

RoleRAG: Enhancing LLM Role-Playing via Graph Guided Retrieval

Yongjie Wang¹ and **Jonathan Leung¹** and **Zhiqi Shen²**

¹Alibaba-NTU Global e-Sustainability CorpLab (ANGEL), Nanyang Technological University

²College of Computing and Data Science, Nanyang Technological University

{yongjie.wang, jonathan.leung, zqshen}@ntu.edu.sg

Abstract

Large Language Models (LLMs) have shown promise in character imitation, enabling immersive and engaging conversations. However, LLMs often generate content that is irrelevant or inconsistent with a character’s background. We attribute these failures to: 1) the inability to accurately recall character-specific knowledge due to entity ambiguity; and 2) a lack of awareness of the character’s cognitive boundaries. This paper introduces RoleRAG, a retrieval-based framework that combines efficient entity disambiguation for knowledge indexing with a boundary-aware retriever to extract contextually appropriate content from a structured knowledge graph. We conducted extensive experiments on role-playing benchmarks and demonstrate that RoleRAG’s calibrated retrieval enables both general LLMs and role-specific LLMs to exhibit knowledge that is more aligned with the given character and reduce hallucinated responses.

1 Introduction

The advent of Large Language Models (LLMs) has significantly enhanced the capabilities of conversational AI agents due to their proficiency in understanding and generation. To further promote user engagement and entrainment (Park et al., 2023), role-playing LLMs are designed to mimic the traits and experiences of specific characters, producing interactions that are role-consistent, emotionally deep, and contextually aware.

To improve imitation capabilities, recent studies (Shao et al., 2023; Tu et al., 2024; Tao et al., 2024; Lu et al., 2024; Zhou et al., 2024) have fine-tuned LLMs on datasets specifically curated for role-playing scenarios. However, due to the labor-intensive nature of data collection and the high computational costs associated with fine-tuning, an alternative line of research explores the use of in-context learning by, for example, providing few-shot examples (Li et al., 2023) or using static user

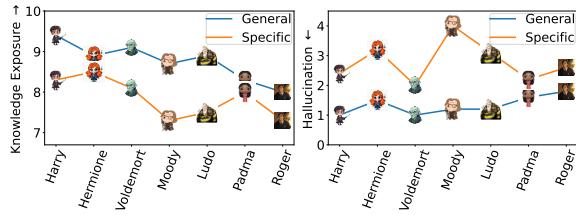


Figure 1: This figure illustrates that LLMs perform worse on role-specific questions, particularly when imitating lower-frequency characters.

profiles (Wang et al., 2024a), to provide role-related background information.

However, LLMs still struggle to align accurately with a character’s knowledge, often generating responses that lack appropriate traits or include fabricated content—particularly in certain role-playing scenarios where factual consistency is critical. As shown in Figure 1, we tasked GPT-4o-mini with playing seven characters from the Harry Potter series, selected based on their frequency of appearance. Each character was presented with 10 general questions (e.g., interests, attitudes) and 10 role-specific questions (e.g., experiences, activities). We then recruited human raters to assess whether the language models accurately reflected each character’s traits and to rate the severity of hallucinations, using a ten-point scale. From these results, we observe that LLMs perform worse on role-specific questions that require detailed character knowledge, particularly for less frequent characters.

Our failure case analysis reveals two key factors contributing to these issues. (1) *Entity ambiguity during knowledge extraction*: a single character may be referred to by different names across different stages or contexts. For example, ‘Anakin Skywalker’ is also known as ‘Darth Vader’ or ‘Lord Vader’ in various installments of the Star Wars series. If not properly unified, such name variation can cause critical information to be missed during knowledge retrieval ; (2) *Character-related cognition-boundary unawareness*: LLMs encode

vast amounts of knowledge beyond the scope of the character they are portraying and often rely on LLM internal knowledge when responding to user queries. This can result in fabricated responses, particularly when the question falls outside the character’s original knowledge boundaries. Such hallucinations are unique to the role-playing setting.

To address these issues, we introduce RoleRAG, a retrieval-based framework specifically designed for role-playing tasks. Our approach is built on knowledge graph-enhanced retrieval, motivated by the observation that answering a single question may require reasoning over a broad range of dispersed textual knowledge. In the context of role-playing, the knowledge graph is constructed from character-centric corpora such as Wikipedia profiles and books. Each node in the graph represents an entity (e.g., character, location), and each edge encodes a semantic relationship between two entities (e.g., interactions between characters). To normalize duplicated names referring to the same entity, we propose an efficient semantic entity normalization algorithm. It first links name variants based on their local context, then clusters them into groups representing the same entity. Finally, an LLM is prompted to generate a unified canonical name for each group. The knowledge graph is then constructed using these normalized entities.

Our retrieval module, built on this graph-based indexing system, is designed to extract both specific and general entities mentioned in user queries while rejecting those that fall outside the scope of the character’s knowledge. Subsequently, information relevant to the designated role is retrieved from the knowledge graph and provided to the LLM, equipping it with detailed contextual information to generate accurate responses. In contrast, out-of-scope questions are encouraged to be rejected to prevent the model from generating hallucinated or fabricated content.

Our contributions can be summarized as follows:

- We propose an efficient entity normalization algorithm that merges duplicated names referring to the same entity, thereby facilitating high-quality graph-based indexing over large character corpora.
- We introduce an effective retrieval module that not only retrieves both general concepts and character-specific details, but also helps reject out-of-scope questions to reduce hallucinations.
- Extensive experiments that demonstrate that Rol-eRAG outperforms relevant baselines by exhibit-

ing aligned character knowledge and reducing hallucinations.

2 Related Works

LLM-based Role-Play enables LLMs to embody user-specified characters, enhancing engagement through conversation. Existing research falls into three main directions: (1) Fine-tuning-based approaches (Shao et al., 2023; Tu et al., 2024; Wang et al., 2024a; Zhou et al., 2024; Lu et al., 2024) involve training open-source LLMs on curated character corpora. The training data is either synthetic—generated specifically for character conditioning (Shao et al., 2023; Tu et al., 2024; Wang et al., 2024a)—or extracted from real-world datasets using LLMs (Zhou et al., 2024). (2) Retrieval-based approaches (Salemi et al., 2024; Weir et al., 2024; Zhou et al., 2024) fetch relevant documents from a character corpus to serve as contextual input to the LLM, thereby enhancing its ability to generate accurate and character-specific responses. The performance of these methods heavily depends on the quality and relevance of the retrieved content. (3) Plugin-based methods (Liu et al., 2024) freeze the LLM while encoding each user’s characteristics using a lightweight plugin model. The resulting user embedding is then concatenated with the embedding of the user’s query to guide the LLM in generating personalized responses. A comprehensive comparison of the three categories is provided in the Appendix A. In this work, we follow retrieval-based approaches, aiming to provide character-relevant content while reducing hallucinations in responses.

Persona-based Dialogue. Persona-based dialogue tasks require LLMs to exhibit general human-like traits such as humor, empathy, or curiosity, rather than adhering to specific role characteristics. Unlike role-playing, the focus is on consistent personality expression. Personas can be assigned via Big Five trait prompts (Jiang et al., 2023), character profiles (Tu et al., 2024; Zhou et al., 2024), or dialogue history (Zhong et al., 2022). Evaluation is typically conducted through personality assessments or interviews (Wang et al., 2024b). A comprehensive comparison between role-playing and persona-based dialogue refer to (Tseng et al., 2024). Our work focuses on enabling role-playing LLMs to produce character-faithful responses.

Retrieval-Augmented Generation (RAG) enhances LLMs by retrieving external knowledge

to support informed, accurate, and contextually grounded responses (Lewis et al., 2020; Liu et al., 2022; Zhuang et al., 2023; Li et al., 2024). Standard RAG struggles with capturing complex inter-entity relationships across multiple chunks (Guo et al., 2024) and often fails on general queries requiring comprehensive understanding of large knowledge bases (Edge et al., 2024). To address these limitations, recent work (Edge et al., 2024; Sarmah et al., 2024; Wu et al., 2024; Guo et al., 2024) leverages LLMs to construct knowledge graphs (KGs), where nodes represent entity attributes and edges encode inter-entity relationships.

However, knowledge graph-enhanced methods overlook the entity ambiguity issue, where multiple names refer to the same character, and their retrieval process typically ignores the character’s knowledge boundary, leading to responses that go beyond the intended role and produce out-of-character content.

3 RoleRAG

Our overall framework for RoleRAG is illustrated in Figure 2 and consists of two novel modules specifically designed for the role-playing task: (1) an entity normalization module that removes semantically duplicated entities, and (2) a retrieval module that fetches question-relevant information while rejecting out-of-scope queries.

3.1 Entity and Relation Extraction

In role-playing, character corpora often originate from novels, TV series, or biographies, with descriptions that exceed LLM token limits. Following prior work (Edge et al., 2024; Wu et al., 2024; Guo et al., 2024), we split descriptions into chunks $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n\}$, process them independently, and aggregate the results into a unified character profile.

For each chunk \mathcal{D}_i , we employ LLMs to meticulously extract entities, adhering to a predefined data structure: $\{\text{name}, \text{type}, \text{description}\}$, denoted by \mathbf{n}_i . Furthermore, we prompt LLMs to identify structural relations between two entities, specifically, $\{\text{source}, \text{target}, \text{description}, \text{strength}\}$, denoted by \mathbf{r}_i , where *description* and *strength* denote the textual relationship and its intensity between the source and target nodes, respectively. After all chunks are processed, all entities and relations are stored in global databases \mathcal{N} and \mathcal{R} .

To enable semantic retrieval, each entity \mathbf{n}_i is encoded into a high-dimensional vector \mathbf{v}_i using a

Algorithm 1 Entity Normalization Algorithm

Require: Entity Database \mathcal{N} .

Ensure: a unified name for each name group.

```

1: Initialize empty entity graph  $\mathcal{G}$ .
2: Initialize empty vector database  $\mathcal{V}$ .
3: for  $\mathbf{n}_i \in \mathcal{N}$  do
4:   if  $\mathbf{n}_i \in \mathcal{V}$  then
5:     continue;                                 $\triangleright$  node exists
6:   else
7:      $\mathcal{N}_k = f_k(\mathbf{n}_i, \mathcal{V})$ 
8:     Insert  $\mathbf{n}_i$  to  $\mathcal{V}$ 
9:     Insert  $\mathbf{n}_i$  to  $\mathcal{G}$ 
10:    end if
11:    for  $\mathbf{n}_j \in \mathcal{N}_k$  do
12:      if  $\mathbf{n}_i == \mathbf{n}_j$  then       $\triangleright$  LLM prompt
13:        Insert  $\mathbf{n}_j$  to  $\mathcal{G}$ 
14:        Connect  $\mathbf{n}_i$  and  $\mathbf{n}_j$  in  $\mathcal{G}$ 
15:      else
16:        continue
17:      end if
18:    end for
19:  end for
20: Count the number of connected components in
21:  $\mathcal{G}$ 
22: for each connected components  $G$  in  $\mathcal{G}$  do
23:   Select the unified name in  $G$            $\triangleright$  LLM
      prompt
23: end for

```

text embedding model applied to both its name and description. The resulting pairs $\mathbf{n}_i, \mathbf{v}_i$ are stored in a vector database \mathcal{V} for efficient similarity-based retrieval. We define the retrieval interface as $f_k(\mathcal{V}, \mathbf{n})$, which returns the top k entities most similar to a query entity.

3.2 Entity Normalization

To mitigate entity ambiguity, we introduce a semantic entity normalization procedure, detailed in Algorithm 1. Given all extracted entities, our algorithm iterates through each entity, retrieving the k most semantically similar entities from the entity vector database. Next, we present these entity pairs, along with their names and descriptions, to the LLM, prompting it to determine if they refer to the same character. If the LLM identifies two entities as the same individual, we connect their corresponding nodes with an edge in the entity graph \mathcal{G} . After processing all entities, we partition the entity graph into multiple connected subgraphs, each representing a distinct individual, as illustrated in

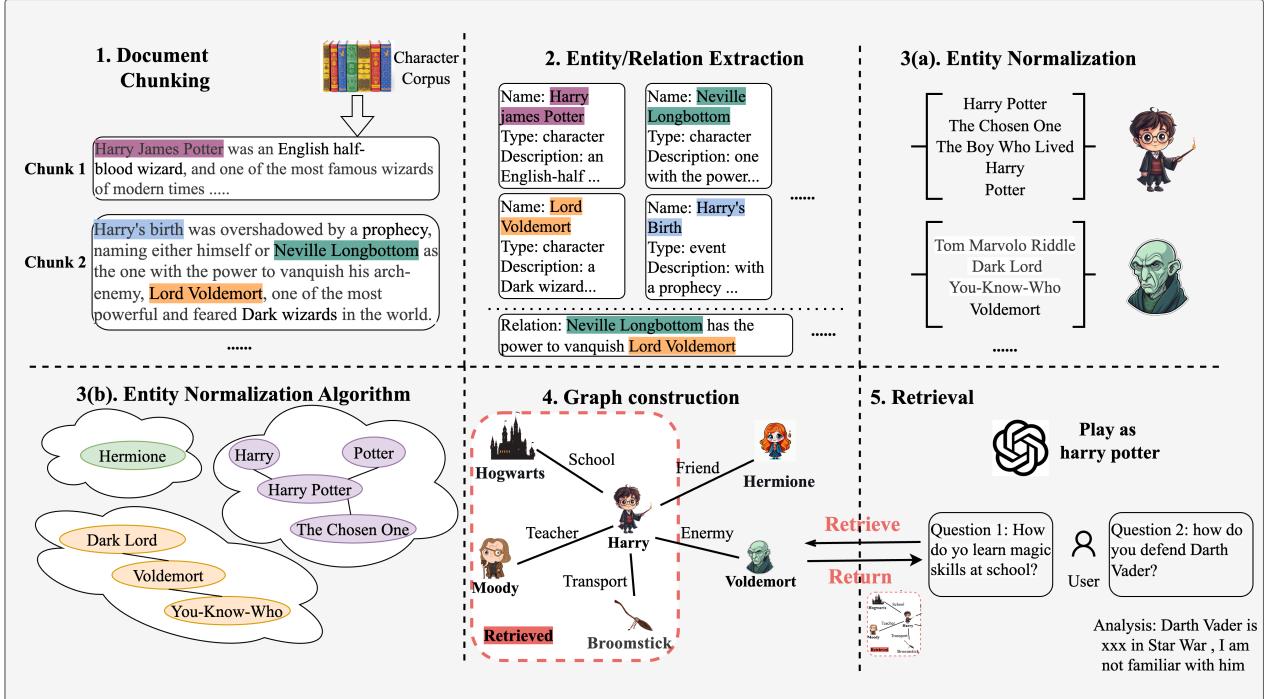


Figure 2: Workflow of our proposed RoleRAG.

Figure 2. Finally, we prompt the LLM once more to generate a unified canonical name for each connected subgraph.

Compared to brute-force LLM-based pairwise comparisons, our method reduces LLM calls by a factor of $|\mathcal{N}|/k$, where $|\mathcal{N}|$ is the total number of entities and k the number of entities retrieved from the vector database. We further leverage modern vector embedding techniques to accelerate retrieval by reducing the number of semantic dissimilar entities.

3.3 Graph Construction

After identifying entity groups referring to the same character and assigning each group a unified canonical name, we construct a mapping table linking source names to their canonical forms. Subsequently, we normalize all raw names across both entity and relationship databases to facilitate effective retrieval. Since normalization reduces duplicate entities and relationships in \mathcal{N} and \mathcal{R} , we summarize their descriptions using LLMs to preserve contextual details.

Finally, we formally construct the knowledge graph from character database as follows,

$$\hat{\mathcal{G}} = \{\hat{\mathcal{N}}, \hat{\mathcal{R}}\} \quad (1)$$

where $\hat{\mathcal{N}}, \hat{\mathcal{R}}$ denote nodes and relationships after de-duplication.

3.4 Retrieval Module for Role-playing

Given a user query, we first prompt an LLM to infer hypothetical contexts relevant to the desired response, inspired by HyDE (Gao et al., 2023). Subsequently, we prompt the LLM with character profiles summarized from our knowledge graph to identify entities appearing in both the original query and the inferred hypothetical context. For each entity, the LLM returns its *name*, *entity type*, *relevance to the designated character* (along with the underlying rationale), and *specificity level* (either specific or general). Leveraging this information, we develop three distinct retrieval strategies to gather contextually appropriate content from the knowledge graph, supplementing the character summary provided to the LLM:

- For entities identified as outside the character’s knowledge scope (e.g., querying an ancient figure about Apollo 11), we explicitly inform the LLM of their irrelevance along with the underlying rationale, thereby discouraging the LLM from providing hallucinatory responses.
- For specific entities, we first retrieve the top semantically similar entities from the vector database \mathcal{V} based on the entity embeddings. Subsequently, we extract detailed descriptions of these entities and their relationships with the designated character from the knowledge graph to

form the context.

- For general entities (e.g., interests, hobbies), we retrieve entities from the 1-hop neighborhood of the target character, filtering out irrelevant entities based on their types. Descriptions of the remaining entities are then used to provide contextual details for response generation.

Our retrieval strategy not only enriches character-related responses with detailed knowledge but also rejects out-of-scope questions that exceed the character’s cognitive boundaries, thereby enhancing knowledge exposure and reducing hallucinations in role-playing.

4 Experimental Setup

4.1 Baselines

We compare RoleRAG against the following set of baselines: **Vanilla**, it prompts an LLM to role-play as a character with task description; **RAG** (Lewis et al., 2020) retrieves chunks most semantically similar to a user query and provides them as context for LLM-based response generation; **Character profile** (Zhou et al., 2024), which provides the LLM with a profile of the character that the LLM is portraying; **GraphRAG** (Edge et al., 2024) retrieves relevant information from an indexed entity-relation knowledge graph.

We collect source materials from Wikipedia, Baidu Baike, and novels to construct the retrieval databases for both RAG and RoleRAG. For character profiles, we prompt GPT-4 to summarize the corresponding Wikipedia or Baidu Baike biography into a short paragraph, which is prepended to user queries to provide background context.

4.2 Evaluation Metrics

Role-play LLMs should consistently embody the target role, provide accurate responses, maintain character integrity, and avoid factual errors. Following existing studies (Tu et al., 2024; Lu et al., 2024), we perform our evaluation with the following metrics in Figure 3.

Knowledge Exposure measures the extent to which personalized traits—such as background, behavior, knowledge, and experiences—are accurately recalled from the character profile. *Knowledge Hallucination* evaluates the precision of responses, focusing on the model’s ability to avoid generating incorrect, misleading, or out-of-scope information. This is essential for maintaining the credibility and consistency of the LLM within the

designed role. *Unknown Questions Rejection* measures the model’s self-awareness in role-playing by assessing its ability to recognize and communicate the boundaries of the character’s knowledge.

To quantitatively evaluate these metrics, we follow prior work (Shao et al., 2023; Dai et al., 2024; Lu et al., 2024; Wang et al., 2024a) and employ GPT-4o as a judge (Zheng et al., 2023) to rate the responses. We prompt GPT-4o to rate knowledge exposure and hallucination on a 1–10 scale. A higher knowledge exposure score indicates that the LLM demonstrates deep understanding of the character, while a lower hallucination score reflects responses free from misinformation about the character’s background. For self-awareness measurement, we prompt the LLM to assign a score of 1 if the response adheres to the character’s cognitive scope, and 0 otherwise. Since judge LLMs may exhibit biases during evaluation—such as the “self-enhancement bias” (Zheng et al., 2023)—we include human evaluators in the loop to verify and correct the scores produced by the judge LLM. The detailed evaluation process is described in Appendix section C.

4.3 Datasets

To evaluate performance of our RoleRAG framework, we conducted experiments on three role-playing datasets: (1) **Harry Potter Dataset**, collected by us, this dataset contains seven characters from the Harry Potter series. Each character is presented with 20 role-specific questions (10 general questions about their interests and values, as well as 10 detailed questions about their experiences and relationships with others). (2) **RoleBench-zh**, a subset of the RoleBench evaluation, this dataset includes five historical and fictional Chinese characters. This dataset contains both role-related and out-of-scope questions, 357 in total. For example, it includes a question about Apollo 11 directed at an ancient figure. (3) **Character-LLM** (Shao et al., 2023), contains 859 questions, including role-related and out-of-scope questions. The statistics of the three datasets are provided in Appendix B.

Our experiments are conducted on relatively small datasets featuring well-known characters or those from famous novels to ensure that details can be easily verified by human evaluators.

4.4 Implementation Details

In RoleRAG, we split the character profile into chunks of 600 tokens with an overlap of 100 to-

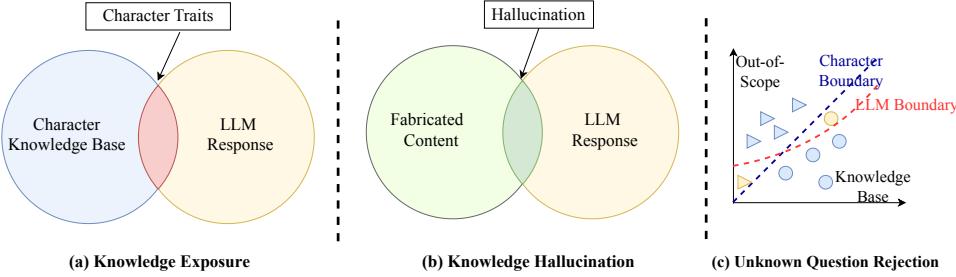


Figure 3: Illustration of evaluation metrics. We encourage LLMs to exhibit more personal traits, minimize fabricated content, and align more closely with the boundaries of character cognition.

kens. GPT-4o mini is used as the LLM to extract entities and their relationships, perform entity normalization, and merge descriptions of duplicate entities. We use OpenAI’s “text-embedding-3-large model” to encode entity descriptions into vector representations with an embedding dimension of 3,072. Cosine distance is used to measure the similarity between entities.

To assess RoleRAG’s usability, we perform experiments with various LLMs, including open-source LLMs (including Mistral-Small 22b ([Mistral, 2025](#)), Llama3.1 8b, Llama3.3 70b ([Dubey et al., 2024](#)), Qwen2.5 14b ([Yang et al., 2024](#))), proprietary LLMs (OpenAI GPT series ([OpenAI, 2024](#))), and LLMs specifically tailored for role-playing tasks (Doubao Pro 32k¹).

5 Experimental Results

5.1 Main Results

Our main results are shown in Table 1. Overall, the results show that RoleRAG performs better than the baseline methods. In many instances, a smaller LLM with RoleRAG, e.g., Qwen 2.5 (14b), can outperform larger LLMs, e.g., Llama 3.3 (70b), without it, demonstrating the effectiveness of RoleRAG. While adding character background improves knowledge exposure and reduces hallucination compared to vanilla approaches, RoleRAG outperforms other retrieval-based baselines by structuring information for efficient access to character details and relationships, enabling more accurate role-playing. For unknown questions, RoleRAG outperforms baseline methods, even when those are explicitly instructed not to answer out-of-scope queries. We attribute this to RoleRAG’s relevance analysis during retrieval, along with rationale generation, which helps prevent implausible responses—such as asking Harry Potter about events in Star Wars.

¹<https://www.volcengine.com/product/doubao>

Fine-tuning LLMs for role-playing can improve performance, as shown by Doubao Pro on the RoleBench-zh dataset. However, the vast number of characters makes it impractical to fine-tune models for all possible roles. Additionally, defining and enforcing cognitive boundaries during fine-tuning remains a challenging, unsolved problem. These limitations are evident in Doubao Pro’s weaker performance on the Harry Potter and CharacterLLM datasets, along with its lower self-awareness. In contrast, RoleRAG enables both general-purpose and fine-tuned LLMs to access character-specific knowledge effectively.

The results in Table 1 appear only marginally improved due to the judge LLM’s tendency to assign high knowledge exposure scores and low hallucination scores when responses lack major errors. For example, scores of 8–9 are often given for generally appropriate answers, while human evaluators tend to adjust scores only in cases of significant faults rather than making fine-grained changes. As a result, the high baseline scores from LLM judges leave limited room for observable improvement.

5.2 Ablation Studies

Note that our method is built upon knowledge graph enhanced retrieval. Different from GraphRAG, we introduce the entity normalization to merge duplicated entities during graph construction and a retrieval strategy for role-playing. In this ablation study, we disable entity normalization and adopt the local search that starts from the most similar nodes from query embedding, and expanding through its neighborhood and community in GraphRAG. The experiment results are illustrated in Table 2, we can see that: 1) the most significant improvement comes from the combination of RoleRAG and the novel retrieval strategy; 2) the retrieval method could clearly enhance the boundary awareness by providing relevance to the character.

Table 1: Our main experimental results on the Harry Potter, RoleBench-zh, and CharacterLLM datasets. The reported scores are the average across all questions in each dataset, and \uparrow / \downarrow means higher/lower results are better. Human evaluators are recruited to verify and correct GPT-4o’s score.

Model	Method	Harry Potter			RoleBench-zh			CharacterLLM \ddagger		
		KE \uparrow	KH \downarrow	UQR \uparrow	KE \uparrow	KH \downarrow	UQR \uparrow	KE \uparrow	KH \downarrow	UQR \uparrow
Open-source General Models										
Mistral-Small (22b)	Vanilla	7.457	2.229	—	4.398	5.731	0.510	8.535	1.794	0.894
	RAG	7.786	2.486	—	4.905	5.367	0.580	8.871	1.538	0.929
	User profile	7.650	2.293	—	5.182	3.890	0.711	8.861	1.570	0.932
	GraphRAG	7.356	2.488	—	5.328	4.459	0.613	8.963	1.572	0.925
	RoleRAG	7.550	2.150	—	5.585	3.961	0.678	9.057	1.404	0.959
Llama 3.1 (8b)	Vanilla	7.579	2.200	—	4.115	6.232	0.462	7.932	2.613	0.819
	RAG	7.486	3.214	—	4.728	5.389	0.600	8.505	2.084	0.884
	User profile	7.057	3.657	—	5.047	4.843	0.569	8.292	2.174	0.875
	GraphRAG	7.373	2.833	—	5.479	4.367	0.678	8.543	2.019	0.900
	RoleRAG	7.750	2.352	—	5.608	4.126	0.661	8.653	1.961	0.908
Qwen 2.5 (14b)	Vanilla	7.614	2.129	—	6.238	3.352	0.734	8.709	1.656	0.907
	RAG	7.707	2.371	—	6.583	3.020	0.773	9.067	1.356	0.959
	User profile	7.764	2.693	—	6.605	3.020	0.818	9.039	1.382	0.953
	GraphRAG	7.762	2.433	—	6.686	2.888	0.790	9.230	1.321	0.956
	RoleRAG	7.986	2.071	—	6.798	2.538	0.832	9.238	1.231	0.974
Llama3.3 (70b)	Vanilla	7.414	2.279	—	6.034	3.709	0.689	8.811	1.419	0.929
	RAG	8.243	2.071	—	6.031	3.546	0.751	9.198	1.352	0.962
	User profile	8.021	2.050	—	6.457	3.014	0.754	9.258	1.272	0.964
	GraphRAG	8.352	2.070	—	6.092	3.521	0.714	9.302	1.275	0.967
	RoleRAG	8.564	1.743	—	6.723	2.622	0.837	9.270	1.265	0.974
Close-source General Model										
GPT-4o-mini	Vanilla	7.643	2.121	—	5.863	4.202	0.714	8.789	1.492	0.925
	RAG	8.493	1.750	—	5.986	3.930	0.709	8.996	1.311	0.954
	User profile	8.221	2.021	—	6.232	3.754	0.733	9.009	1.317	0.945
	GraphRAG	8.729	1.776	—	6.445	3.429	0.717	9.136	1.308	0.958
	RoleRAG	8.821	1.571	—	6.994	2.697	0.857	9.138	1.211	0.978
Close-source Role-playing Model										
Douba Pro 32K	Vanilla	7.193	2.257	—	6.840	3.745	0.860	8.522	1.639	0.891
	RAG	8.179	1.814	—	7.170	2.246	0.880	8.836	1.379	0.939
	User profile	7.450	2.179	—	7.207	2.429	0.905	8.927	1.351	0.932
	GraphRAG	8.040	1.780	—	6.866	2.087	0.902	8.929	1.361	0.932
	RoleRAG	8.221	1.564	—	7.733	1.689	0.952	8.970	1.313	0.956

KE: Know exposure [0, 10], KH: Knowledge hallucination [0, 10], UQR: Unknown question rejection {0, 1}.

\ddagger Human evaluation takes extremely longer on this dataset, we average scores from two trials of GPT4o.

5.3 RoleRAG for General Questions

Table 3 presents knowledge exposure and hallucination scores for general questions in the Harry Potter dataset. While LLMs show low hallucination, they reveal few character-specific traits. We hypothesize that LLMs have internalized general knowledge from large-scale pretraining but lack role-specific details. In our RoleRAG, we retrieve 1-hop neighbors of the character matching the type of general keywords, enriching the response with relevant context and significantly improving knowledge exposure while keeping low hallucination.

5.4 RoleRAG for Specific Questions

Table 4 demonstrates knowledge exposure and hallucination scores for specific questions from the

Harry Potter dataset. Compared with responses to general questions, when asked about details, LLMs tend to fabricate stories or are reluctant to provide specific information. With our RoleRAG, we observe a clear improvement in knowledge exposure and hallucination scores after retrieving detailed entity information mentioned in user questions from the knowledge base. We also observe an interesting phenomenon: smaller LLMs tend not to incorporate the retrieved knowledge into their responses as effectively as larger LLMs.

5.5 RoleRAG for Minority Groups

Table 5 reports performance across characters in the Harry Potter series, sorted by their frequency of appearance. The results demonstrate that for pop-

Table 2: Ablation studies on RoleBench-zh datasets.

Entity Normalization	Retrieval	KE	KH	UQR
Without	Local search	6.006	4.126	0.745
With	Local search	6.431	3.409	0.770
Without	Our retrieval	6.154	3.454	0.762
With	Our retrieval	6.994	2.697	0.857

Table 3: Performance of RoleRAG on general questions on Harry Potter dataset.

Model	KE		KH	
	Vanilla	RoleRAG	Vanilla	RoleRAG
Mistral-Small (22b)	7.486	7.685	1.457	1.485
Llama3.1 (8b)	7.714	8.342	1.343	1.614
Qwen 2.5 (14b)	7.614	8.157	1.414	1.371
Llama 3.3 (70b)	7.414	8.814	1.557	1.086
GPT-4o mini	7.671	8.957	1.371	1.157
Douba Pro 32K	7.300	8.414	1.586	1.057

Table 4: Performance of RoleRAG on specific questions on Harry Potter dataset.

Model	KE		KH	
	Vanilla	RoleRAG	Vanilla	RoleRAG
Mistral-Small (22b)	6.587	7.414	2.6	2.814
Llama3.1 (8b)	6.842	7.157	3.058	3.070
Qwen 2.5 (14b)	7.425	7.902	2.842	2.771
Llama 3.3 (70b)	7.213	8.314	3.000	2.400
GPT-4o mini	7.314	8.686	2.871	1.986
Douba Pro 32K	7.085	8.029	2.929	2.071

Table 5: Performance of RoleRAG across characters with varying frequencies in the Harry Potter series, listed from highest to lowest frequency.

Model	KE		KH	
	Vanilla	RoleRAG	Vanilla	RoleRAG
Harry Potter	7.77	8.11 _{+0.34}	1.69	1.97 _{+0.28}
Hermione Granger	7.57	8.23 _{+0.66}	2.58	2.28 _{-0.3}
Voldemort	7.99	8.37 _{+0.38}	1.85	1.98 _{+0.13}
Alastor Moody	7.47	7.83 _{+0.36}	2.77	2.63 _{-0.14}
Ludovic Bagman	7.08	8.18 _{+1.1}	2.46	1.68 _{-0.78}
Padma Patil	7.14	8.4 _{+1.26}	2.21	1.34 _{-0.87}
Roger Davies	7.24	7.94 _{+0.7}	2.08	1.83 _{-0.25}

ular characters like ‘Harry Potter’, LLMs exhibit higher knowledge exposure and lower hallucination rates. Conversely, less commonly mentioned characters tend to show reduced knowledge accuracy and increased instances of fabricated content. These results show that with the aid of RoleRAG, characters that appear less frequently, such as ‘Ludovic Bagman’ and ‘Padma Patil’, benefit significantly in terms of enhanced knowledge exposure and reduced fabrication of content.

5.6 RoleRAG for Out-of-scope Questions

Figure 4 shows that when role-playing, LLMs tend to answer all questions—even those beyond the

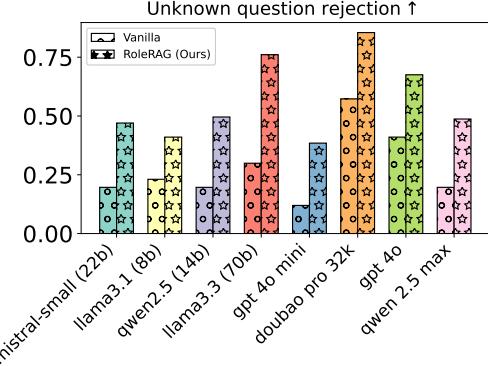


Figure 4: Experiments of out-of-scope questions in RoleBench-zh dataset.

character’s knowledge scope. This suggests that LLMs often fail to fully adopt the perspective of the target character, instead relying on their internalized knowledge—an issue observed even in larger models like GPT-4o and Qwen2.5-Max. While the strong performance of Douba Pro shows that fine-tuning can improve awareness of a character’s cognitive boundary, it lacks adaptability to new characters without task-specific data. Overall, regardless of model size or fine-tuning, the results demonstrate that RoleRAG equips LLMs with the information needed to correctly reject out-of-scope questions, better aligning their cognitive boundaries with the intended character.

6 Conclusion

When tasked with role-playing, LLMs often generate responses that lack depth in character knowledge and introduce information outside the character’s known universe—a role-specific form of hallucination. To address these issues, in this paper, we introduce RoleRAG, a novel framework for role-playing that merges duplicated entities and enhances the retrieval of relevant information. Additionally, our retrieval module assesses entity relevance to the target character, enabling accurate content generation while effectively rejecting unrelated questions. Through rigorous experimentation, we demonstrated that RoleRAG consistently outperforms relevant baselines. The success of RoleRAG highlights its potential as a powerful tool for improving the reliability and authenticity of role-playing models, paving the way for more sophisticated, context-aware conversational agents in a variety of applications.

7 Limitations

A minor concern in our work is the evaluation of the responses generated by LLMs. It is difficult to recruit human evaluators who have deep knowledge about the characters and stories used in our evaluations. Even if evaluators are familiar with the characters and stories, they may need more detailed information to accurately judge whether a generated response is sensible and does not contain hallucination. Therefore, we use LLMs as evaluators in our experiments, then verified by human annotators. However, we observed that LLMs tend to assign over-confident scores, which can mislead human evaluators and render the scores insufficiently discriminative in our experiments.

A possible direction to explore is how to prompt an LLM to recognize and understand the limits of character knowledge when engaged in role-play. Given that LLMs are trained on massive, diverse datasets, they often possess knowledge far beyond what the characters they are asked to portray would realistically know. As a result, managing these knowledge boundaries becomes crucial to ensuring more authentic role-playing. Defining the scope and limits of a character’s knowledge is not only necessary to prevent the model from introducing irrelevant or inaccurate information, but it also directly improves the accuracy of knowledge exposure within the context of the character. Ultimately, addressing this challenge could significantly enhance the believability and effectiveness of LLMs in role-playing scenarios, fostering more realistic and emerging interactions.

Another limitation of our work is that we focused on single-turn conversations. Multi-turn conversations present unique challenges, including maintaining consistency across turns, ensuring that the LLM remains in-character, and effectively managing the dialogue history. As multi-turn conversations often require the model to recall and build upon previous interactions, there is an increased risk of the model deviating from the character’s personality or losing track of essential details. In the future, we plan to investigate how to address these challenges.

In retrieval-based methods, the quality of the response generated by an LLM depends on the model’s ability to utilize the information retrieved. However, it is not fully understood how LLMs incorporate this retrieved knowledge into their responses. We have observed numerous instances

where LLMs contradict the retrieved information. Thus, gaining a deeper understanding of the internal mechanisms of in-context learning is crucial to improving retrieval-based approaches.

8 Ethics

We will release our code base publicly as part of our commitment to the open source initiative. However, it is important to recognize that role-playing with these tools can lead to jailbreaking, and misuse may result in the generation of biased or harmful content, including incitement to hatred or the creation of divisive scenarios. We truly hope that this work will be used strictly for research purposes.

With our proposed RoleRAG, we aim to effectively integrate role-specific knowledge and memory into LLMs. However, we must acknowledge that we cannot fully control how LLMs utilize this knowledge in dialogue generation, which could still result in harmful or malicious responses. In the future, we plan to investigate the mechanisms of prompting to more deliberately control response generation. Additionally, it is crucial to scrutinize responses in high-stakes and sensitive scenarios to ensure safety and appropriateness.

References

- Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. [Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents](#). *arXiv preprint arXiv:2408.04203*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. [From Local to Global: A Graph RAG Approach to Query-Focused Summarization](#). pages 1–15.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. [Precise zero-shot dense retrieval without relevance labels](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.
- Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. [Lightrag: Simple and fast retrieval-augmented generation](#). *arXiv preprint arXiv:2410.05779*.

- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kütller, 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. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.
- Zhuohang Li, Jiaxin Zhang, Chao Yan, Kamalika Das, Sricharan Kumar, Murat Kantacioglu, and Bradley A. Malin. 2024. Do you know what you are talking about? characterizing query-knowledge relevance for reliable retrieval augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6130–6151, Miami, Florida, USA. Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2024. LLMs + Persona-Plug = Personalized LLMs.
- Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.
- Mistral. 2025. Mistral small 3.
- OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.
- 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.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.
- Bhaskarjit Sarmah, Dhagash Mehta, Benika Hall, Rohan Rao, Sunil Patel, and Stefano Pasquali. 2024. Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance, ICAIF ’24*, page 608–616, New York, NY, USA. Association for Computing Machinery.
- 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, Singapore. Association for Computational Linguistics.
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.
- Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie. 2024. RoleCraft-GLM: Advancing personalized role-playing in large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 1–9, St. Julians, Malta. Association for Computational Linguistics.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.
- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhuan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*,

pages 14743–14777, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. **InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

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

Nathaniel Weir, Ryan Thomas, Randolph d’Amore, Kellie Hill, Benjamin Van Durme, and Harsh Jhamtani. 2024. **Ontologically faithful generation of non-player character dialogues**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9212–9242, Miami, Florida, USA. Association for Computational Linguistics.

Junde Wu, Jiayuan Zhu, and Yunli Qi. 2024. **Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation**. *arXiv preprint arXiv:2408.04187*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. **Judging LLM-as-a-judge with MT-bench and chatbot arena**. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. **Less is more: Learning to refine dialogue history for personalized dialogue generation**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guan-qun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. **CharacterGLM: Customizing social characters with large language models**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry*

Track, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

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 Comparison of LLM-based Role-playing approaches

Table 6 shows a comparison of different methods used for using LLMs in role-playing tasks. Fine-tuning-based approaches require extensive data collection and are computationally expensive, and they often fail to generalize to roles beyond the training corpus, as each character has a distinct knowledge. Moreover, LLMs inherently encode vast general knowledge, which they may draw upon when answering queries—often leading to fabricated or out-of-scope content. Defining clear character boundaries remains a challenge for fine-tuning-based approaches. Retrieval-based methods eliminate the need for model training and costly data labeling. However, their effectiveness depends on efficiently retrieving query-relevant context from a large character knowledge base through a robust indexing system.

B Dataset Statistics

The statistics of our experimental datasets are illustrated in Table 7. In our experiment, recruiting evaluators who can recall the complete knowledge base of a specific character is challenging, and web searches are often required during evaluation. For instance, assessing a batch of 357 response in the RoleBench-Zh dataset takes approximately **three hours** per evaluation session; The cost of evaluating LLM generation of CharacterLLM dataset with GPT-4 is approximately 5 US dollars.

Table 7: Statistics of the experimental datasets.

Datasets	#Roles	In Scope	Out of Scope
Harry Potter	7	140	-
RoleBench-Zh	5	240	117
Character-LLM	9	814	45

C Evaluation Process

To judge the generated responses according to the above metrics, we make use of GPT-4o to act as a judge LLM by rating the responses. Powerful LLMs such as GPT-4 have been widely employed as evaluators in recent studies (Shao et al., 2023; Dai et al., 2024; Lu et al., 2024; Wang et al., 2024a) where GPT-4 is prompted to give scores for generated output on a defined scale, or to compare

responses and select which one is better. However, there are some concerns about the reliability of LLMs to rate generated responses. Therefore, based on recent works that explore the use of LLMs as judges, we adopt a few measures to increase the reliability of the scores in our experiments. First, we prompt the LLM to generate an analysis before it scores the response. This approach follows recent research (Shen et al., 2023; Zheng et al., 2023) and is based on the success of Chain-of-Thought prompting (Wei et al., 2022). Following Ditto (Lu et al., 2024), we set the temperature of GPT-4o to 0.2 to penalize creativity during evaluation.

To avoid biases that judge LLMs may have, such as the “self-enhancement bias” (Zheng et al., 2023), we include humans in the evaluation process to verify the scores produced by the judge LLM. The human evaluator can use the analysis produced by the judge LLM, as well as any other information sources they want to use, to determine whether the score is sensible. The human evaluator can adjust the score if they feel that it is not correct. We use three different prompts to generate scores for each metric, which can be found in Appendix E.



Figure 5: Word cloud for responses generated by GPT-4o mini when role-playing as Harry Potter.



Figure 6: Word cloud for responses generated by GPT-4o mini when role-playing as Voldemort.

Table 6: Comparison of different LLM role-playing approaches.

Methods	Fine-tuning Based	Retrieval Based	Plugin Model
LLM Training	YES	No	YES
Character Data Labeling	YES	No	YES
Computational Burden	High	Low	Moderate
Character Data Organization	All characters shared	One character, one corpus	One character, one plugin
Adaptation to Unseen Roles	Hard	Easy	Hard
Modifying LLMs’ Knowledge	Hard	Easy	Hard

D Additional Experiments

D.1 Word Clouds

In this experiment, we collect all responses generated by GPT-4o-mini when role-playing characters from the Harry Potter series and visualize them using word clouds to highlight frequently used terms. Larger words in the figures indicate higher frequency. As shown in Figure 5, when playing as Harry Potter, the model emphasizes terms such as friends, Hogwarts, Hermione, Voldemort, and brave. In contrast, Figure 6 shows that when role-playing Voldemort, the most frequent terms include world, power, death, and dark.

D.2 Demonstration of RoleRAG retrieval

Figure 7 illustrates the types of information retrieved by RoleRAG in response to an interview question posed to an LLM role-playing Ludwig van Beethoven. The question asks about Beethoven’s relationship with Haydn and Mozart. RoleRAG first identifies the entities in the query, assesses their relevance to Beethoven, and determines their specificity.

Since all three are specific entities related to Beethoven, we locate them in the knowledge graph using semantic matching and the name mapping table. We then retrieve detailed information about each entity and their relationships with Beethoven. Finally, the retrieved context is aggregated and provided to the LLM to answer the user’s question.

E Prompts in our experiments

This section contains the prompts used in our experiments. Figures 8, 9, 10, and 11 show the prompts used for generation and scoring. Green text in curly braces represent text that is replaced based on the context.

Demonstration of RoleRAG retrieval step

Question: {Beethoven, What was the nature of your relationship with Haydn and Mozart?}

Entity Extraction: {entity type, name, familiarity with Beethoven (reason, YES|No), Level (specific|general)}

(character<|>Beethoven<|>Beethoven must know himself<|>Yes<|>specific)

(character<|>Joseph Haydn<|>Haydn was Beethoven's teacher and a significant influence on his early compositions<|>Yes<|>specific)

(character<|>Wolfgang Amadeus Mozart<|>Mozart was a contemporary of Beethoven and a major influence on his work, although they did not have a direct relationship<|>Yes<|>specific)

Retrieved Content:

— Entity Information —

BEETHOVEN,CHARACTER, “Ludwig van Beethoven was a renowned German composer and pianist who significantly contributed to the landscape of classical music, transitioning between the late Classical and early Romantic periods. Born in Bonn, he moved to Vienna, ... ”

JOSEPH HAYDN,CHARACTER, “Joseph Haydn was a prominent composer who played a significant role in Beethoven’s musical education, particularly in the area of counterpoint. He was an esteemed figure in the Classical music genre and served as a pivotal influence during Beethoven’s early years in Vienna, introducing him to various musical concepts and techniques.”

MOZART,CHARACTER, “Wolfgang Amadeus Mozart was a prominent and renowned composer whose significant success and legacy played a crucial role in influencing Johann, particularly in promoting Beethoven as a child prodigy. His impact extended deeply into Beethoven’s formative years as a composer, shaping the young musician’s development and artistic direction.”

— Relation Information —

BEETHOVEN,JOSEPH HAYDN, “Beethoven, an influential composer in the history of classical music, studied under the guidance of Joseph Haydn, a prominent figure known for his significant contributions to the development of the symphony and string quartet. Under Haydn’s direction, Beethoven acquired essential skills in counterpoint and composition, which greatly influenced his early works. Additionally, Haydn played a crucial role in introducing Beethoven to a broader musical community and providing mentorship in Vienna, marking a pivotal development in Beethoven’s career. This relationship between the two composers highlights the impact of mentorship and collaboration in the evolution of classical music.”

BEETHOVEN,MOZART, “Beethoven, a prominent composer, was notably influenced by the style of Mozart, which played a significant role in his artistic development. This relationship highlights the profound impact that Mozart’s musical elements had on Beethoven’s compositions.”

Figure 7: Use case of retrieval step in our RoleRAG.

Prompt for Generating Knowledge Exposure Scores

Play act as {character}, who is {description}. I will provide questions from users and responses to those questions, where the responses are created in the style of you by other LLMs. You are required to judge and assess whether the response to a user's question match the knowledge and experience of you. To evaluate the response, consider the following aspects:

- (1) Whether stories/events discussed occurs in the same period as you.
- (2) Whether objects in the response have relevance to you.
- (3) Whether locations in the response are correct in your experience.
- (4) Whether persons mentioned have accurate relationships with you.

Based on the given instructions, provide a brief analysis of the given response. Then rate the response using a single score from 1 to 10, where a higher score indicates greater consistency with your knowledge.

Please provide your output in the following format:

Analysis: <analysis>

Rating: <rating>

Test Begin

User Question: {question}

Response: {response}

Output:

Figure 8: The prompt used for generating knowledge exposure scores.

Prompt for Generating Knowledge Hallucination Scores

Play act as {character}, who is {description}. I will provide questions from users and responses to those questions, where the responses are created in the style of you by LLMs. Based on your knowledge and experience, you must judge and assess whether the response to the question contains hallucination (fabricated or incorrect information). To evaluate hallucination, consider the following aspects:

- (1) Whether the events, objects, locations, or persons mentioned are consistent with your established story and background. A response that is not consistent with your lore is considered as hallucination.
- (2) Whether the response demonstrates a deep level of knowledge about a topic or concept that does not make sense for you to have, due to factors such as the topic not existing in your time period or universe. A response may refer to a topic if the question directly asks about the topic. However, answering the question with great detail would be considered as hallucination.

Based on the given instructions, provide a brief analysis of the given response. Then rate the response using a single score from 1 to 10, where a score of 1 represents a response with no hallucination and a 10 represents a response that does not make sense at all.

Please provide your output in the following format:

Analysis: <analysis>

Rating: <rating>

Test Begin

User Question: {question}

Response: {response}

Output:

Figure 9: The prompt used for generating knowledge hallucination scores.

Prompt for Generating Unknown Question Rejection Scores

Play act as {character}, who is {description}. I will provide questions from users and responses to those questions, where the responses are created in the style of you by other LLMs. You must judge and assess the ability whether the responses appropriately adhere to cognitive boundaries and reject questions that fall outside the scope of your knowledge and experience. Use the following guidelines to assess each response:

(1)Lore consistency: If the events, objects, locations, or persons mentioned are consistent with your established story and background, the response must align with your lore.

(2)Logical Rejection: For questions about topics or concepts that does not make sense for you to have, due to factors such as topics not existing in your time period or universe, the response should logically and explicitly reject the question rather than providing detailed or fabricated answers.

Based on the given instructions, provide a brief analysis of the given response. Then rate the response with a binary score [0, 1], where 1 indicates that the response follows the guidelines above, 0 means not follow the guideline.

Please provide your output in the following format:

Analysis: <analysis>

Rating: <rating>

Test Begin

User Question: {question}

Response: {response}

Output:

Figure 10: The prompt used for generating unknown question rejection scores.

Prompt for Response Generation on the Harry Potter Dataset

Please play as `{character}` in “Harry Potter” series and generate a response based on the dialogue context, using the tone, manner and vocabulary of `{character}`. You need to consider the following aspects to generate the character’s response:

- (1) Feature consistency: Feature consistency emphasizes that the character always follows the preset attributes and behaviors of the character and maintains consistent identities, viewpoints, language style, personality, and others in responses.
- (2) Character human-likeness: Characters naturally show human-like traits in dialogue, for example, using colloquial language structures, expressing emotions and desires naturally, etc.
- (3) Response interestingness: Response interestingness focuses on engaging and creative responses. This emphasizes that the character’s responses not only provide accurate and relevant information but also incorporate humor, wit, or novelty into the expression, making the conversation not only an exchange of information but also comfort and fun.
- (4) Dialogue fluency: Dialogue fluency measures the fluency and coherence of responses with the context. A fluent conversation is natural, coherent, and rhythmic. This means that responses should be closely related to the context of the conversation and use appropriate grammar, diction, and expressions.

Please answer in ENGLISH and keep your response simple and straightforward. If the question is beyond your knowledge, you should decline to answer and provide an explanation. Format each dialogue as: character name`{tuple_delimiter}`response. Remember do not provide any content beyond the character response.

```
#####context#####
{context_data}
```

----- Test Data -----

Character name: {character}

Question: {question}

Output:

Figure 11: The prompt used for generating responses on the Harry Potter dataset. We use the colon character (":") for `{tuple_delimiter}`.