DRMR: An Immersing Oriented Role-Playing Framework with Duplex Relationship Modeling

Anonymous ACL submission

Abstract

Role-playing is an emerging application of large language models (LLMs), allowing users to be immersed in conversations with virtual characters by mimicking their tones and background knowledge. It can be applied in various scenarios such as gaming and virtual reality systems. However, existing methods ignore two challenges: (1) ignoring the relationship with the role played by the user will diminish the immersive experience of the user; (2) insufficient understanding of the character's background knowledge may lead to inconsistent dialogue. In this paper, we introduce the **D**uplex **R**elationship **M**odeling based **R**ole-play framework (DRMR), a novel role-playing framework designed to enhance the immersion of user when interacting with the role-play model. We first propose a graph-based relationship modeling method, utilizing graph structures to model the duplex relationship between the user and the model's played characters. In order to better extract useful personalized information about roles from historical dialogues, we construct a role memory consisting of the description of the duplex relationship. To avoid generating an inconsistent response, we iteratively verify the generated response by updating the role memory according to the current dialogue context. Extensive experiments on benchmark dataset demonstrate the effectiveness of DRMR in enhancing user immersion in role-playing interactions[1].

1 Introduction

In recent years, large-scale language models (LLMs) have made significant advancements in numerous classical natural language processing tasks (Zhang et al., 2020; Wei et al., 2022a; Lei et al., 2023; Zhang et al., 2023). This has also brought several new paradigms in natural language processing, transitioning gradually from better accomplishing traditional natural language tasks to some new applications such as tool usage (Qin et al., 2023a; Zhuang et al., 2023; Qin et al., 2023b), LLM-based multi-agent systems (Park et al., 2023; Li et al., 2023b; Chen et al., 2023a), embodied intelligence methods for manipulating robots (Huang et al., 2023; Jang et al., 2021; Mahadevan et al., 2024) and role-playing (Li et al., 2023a; Wang et al., 2023b; Chen et al., 2023b). Role-playing aims to enable LLMs to portray specific characters/roles[2] (*e.g.,* characters in movies and TV dramas, historical figures, etc.) to meet user needs. These methods have been widely used in interactive games (Light et al., 2023; Xu et al., 2023b), virtual reality systems (Sapkaroski et al., 2022), and psychological counseling (Zheng et al., 2023; Hsu et al., 2023).

On one hand, some of the existing role-playing

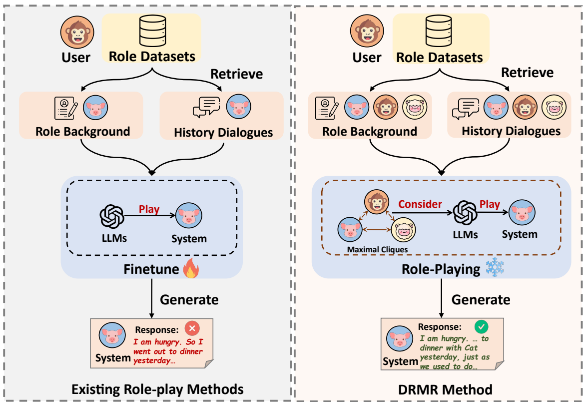

Figure 1: Comparison between existing role-play methods and our proposed DRMR. Most previous methods usually annotate large amounts of data and then fine-tune the LLM, and they typically consider only the information of the role played by the model, neglecting the duplex relationship information between the roles played by the user.

[1]Code is available at https://anonymous.4open.science/r/DRMR.

[2]These two terms are interchangeably used.

methods (Li et al., 2023a; Chen et al., 2023b; Zhou et al., 2023; Wang et al., 2023b) focus on fine-tuning LLMs by either constructing more role-playing datasets or data augmentation. This enables large models to understand the background knowledge and language style characteristics of roles, thus achieving better role imitation. However, this not only relies on acquiring a large amount of data but also considerable training time and GPU resources for fine-tuning LLM. On the other hand, some methods (Zhang et al., 2018; Zhong et al., 2020; Xu et al., 2023a) attempt to achieve this by allowing users to define role profiles as in-context instructions, but this requires lengthy input from users to define roles, which adversely affects user experience.

In real-world applications, users will provide only a brief role profile and several previous dialogues for the role-play model. Intuitively, since not all the details of the role can be comprehensively defined in the profile, models often struggle to generate consistent responses, such as an ancient figure writing code. This deviation from the background of the character in responses also diminishes the user immersion. Therefore, the **first challenge** lies in deeply understanding the brief profile and making full use of the given data to generate dialogues that are consistent with the background of the character.

Furthermore, the majority of existing methods only incorporate personalized information about the role played by the model (*a.k.a.,* **simplex** relationship), ignoring the role profile and relationship played by the user (*a.k.a.,* **duplex** relationship). However, an immersive role-playing experience requires not only mimicking the tone and knowledge background of the character being played but also involving the user in the scenario where the character is situated. It is crucial for role-playing models to understand the duplex relationship between the character played by the user and the character played by the model, as this greatly contributes to the immersive experience of role-playing. Thus, the **second challenge** lies in how to model the interpersonal relationships between the roles played by the user and the model when role-playing.

In this paper, we propose the **D**uplex **R**elationship **M**odeling based **R**ole-play framework (DRMR) method, a role-playing framework aimed at enhancing immersion of user experience. Given a brief role profile provided by the user and several historical dialogues, our approach employs two novel methods to enhance understanding of the character's background and achieve duplex relationship modeling for both the model and the user-played roles. To achieve duplex relationship modeling, we propose a *maximal-cliques-based role relationship modeling* method based on a role relation graph. By using the maximal cliques representing both the model and the user's played characters along with their shared background information, we construct a role memory to summarize the useful relationship information, thereby enhancing user immersion.

Then we introduce an *iterative response revision* method, which iteratively revises the model responses by retrieving more related dialogues and updating the role memory, thus generating responses that align with the background of the character. Extensive experiments conducted on a benchmark dataset demonstrate the effectiveness of our proposed DRMR, and we can find that our proposed model can enhance the immersion of the user when chatting with the role-playing system. Our contributions of this work are as follows:

- We propose DRMR, which is a role-playing framework to enhance the immersion of user experience.
- We introduce maximal-cliques-based role relationship modeling to incorporate the duplex relationship of both characters played by the user and the model.
- We propose the iterative response revision method which iteratively verifies the consistency of the response and revises the response by using updated role memory.
- Experimental results on benchmark dataset illustrate the superiority of DRMR.

2 Related Work

Role-playing is an important application of LLMs, aimed at simulating a character comprehensively by using events from movies, TV shows, or historical figures to achieve immersive interaction with users. ChatHaruhi (Li et al., 2023a) is an earlier method that utilizes LLMs to implement role-playing which establishes a character dialogue database and introduces a retrieval-enhanced role-playing framework. Character-LLM (Shao et al., 2023) focuses on modeling character memories, reconstructing scene-based memories using WikiData, and

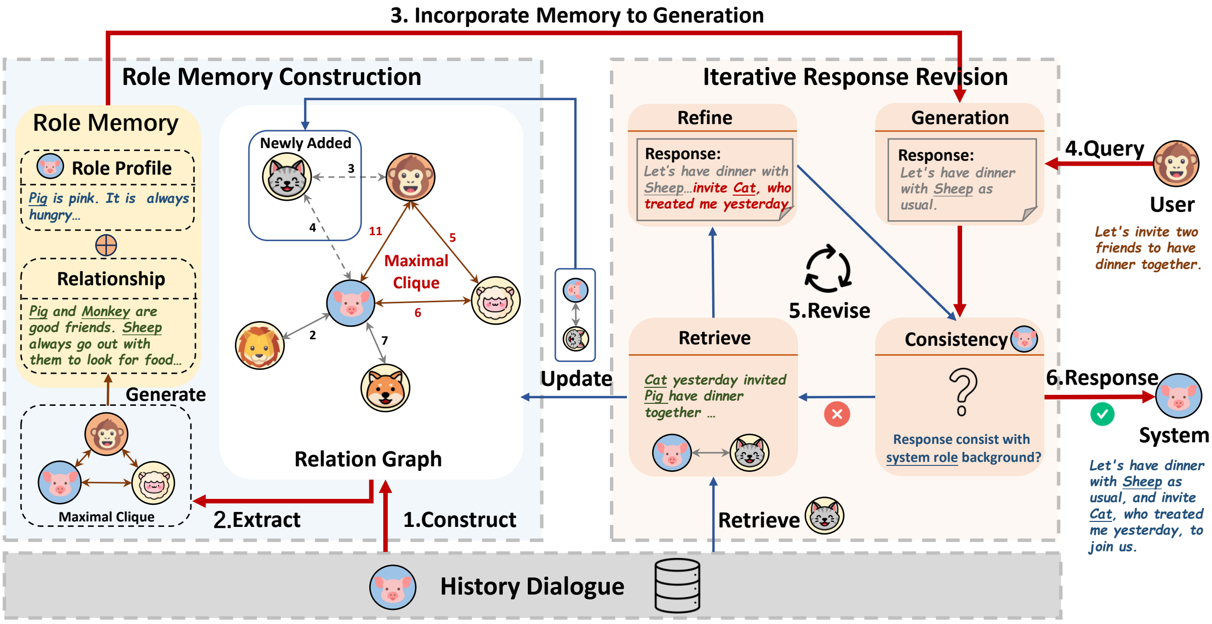

Figure 2: Overview of **D**uplex **R**elationship **M**odeling based **R**ole-play framework (DRMR) which has three main steps: (1) We first construct a relation graph using the historical dialogue of the roles and extract the maximal cliques from the graph to build the role memory; (2) We generate the response to the user by incorporating the role memory; (3) We employ the iterative response revision framework to verify the revise the response which ensures the response is consistent with the background of the role.

adopting protective experiences to mitigate the hallucination of response. CharacterGLM (Zhou et al., 2023) further develops a multi-turn role-playing dialogue system based on fine-tuning LLMs, using character profiles, dialogues, and a large amount of crowd-sourcing data as training dataset. HPD (Chen et al., 2023b) is a dataset for playing the role of Harry Potter integrating extensive and detailed background information to better match LLMs with the characteristics of Harry Potter. RoleLLM (Wang et al., 2023b) proposes a role-playing model based on instruction tuning by maintaining specific knowledge and speaking tones of characters by combining in-context instructions.

However, the majority of existing role-playing methods require fine-tuning LLMs through annotating large datasets, which demands significant computational resources and data labeling efforts. Moreover, most existing works do not consider the interpersonal relationships between the characters portrayed by the user and the model, leading to model-generated responses that may not align with the current conversational context, thereby diminishing the immersive experience of the user.

3 DRMR Methodology

In this section, we detail the **D**uplex **R**elationship **M**odeling based **R**ole-play framework (DRMR). An overview of DRMR is shown in Figure 2.

3.1 Problem Formulation

Given the brief profiles P_m and P_u of the role E_m to be played by the model and the role E_u played by the user, along with several historical dialogues $D = \{(E_1, U_1), (E_2, U_2), \dots, (E_L, U_L)\}$ as the input to our DRMR, where E_i denotes the speaker of utterance U_i.

The user plays the role E_u and engages in a dialogue of T turn with the role E_m played by the model, denoted as the current dialogue context $C = \{(c_1^u, c_1^m), (c_2^u, c_2^m), \dots, (c_T^u, c_T^m), (c_{T+1}^u)\}$, where c_i^u represents the i-th utterance of the user, and c_i^m represents the i-th response of the model. Based on this input, our model aims to generate responses c_{T+1}^m of role E_m to the user query c_{T+1}^u.

3.2 Role Memory Construction

When we engage in conversation with others, our minds not only contain information about ourselves but also the profile of the other person. And we also recall past experiences with this person (*e.g.*, travels together previously) and information about people associated with them (*e.g.,* their parents). Intuitively, it is crucial for humans to recall this information from memory during conversations which makes human-to-human dialogue natural; otherwise, conversations would become disjointed. Therefore, to enhance immersion in role-playing systems, we propose a *role memory* M to store relation information about the user role E_u, model role E_m, and other related roles. To construct the role memory M, we propose using the graph to explicitly model the relationship of the roles, and summarize the structures of the graph into natural language descriptions of the relationship between roles. The role memory M contains several paragraphs describing the detailed profile of role E_u and the relationship between role E_u and E_m to mimic the mind of people when chatting with others. Initially, we use the role profile P_m as the initialization for the role memory M.

To recall the most related role information, we firstly utilize the dialogue context C as query to retrieve N dialogues from historical dialogues D which are relevant to the roles E_u and E_m, denoted as $D^c = \{(E_1^c, U_1^c), (E_2^c, U_2^c), \dots, (E_N^c, U_N^c)\}$. Specifically, we leverage the dense retrieval method as the semantic similarity measurement to retrieve the most relevant dialogues D^c from the historical dialogue D:

$$\phi = \cos\left(\text{Emb}(U), \text{Emb}(C)\right), U \in D, \quad (1)$$

where ϕ is the similarity score. We employ the pre-trained LLM as the text embedding function Emb and use the cosine to measure the similarity between dialogue representations. Finally, we take the top-N dialogues according to the score ϕ as the relevant dialogues D^c.

Then, we construct a user relationship graph G containing the roles E_u and E_m as well as other related roles retrieved $\{E_1^c, E_2^c, \dots, E_N^c\}$. When the two roles have conversations, an edge between these roles is added to the graph G. As the degree of association between roles varies, it is necessary to quantitatively measure the degree of association between roles when constructing the role relationship graph G. In this paper, we propose using LLM to evaluate the relationship weights between nodes:

$$s_{i,j} = \text{EdgeScore}(I_{\text{ES}}\{U_i^c, U_j^c\}) \in \{1, 2, 3, 4, 5\}, \quad (2)$$

where $s_{i,j}$ indicates the relation score between the role E_i and E_j, $\{U_i^c, U_j^c\}$ represents the historical dialogue between the role E_i and E_j, and I_{ES} denotes the instruction we used to prompt the LLM to score the relationship between two roles:

> You are a Character Event Assessment Assistant. Please carefully evaluate and score, reflecting the importance of the characters $\{E_i$ and $E_j\}$ in the following dialogue. Your scoring range is from 1 to 5...
> [history dialogue]
> Refer to the following standards for scoring:
> 1 point: The character barely participates in the event, having no impact on its development.... Please provide a brief explanation for your score, assessing the importance of $\{E_i$ and $E_j\}$ based on the above standards.

Due to the limited context length of the LLM, it is not feasible to consider all the information of nodes and edges in a single dialogue turn. In the graph, since maximal cliques can represent a subset of vertices in a graph where every two distinct vertices are adjacent, providing a dense connection indicative of a strong relationship or relevance among the included vertices. Thus, we employ a relation maximal clique algorithm on graph G to obtain a subgraph G' containing a maximal clique comprising several roles most relevant to the roles E_u and E_m:

$$G' = \underset{G' \in G^\star}{argmax} \sum_{i,j \in G'} s_{i,j}, where\ u, m \in G' \quad (3)$$

where $G^\star$ is the set of the maximal cliques. Since the maximal cliques are not always unique in graph G, we utilize the sum of relationship weights within the maximal clique as a selection criterion.

Subsequently, we utilize the role relation contained in the subgraph G' to expand the role memory M. Each edge in the maximal clique subgraph G' represents a dialogue between two roles. We retrieve the top K dialogues C' most relevant to the current dialogue C from these edges in G' to update the role memory M. And we use the same retrieval method as in Equation 1. In order to enable the role-playing model to better understand the relationships between these relevant roles and incorporate these relationships into dialogue generation, we utilize LLM to extract descriptions of relationships between characters from these relevant dialogue data C' and summarize the events described in the dialogue:

$$m = \text{MemBuild}(C', P_m, P_u), \quad (4)$$

where m is a new role memory record (*a.k.a.,* a paragraph that describes the detailed relation between two roles). And the operation MemBuild is a chain-of-thought (Wei et al., 2022b) based prompting method that prompts the LLM to summarize the relationship between two roles is as follows:

> You will play a role that depicts your relationship with another character through a series of events that have occurred...
> First, you will play $\{E_m\}$...
> Next, briefly describe your relationship with $\{E_u\}$ from $\{E_m\}$'s first-person perspective.
> Third, ...
> To assist you in this task, here are some events that have occurred between $\{E_m\}$ and $\{E_u\}$:
> $\{C'\}$
> Please output $\{E_m\}$'s first-person evaluation of$\{E_u\}$. The description should be concise and relevant.

Finally, we append the new memory record m into the role memory M.

3.3 Iterative Response Revision

Based on the role memory M and dialogue context C, we prompt LLM to generate responses c^m_{T+1} for the role E_m:

$$c^m_{T+1} = \text{GenResp}(M, C). \tag{5}$$

However, existing works (Marcus, 2020; Ji et al., 2023), have found that directly generating the response of the role may sometimes be inconsistent with the character background, such as an ancient figure writing the Python code. Therefore, we propose an *iterative response revision* method. After generating a response c^m_{T+1}, we employ an LLM to first validate whether the generated response c^m_{T+1} is consistent with the content of the role memory with the character background:

$$h = \text{Verify}(M, c^m_{T+1}, c^u_{T+1}) \in \{1, 2, 3, 4, 5\}, \tag{6}$$

where h represents the consistency score, where a score of 1 indicates the lowest consistency and 5 indicates the highest consistency. The instruction of the Verify is:

> You are a helpful director, focused on the setting of the character $\{E_m\}$. Please give a score following the steps, your scoring range is from 1 to 5...
> $\{E_u\}$'s question is $\{c^u_{T+1}\}$, and the $\{E_m\}$'s response is: {answer}.
> The setting for $\{E_m\}$ is {M}.
> Please assess how well the answer matches the setting of $\{E_m\}$. Explain the reason and then give a score.
> I will provide you with some sample outputs. Their main purpose is to help you understand the output format and judgment criteria:
> {Examples}

When the consistency score $h \leq \alpha$, where α is a threshold hyper-parameter, we revise the generated responses to align them with the background information of the role. To give more personalized information about the role for better revising the response, we retrieve K relevant dialogues $D^r = \{(E^r_1, U^r_1), (E^r_2, U^r_2), \ldots, (E^r_K, U^r_K)\}$ from the historical dialogues D by using the user's last utterance c^u_{T+1} as the query. The newly retrieved dialogues D^r are then used to update the role relationship graph G, and the weights of the newly added nodes and their associated edges are updated according to Equation 2. After updating the relation graph G, following the previous steps, the maximal clique is extracted again, and we generate a new role memory record m and append it into the role memory M (introduced in Equation 4). Finally, we re-generate the response c^m_{T+1} based on the updated role memory M.

4 Experimental Setup

4.1 Evaluation Metrics

Role-play aims to customize LLMs to simulate various characters or personas with distinct and precise attributes, which provides a more nuanced interaction experience for users and makes LLMs more familiar (Shanahan et al., 2023; Wang et al., 2023a). Consequently, immersion can be defined as the consistency of the model's responses with the role's personality and memory, as well as the familiarity felt by the user. This familiarity arises from the relationship between the model and the user's role. For example, family members feel familiar with each other due to their relationship, whereas passersby feel alienated from each other because of the absence of a relationship and shared experiences. Therefore, to quantitatively measure the immersion performance of DRMR, we propose three evaluation metrics in our paper: (i) **Personality** (*Pers.*): Evaluate whether the responses align with the personality traits and linguistic habits. It also verifies whether their attitude towards current events is reasonable according to the dialogue history. (ii) **Memorization** (*Mem.*): Assess the recollection of character-relevant experiences and knowledge, ensuring alignment with the background of the character. Relevant historical dialogues are retrieved to determine whether specific events mentioned in the dialogue history are reflected in the responses. (iii) **Relation** (*Rela.*): Evaluate the degree to which the responses correspond to the relationship between the user's and the model's portrayed character. Considering the relationship between both roles (such as lover, family member, foe, etc.), it judges whether the generated responses align with the relationship. To evaluate the generated response according to the above criteria, we employ an LLM and prompt it with elaborate descriptions of the criteria to quantitatively evaluate the response. The LLM scores each response for the above three aspects separately using a scale of 1-5. Detail instructions can be found in Appendix A.

4.2 Dataset

In the experiments, we employ a Chinese benchmark role-play dataset CharacterEval (Tu et al., 2024), which contains 77 characters and 1,785 high-quality multi-turn dialogue contexts. Following Tu et al. (2024), we use the whole CharacterEval dataset as the test set to evaluate our model and baselines.

4.3 Baselines

We compare our method with several LLM-based role-play methods, including: **`RoleGPT`** (Wang et al., 2023b), **`CharacterGLM`** (Wang et al., 2023b), **`CharacterGLM`** (Wang et al., 2023b), **`Qwen`** (Bai et al., 2023), **`ChatGLM`** (Zhipuai, 2023). More descriptions about this method can be found in Appendix B

We employ three variants of DRMR: **DRMR-C**, **DRMR-Q** and **DRMR-G** with `ChatGPT`, `Qwen` and `ChatGLM` as the backbone respectively. And we also employ two ablation models: (i) **`DRMR w/o Revison`**: We remove the verify step (introduced in Equation 6) and directly use the output of the model as the response. (ii) **`DRMR w/o RoleMem`**: We remove the graph-based role memory construction module and directly use the related dialogue as a prompt to the LLM.

4.4 Implementation Details

In our experiments, all DRMR-C variants and the `RoleGPT` use the gpt-3.5-turbo-0125 version, the DRMR-G variant and `ChatGLM` baseline use the glm-3-turbo API[3], and the DRMR-Q and `Qwen` are implemented using open-source Qwen-14B-chat as the backbone. In our model, we use the temperature 1.0 in most steps, and the temperature 0.1 during the verify step in Equation 6. For the consistency threshold used in the verify step, we set $\alpha = 4$. And we employ $N = 3$ and $K = 2$ retrieved dialogues when constructing role memory and revising the response respectively. We use the *text-embedding-ada-002* model of OpenAI as the embedding model used in Equation 1. We use the Bron-Kerbosch algorithm (Bron and Kerbosch, 1973) in Equation 3 to find the maximal clique.

[3] `https://maas.aminer.cn/dev/api#glm-3-turbo`

5 Experimental Results

5.1 Overall Performance

Table 1 shows the performance of our proposed DRMR and baselines in terms of three metrics. We can find that DRMR shows superior performance in terms of all metrics compared to their corresponding backbone LLM. Furthermore, we find that our DRMR achieved a greater improvement in terms of the relation metric compared to its backbone LLM, indicating that our role memory construction contributes to generating higher-quality responses. This phenomenon indicates that our proposed DRMR is capable of generating a response that mimics the personality of the role in a higher quality.

5.2 Human Evaluation

To better examine whether the generated responses align with human intuitive judgment and further evaluate the enhancement of immersion, we employ the human evaluation to further assess the baselines. We use three criteria for the human evaluation, including: (i) **Personality**: Assessing whether the responses align with the personality of the character; (ii) **Contextualize**: Determining if the responses correspond to the relevant events or background associated with the character in the ongoing conversation; (iii) **Relationship Consistency**: Evaluating whether the responses of the character align with the relationship between the two roles in the current conversation. We randomly select 300 generated results from each model and hire 3 educated annotators with master's degree to conduct double-blind annotation on randomly shuffled results. The score range of each aspect is 1-5. From Table 1, we observed that the DRMR outperforms all baselines. And the ranking of human evaluation is also consistent with the LLM-based automatic metrics, which

Method	**Pers.** (↑)	**Mem.** (↑)	**Rela.** (↑)	**Human** (↑)
`CharacterGLM`	3.21	3.45	3.41	0.80
`ChatGLM`	3.68	4.01	3.67	0.81
`Qwen`	3.78	4.08	3.71	0.83
`RoleGPT`	3.39	3.47	3.49	0.75
DRMR-G	3.83(4.08%)	4.12(2.74%)	**4.05**(10.35%)	**0.87** (7.41%)
DRMR-Q	**3.89**(2.91%)	**4.15**(1.72%)	3.93(5.93%)	0.85(2.41%)
DRMR-C	3.65‡(7.67%)	3.76‡(8.36%)	3.93‡(12.61%)	0.84‡(12.00%)
`DRMR-C w/o Revision`	3.62	3.68	3.83	0.75
`DRMR-C w/o RoleMem`	3.61	3.65	3.72	0.76

Table 1: Comparison of the response quality. ‡ indicates significant improvement over `RoleGPT` with $p \leq 0.01$ according to a Student's t test. The value in parentheses indicates the proportion of improvement compared to the LLM backbone.

also demonstrate the effectiveness of our proposed LLM-based evaluation method.

5.3 Ablation Study

To evaluate the effectiveness of each module in DRMR, we also conduct ablation studies with model DRMR-C, and the results are shown in Table 1. From this table, it can be observed that both ablation models perform worse than DRMR-C in terms of all metrics, indicating the effectiveness of the role memory and iterative response revision. We found that the `DRMR w/o RoleMem` method achieved lower scores compared to other ablation models, indicating the effectiveness of modeling the relationship between roles in our approach.

5.4 Case Study

We analyzed the impact of role memory and history dialogues in our methods through two cases.In the first case, police officer Anxin interrogates Gao, who is a criminal. In `RoleGPT`'s response, Gao admit to criminal interactions, which misaligns with Gao's background. In contrast, in the response generated by DRMR, Gao denies such interactions and cheats Anxin, maintaining character consistency and role-play immersion. In the second case, Tong interacts with Bai and mentions Zhan. DRMR effectively extracted the relationships among the three characters from historical dialogues and applied them smoothly in the conversation. However, influenced by these relationships, it did not adequately capture the subtle emotions in the dialogue. More details and analysis can be found in Appendix D

Method	**Unseen** (↑)	**Seen** (↑)
`RoleGPT`	0.58	0.77
DRMR-C	**0.67**(15.51%)	**0.85** (10.38%)

Table 2: Comparison of the response quality on two subsets of the CharacterEval. The subset "Unseen" indicates that the content of the TV show has not been used as the pre-train data of the backbone LLM, while the characters in the "Seen" subset have been shown when pre-training LLM.

6 Discussion

6.1 Analysis on Unseen Character

Due to the extensive use of web data for pre-training, LLM backbone is already familiar with most of the roles in the dataset CharacterEval. To validate the generalization ability of the model, we separate the data from CharacterEval for some newly released TV shows, which have not been trained on LLM. Thus, we divided the CharacterEval into two subsets, **seen** and **unseen**, not only based on the release time of the TV show but also by asking LLM if it knows the characters in the script. Table 2 shows the comparison between our proposed DRMR and `RoleGPT` on these two subsets. From the results, it can be seen that our method exhibits superior performance on both subsets, demonstrating better generalization ability of our DRMR. We can also find that both methods achieve higher scores on the seen dataset compared to the unseen dataset. As LLM has been trained on many data related to the role during the pre-training phase, it has a better understanding of

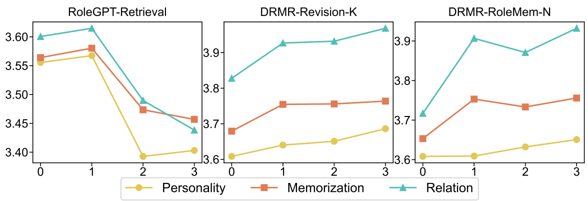

Figure 3: Performance of using different numbers of retrieved historical dialogues. The middle and right figures show the performance of retrieving different historical dialogues when revising responses and constructing role memory, respectively.

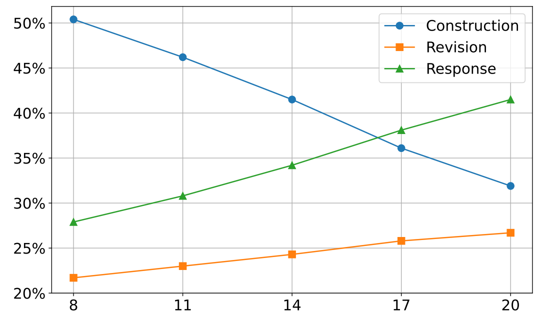

Figure 4: Token consumption proportion in different modules with different dialogue turns. Construction, Revision, and Response respectively represent the proportion of role memory construction, response generation, and iterative response revision modules.

the role compared to simply providing in-context information about the role. Due to the same reason, LLM may not fully understand the background of unseen characters, it cannot assess the quality of the response comprehensively. In this experiment, we employ human evaluation on 150 generated responses for each subset respectively, which uses the same criteria as in § 5.2.

6.2 Analysis of Using Different Numbers of Retrieved Dialogues

In § 3.2 and § 3.3, we employ the dense retrieval method (Lewis et al., 2020) to find semantically related dialogues from historical dialogues of the role to enhance the role memory and revise the response. In this section, we explore the influence of using different numbers of retrieval dialogues on the final performance. The baseline `RoleGPT` also employs a similar retrieval approach to extract relevant information about the roles from historical dialogue data. Figure 3 illustrates the impact of using different numbers of retrieval dialogues on the performance of `RoleGPT`, our model in the revision stage, and our model in the role memory construction stage, respectively. From Figure 3, we observe that our approach effectively enhances response quality by using more retrieval dialogues in both stages. This demonstrates that our method leverages prompting LLM to construct role memory more effectively, thus utilizing data more efficiently. On the other hand, the baseline method `RoleGPT` struggles to extract useful information from excessive data, leading to a decline in the quality of generated responses.

6.3 Analysis of Efficiency

Our method constructs and iteratively updates role memory using retrieved dialogues, which increases token consumption. Appendix Table 3 presents the token consumption details. The results show that, compared to `RoleGPT`, DRMR consumes more tokens but performs much better, especially in long dialogues. We also analyze the token consumption of different modules in DRMR. As shown in Figure 4, the memory construction module consumes more tokens than the other two modules in short dialogues. However, as the number of dialogue turns increases, its proportion continuously decreases. This is because memory construction is frequent only at the beginning of a dialogue. Once the relationship is built, token consumption for this part will no longer increase.

7 Conclusion

In this paper, we present the **D**uplex **R**elationship **M**odeling based **R**ole-play framework (DRMR), an LLM-based role-playing framework aiming at enhancing the immersion of the user. We first introduce a novel maximal-cliques-based graph method to establish a duplex role relationship between characters played by the user and the model. Next, we propose to leverage the reasoning ability of the LLM to summarize useful relationship information from the maximal cliques as a role memory, and then generate the response by incorporating the role memory. To enhance the consistency between the generated responses and the background knowledge of the role, we propose the iterative response revision which first verifies the consistency of the response with the background knowledge of the role and then retrieves related dialogues to update the role memory and revise the response. Experimental results on the benchmark dataset demonstrate the superiority of the DRMR in elevating user immersion in role-playing interactions.

Limitations

In this paper, we only use the text input of our model. In real-world scenarios, the multi-modal input and output (*e.g.,* images and videos) is a more popular form for users. As existing multi-modal LLMs are capable of encoding both textual and multi-modal information into vector representations and unifying modeling, our method can be readily adapted to accommodate multi-modal inputs in the role-play task. We plan to incorporate multi-modal information into role-playing tasks in our future work.

Ethics Statement

While LLMs have the potential to generate hallucination information, our proposed method employs an iterative response revision framework to generate the response that aligns with the role identity as much as possible. As role-playing methods are mostly applied in non-critical domains such as gaming, they are unlikely to raise significant ethical concerns. However, if such role-playing methods were to be applied for therapeutic purposes like psychological counseling, they should be used under the guidance of a mental health professional.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Coen Bron and Joep Kerbosch. 1973. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. 2023a. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023b. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. 2023. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. 2021. BC-z: Zero-shot task generalization with robotic imitation learning. In *5th Annual Conference on Robot Learning*.

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

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023a. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. From text to tactic: Evaluating LLMs playing the game of avalon. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative expressive robot behaviors using large language models. *arXiv preprint arXiv:2401.14673*.

Gary Marcus. 2020. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2023a. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023b. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Daniel Sapkaroski, Matthew Mundy, and Matthew Richard Dim, mock. 2022. Immersive virtual reality simulated learning environment versus role-play for empathic clinical communication training. *Journal of medical radiation sciences*, 69(1):56–65.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, et al. 2023a. Interactive natural language processing. *arXiv preprint arXiv:2305.13246*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023a. Towards zero-shot persona dialogue generation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023b. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via chatgpt for faithful summary generation. In *Conference on Empirical Methods in Natural Language Processing*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.

Zhipuai. 2023. Chatglm. `https://chatglm.cn`.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. ToolQA: A dataset for LLM question answering with external tools. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Prompt of Evaluation Method

To quantitatively evaluate the performance of the generated response, we prompt the LLM to score the response according to the role profile and historical dialogues. We design different prompts for each aspect of our evaluation criteria.

Prompt for personality evaluation

You will receive a response generated by an AI assistant that plays the role {model_role}. Your task is to evaluate whether the answer is consistent with the personality of {model_role} based on specific criteria and evaluation steps. The data provided is as follows:
[Personal Background]
{role_profile}
[Conversation History]
{context}
[Answer]
{model_output}
[Evaluation Steps]
1. Check the personal background to select the personality traits and preferences of the real character. 2. Examine the dialog history to identify the character's personality traits and preferences.
...
Finally, repeat the selected score in a new line.
[Example]
The following is an example, intended only as a reference for the output format and not included in the judgment. Based on his background and dialogue history, Gao is a ruthless and calculating person... It reflects Gao's personality traits and attitudes well, but could probably have more accurately conveyed his fearlessness and determination.
Therefore, the final score is: [4]

Prompt for memorization evaluation

You will receive a response generated by an AI assistant that simulates the character {model_role}. Your task is to evaluate whether the answer is consistent with the character's personal and event background based on specific criteria and evaluation steps. The data provided is as follows:
[Personal Background]
{role_profile}
[Event Background]
{history dialogues}
[Conversation History]
{context}
[Answer]
{model_output}
[Evaluation Steps]
...
[Example]
Based on the Event Background, I can infer that Gao is related to the kidnapping case......the answer is consistent with the relevant memory background of Gao.
Therefore, the final score is: [4]

Prompt for relation evaluation

You will receive a response generated by an AI assistant modeling the role {model_role}. Your task is to evaluate whether the {model_role}'s answer to question from {user_role} is consistent with the role relation between them, based on specific criteria and evaluation steps. The data provided is as follows:
[Role Relation]
{role_relation}
[Conversation History]
{context}
[Answer]
{model_output}
[Evaluation Steps]
...
[Example]
The answer "I do not intend to cooperate" demonstrates Gao's attitude toward Anxin, which is in line with their relationship as enemies...
Therefore, the final score is: [5]

B Descriptions of baselines

- **RoleGPT** (Wang et al., 2023b) elicit role-playing abilities in ChatGPT via dialogue-engineering-based role prompting, utilizing system instruction and retrieval augmentation, to generate customized responses for speaking style imitation.

- **CharacterGLM** (Wang et al., 2023b) is a closed-source LLM-based role-play online, which has been fine-tuned with many role-play corpus.

- **Qwen** (Bai et al., 2023) is an open-source LLM and we use the pre-train model with 14 billion parameters. We conduct role-playing as a prompt learning method that uses a single instruction with the same input data as our DRMR.

- **ChatGLM** (Zhipuai, 2023) is a closed-source LLM and we use the model with 130 billion parameters, and use the same prompt as Qwen.

C Token consumption

Model	Pers. (↑)	Mem. (↑)	Rela. (↑)	Tokn.
ALL				
RoleGPT	3.39	3.47	3.49	**24K**
DRMR-C	**3.65**(7.67%)	**3.76**(8.40%)	**3.93**(7.67%)	46K
Long				
RoleGPT	3.11	3.09	3.47	**62K**
DRMR-C	**3.42**(10.0%)	**3.52**(13.9%)	**4.04**(16.4%)	86K

Table 3: Comparison of token consumption. The "All" is the results on the entire dataset. The "Long" is the result of dialogues with more than 20 turns. The "Token" is the average token consumption for a multi-turn dialogue.

D Case study

D.1 The impact of role memory

Table 4 shows an example of responses generated by RoleGPT and DRMR-C. In this case, Anxin is a police officer, and Gao is a villainous character associated with criminal underworld activities. The dialogue occurs while Anxin, acting as a police officer, is investigating a case related to Gao. In fact, Gao should consider the identity of Anxin as a police officer and should not admit to frequent interactions with Lao Mo, let alone acknowledge himself as a member of the criminal underworld in the response of RoleGPT (indicated in red text). While this dialogue aligns with the facts of the plot, it does not correspond to the context of the conversation at that time, thus diminishing the immersion of the user.

In our DRMR, considering the role of Anxin, DRMR not only refrained from admitting to being a member of the criminal underworld, but also concealed the frequency of interactions with Lao Mo, aligning with the actual dialogue of the characters. We also show the role memory of DRMR when generating this response. The role memory explicitly points out the relationship between Anxin and Gao, and it also deduces the salient fact from historical dialogues that Gao vigorously concealed his acquaintance with Lao Mo (indicated in the green text).

From this case, we can find that our proposed method DRMR can generate a high-quality response that is correctly aligned with the role relationship and correct dialogue context with the help of role memory.

D.2 The impact of history dialogues

Table 5 shows an example of how DRMR transitions from retrieved content to generated responses. Within the DRMR framework, the model extracts relationships from dialogues between Tong and Bai, and between Tong and Zhan, incorporating these relationships into Tong's role memory(indicated in green text). During interaction with the model, when Tong faces a question from Bai, she proactively mentions Zhan, demonstrating that the model effectively extracts the complete relationships among the three characters from the dialogue history (indicated in green text).

Due to Tong's emotional sensitivity, when Bai clarifies that he was "kidding," Tong would typically be angry in the real scenario. However, based on the historical "couple" relation, Tong forgive Bai immediately (indicated in red text). Therefore, our method still needs improvement in terms of the granularity of personality and contextual emotions.

User Role: Anxin, **Model Role**: Gao **Dialogue Context** C: *Anxin*: When did you meet Lao Mo? *Gao*: It's been many years. We met through Tang Xiaohu. He wanted to rent a stall to do business but lacked experience, so Tang Xiaohu brought him to me and asked me to teach him. *Anxin*: Have you been in touch since then?
RoleGPT: There was a period when we were in close contact. We often discussed business and industry trends together. But later, as we both rose in the underworld, we gradually drifted apart. However, Lao Mo has always been an important friend to me in my heart. **DRMR**: We don't stay in touch very often, after all, we each have our own busy lives. But when there's a need, we still help each other out. ***RoleMemory of DRMR***: I see Anxin as a thorn in my side, a threat to my position and power. He's a police officer, and I'm a member of the criminal underworld. We are enemies, and I admit Anxin has surprising insight. He can find the elusive Lao Mo I've been hiding.

Table 4: Examples of response generated by different methods. The text in red indicates conflicting content with the plot, and the text in green indicates the supporting facts for generating consistent response.

User Role: Bai, **Model Role**: Tong

Building role memory through history dialogues:
`Marriage between Bai and Tong (relation: couple)`
Tong: Come and help me
Bai: What do you need help with?
...
Tong: Just put it on at the wedding. If you don't unveil me, I will...
`Conflict between Tong and Zhan due to Bai (relation: rivalry)`
Zhan: Do you need me to do anything?
Tong: Nothing. I just want to tell you that I won't let him go.
...
Tong: Please, I can't be without him.
Zhan: I feel the same.

Role memory of Tong Sometimes I feel that Bai is... but his humor and kindness sometimes make me unable to help but laugh... he is the perfect companion. Zhan is my rival; she has also pursued Bai before...
Dialogue context C: *Bai*: How do you feel about Zhan? ... *Tong*: I'm not trying to oppose you, but if you were to meet me and Zhan at the same time, who would you choose? *Bai*: I would choose Zhan. *Tong*: Why? What am I lacking compared to her? *Bai*: I was just kidding, I choose you, okay? *Tong*: That's good, if you choose me, I will be good to you

Table 5: Impact of the history dialogues in DRMR