y. Song, J. Yu, Y. Huang, L. Peng, J. Yang, X. Xiao *et al.*, "Characterglm: Customizing chinese conversational ai characters with large language models," *arXiv preprint arXiv:2311.16832*, 2023.

[3] C. Li, Z. Leng, C. Yan, J. Shen, H. Wang, W. Mi, Y. Fei, X. Feng, S. Yan, H. Wang *et al.*, "Chatharuhi: Reviving anime character in reality via large language model," *arXiv preprint arXiv:2308.09597*, 2023.

[4] N. Chen, Y. Wang, H. Jiang, D. Cai, Y. Li, Z. Chen, L. Wang, and J. Li, "Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters," *arXiv preprint arXiv:2211.06869*, 2022.

[5] Z. M. Wang, Z. Peng, H. Que, J. Liu, W. Zhou, Y. Wu, H. Guo, R. Gan, Z. Ni, M. Zhang *et al.*, "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models," *arXiv preprint arXiv:2310.00746*, 2023.

[6] J. Chen, X. Wang, R. Xu, S. Yuan, Y. Zhang, W. Shi, J. Xie, S. Li, R. Yang, T. Zhu *et al.*, "From persona to personalization: A survey on role-playing language agents," *arXiv preprint arXiv:2404.18231*, 2024.

[7] K. Lu, B. Yu, C. Zhou, and J. Zhou, "Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment," *arXiv preprint arXiv:2401.12474*, 2024.

[8] M. Shanahan, K. McDonell, and L. Reynolds, "Role-play with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.16367

[9] M. Yan, R. Li, H. Zhang, H. Wang, Z. Yang, and J. Yan, "Larp: Language-agent role play for open-world games," 2023. [Online]. Available: https://arxiv.org/abs/2312.17653

[10] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "Memorybank: Enhancing large language models with long-term memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 724–19 731.

[11] K. Zhang, F. Zhao, Y. Kang, and X. Liu, "Memory-augmented llm personalization with short-and long-term memory coordination," *arXiv preprint arXiv:2309.11696*, 2023.

[12] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22.

[13] Z. Wang, Y. Y. Chiu, and Y. C. Chiu, "Humanoid agents: Platform for simulating human-like generative agents," 2023. [Online]. Available: https://arxiv.org/abs/2310.05418

[14] Z. Zhang, X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen, "A survey on the memory mechanism of large language model based agents," *arXiv preprint arXiv:2404.13501*, 2024.

[15] S. Ge, C. Xiong, C. Rosset, A. Overwijk, J. Han, and P. Bennett, "Augmenting zero-shot dense retrievers with plug-in mixture-of-memories," 2023. [Online]. Available: https://arxiv.org/abs/2302.03754

[16] B. Wang, X. Liang, J. Yang, H. Huang, S. Wu, P. Wu, L. Lu, Z. Ma, and Z. Li, "Enhancing large language model with self-controlled memory framework," 2024. [Online]. Available: https://arxiv.org/abs/2304.13343

[17] Y. Yu, H. Li, Z. Chen, Y. Jiang, Y. Li, D. Zhang, R. Liu, J. W. Suchow, and K. Khashanah, "Finmem: A performance-enhanced llm trading agent with layered memory and character design," 2023. [Online]. Available: https://arxiv.org/abs/2311.13743

[18] X. Zhu, Y. Chen, H. Tian, C. Tao, W. Su, C. Yang, G. Huang, B. Li, L. Lu, X. Wang, Y. Qiao, Z. Zhang, and J. Dai, "Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory," 2023. [Online]. Available: https://arxiv.org/abs/2305.17144

[19] C. Xiao, P. Zhang, X. Han, G. Xiao, Y. Lin, Z. Zhang, Z. Liu, and M. Sun, "Infllm: Training-free long-context extrapolation for llms with an efficient context memory," 2024. [Online]. Available: https://arxiv.org/abs/2402.04617

[20] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "Memgpt: Towards llms as operating systems," 2024. [Online]. Available: https://arxiv.org/abs/2310.08560

[21] A. Modarressi, A. Imani, M. Fayyaz, and H. Schütze, "Ret-llm: Towards a general read-write memory for large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.14322

[22] J. Kang, R. Laroche, X. Yuan, A. Trischler, X. Liu, and J. Fu, "Think before you act: Decision transformers with working memory," 2024. [Online]. Available: https://arxiv.org/abs/2305.16338

[23] L. Liu, X. Yang, Y. Shen, B. Hu, Z. Zhang, J. Gu, and G. Zhang, "Think-in-memory: Recalling and post-thinking enable llms with long-term memory," 2023. [Online]. Available: https://arxiv.org/abs/2311.08719

[24] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.16291

[25] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[26] G. H. Bower, "Mood and memory." *American psychologist*, vol. 36, no. 2, p. 129, 1981.

[27] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, "C-pack: Packaged resources to advance general chinese embedding," 2023.

[28] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[29] X. Wang, Y. Xiao, J. tse Huang, S. Yuan, R. Xu, H. Guo, Q. Tu, Y. Fei, Z. Leng, W. Wang, J. Chen, C. Li, and Y. Xiao, "Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews," 2024.

[30] Q. Tu, S. Fan, Z. Tian, and R. Yan, "Charactereval: A chinese benchmark for role-playing conversational agent evaluation," *arXiv preprint arXiv:2401.01275*, 2024.

[31] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues." 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:149343234

[32] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 320–335.

[33] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[34] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[35] C. Hu, J. Fu, C. Du, S. Luo, J. Zhao, and H. Zhao, "Chatdb: Augmenting llms with databases as their symbolic memory," *arXiv preprint arXiv:2306.03901*, 2023.

[36] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[37] Y. Wu, M. N. Rabe, D. Hutchins, and C. Szegedy, "Memorizing transformers," *arXiv preprint arXiv:2203.08913*, 2022.

[38] W. Wang, L. Dong, H. Cheng, X. Liu, X. Yan, J. Gao, and F. Wei, "Augmenting language models with long-term memory," *Advances in Neural Information Processing Systems*, vol. 36, 2024.