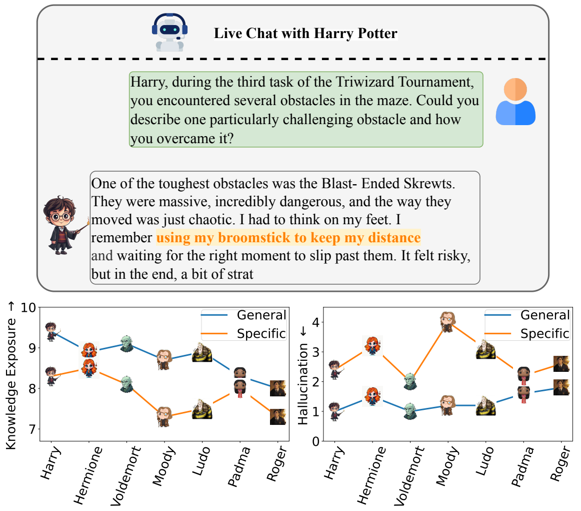
RoleRAG: Enhancing LLM Role-Playing via Graph Guided Retrieval

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have been used to imitate various characters to make conversations more engaging and immersive. However, LLMs fail to accurately capture the extensive knowledge specific to a given character, often generating hallucinated content that is irrelevant or inconsistent with the character's known information. To overcome this, we propose RoleRAG, a retrieval-based approach that includes (1) a graph-based indexing module that extracts the target role's experiences and relationships from a vast knowledge corpus and (2) an adaptive retrieval module that efficiently retrieves relevant information from the indexing system to ensure responses are accurate and contextually appropriate. We conduct 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. Our code is available at https://github.com/AnonymousSub123/RoleRAG.



Live Chat with Harry Potter

Harry, during the third task of the Triwizard Tournament, you encountered several obstacles in the maze. Could you describe one particularly challenging obstacle and how you overcame it?

One of the toughest obstacles was the Blast- Ended Skrewts. They were massive, incredibly dangerous, and the way they moved was just chaotic. I had to think on my feet. I remember **using my broomstick to keep my distance** and waiting for the right moment to slip past them. It felt risky, but in the end, a bit of strat

Figure 1: We tasked GPT-4o-mini to play seven characters from 'Harry Potter', presenting each character with 10 general questions (e.g., interests, attitudes) and 10 role-specific questions (e.g., experiences, activities), as depicted in the top figure, where hallucinated content is highlighted in orange. In the bottom figure, we recruit human raters to evaluate whether the language models accurately exhibit the characters' traits and to identify any instances of hallucination. The higher knowledge exposure and lower hallucination are better.

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.

However, LLMs often fail to accurately represent the character's knowledge, frequently responding without the appropriate character traits and introducing fabricated responses that contain experiences and objects unrelated to those characters. These issues become particularly severe when LLMs deal with role-specific details. As shown in Figure 1, we observe that LLMs exhibit fewer character traits and produce more hallucinations for specific questions than for general questions. We hypothesize that the failing cases could be attributed to knowledge scarcity, i.e., a lack of rich knowledge about the details of the target character.

To inject character-specific knowledge, 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 using datasets collected for role-playing scenarios. Despite the laborious data collection and computational burdens, fine-tuned-based approaches may not perform well for roles beyond the training corpus as different roles have their distinct knowledge base. Moreover, LLMs contain vasts amount of knowledge beyond the character they are portraying, and thus will often utilize that information when answering user queries, which likely contains fabricated elements especially for questions out of the character knowledge scope. Another family of research investigates the use of in-context learning by, for example, providing few-shot examples (Li et al., 2023) and using static user profiles (Wang et al., 2024a). However, given that the knowledge required to answer a given question may span a large amount of text, retrieving sufficient information to accurately answer questions is a challenge for these methods. To the best of our knowledge, in role-playing, few studies have investigated how to effectively pinpoint low-level query-related content from a large character dataset to provide relevant context for LLMs to reduce hallucination.

In this work, we introduce RoleRAG, a retrieval-based framework specifically designed for role-playing tasks. In particular, we incorporate graph structures into the indexing system and facilitate a nuanced retrieval process. The knowledge graph is created from the character knowledge base, such as Wikipedia profiles and books, where each node represents an entity from the character data, and each edge denotes the relationships between two entities. To remove duplicated entities that have different names, we propose an efficient entity normalization algorithm to merge them. Our retrieval module, built upon this graph indexing system, is designed to handle both specific and general entities mentioned in user questions and reject entities are out-of-scope. Information about the entities is retrieved from our knowledge graph, which is then given to the LLM to provide it with detailed context information to generate accurate responses.

Our proposed RoleRAG framework offers several advantages: (i) comprehensive information extraction, ensuring that graph information is extracted across multiple documents that far exceeds the token limits of LLMs; (ii) query-adaptive, allowing the retrieve module to adjust dynamically to the requirements of each query; (iii) rapid adaptation, enabling quick integration of new characters as long as character data is available. The graph creation is designed to be computationally efficient and easy to maintain.

Our contributions can be summarized as follows:

- A graph-based indexing module that extracts entities from a vast corpus for role-play tasks.
- An adaptive retrieval module that enhances LLMs' responses by providing sufficient, relevant character knowledge to the LLM.
- Extensive experiments that demonstrate that RoleRAG outperforms relevant baselines by exhibiting aligned character knowledge and reducing hallucinations.

2 Related Works

LLM-based Role-Play aims to enable LLMs to embody user-preferred characters, thereby enhancing user engagement and interest through conversation. Modern LLMs are pretrained for general purposes and often lack the experiential and emotional depth of characters. Therefore, three branches of approaches are introduced to inject character knowledge: (1) Finetuning-based approaches (Shao et al., 2023; Tu et al., 2024; Wang et al., 2024a; Zhou et al., 2024; Lu et al., 2024). This kind of approaches involve fine-tuning open source LLMs on collected character corpora. The training data is either synthetic (Shao et al., 2023; Tu et al., 2024; Wang et al., 2024a), or extracted from real dataset by LLMs (Zhou et al., 2024). Ideally, this training data should exhibit pronounced role characteristics and encompass a wide range of characters. (2) Retrieval-based method. These approaches (Salemi et al., 2024; Weir et al., 2024; Zhou et al., 2024) retrieve relevant documents from a character corpus to use as context for the LLM, thereby enhancing its ability to generate responses that are accurate and character-specific. Retrieval-based methods heavily depend on the quality of the retrieved content. (3) Plugin model. This method (Liu et al., 2024) keeps the LLM model frozen while encoding each user's corpus using a corresponding lightweight plugin model. The user embedding from the plugin model is concatenated with the embedding of the user's question to facilitate LLM generation. A comprehensive comparison of the three categories is provided in the Appendix A. In this work, we follow retrieval-based approaches, aiming to provide precise content relevant to user questions and reframing the role-playing task as a reading comprehension task, which modern LLMs can effectively handle.

Persona-based dialogue engages LLMs to embody a human-like persona. This method is similar to role-playing, but differs in that it focuses on broad personality traits, such as humor, empathy, or curiosity, rather than adhering to specific characteristics of a particular role. Unlike simple role-playing, persona-based dialogue requires the

LLM to exhibit attributes relevant to the assigned persona, such as displaying specific knowledge or behavioral traits that align with that persona. There are several ways to assign a persona to an LLM. These include prompting the model to act according to characteristics defined by the Big Five personality traits (Jiang et al., 2023), providing a text-based character profile (Tu et al., 2024; Zhou et al., 2024), or leveraging dialogue history (Zhong et al., 2022). Evaluation of such models can be performed using personality assessments and interviews (Wang et al., 2024b). In our work, however, we focus on equipping role-playing LLMs with the necessary information to provide accurate answers, rather than emphasizing personality traits.

Retrieval-Augmented Generation (RAG) is a technique described by Lewis et al. (Lewis et al., 2020) that retrieves information from an external knowledge base to enhance the capabilities of LLMs, particularly in generating responses that are informed, accurate, and contextually relevant (Liu et al., 2022; Zhuang et al., 2023; Li et al., 2024).

However, standard RAG struggles to capture intricate relationships between entities spanning multiple chunks (Guo et al., 2024) and also fails to answer general questions that require a thorough understanding of a vast database (Edge et al., 2024). To address these, a body of recent work (Edge et al., 2024; Sarmah et al., 2024; Wu et al., 2024; Guo et al., 2024) leverages LLMs to create knowledge graphs where each node represents the characteristics of an entity and each edge denotes the relationship between two entities. Consequently, hierarchically clustering was adopted to break the graph into several levels of sub-graphs, also known as communities, to facilitate hierarchical and high-level summarization (Edge et al., 2024). This approach allows the retrieval of higher-order neighborhood nodes, providing more relevant context (Guo et al., 2024; Sarmah et al., 2024; Wu et al., 2024). More discussion between knowledge graph and hallucination refers to (Agrawal et al., 2024). In our work, we propose a novel framework that leverage knowledge graph to retrieve precise entities associated with the target characters.

3 RoleRAG

Our overall framework for RoleRAG is depicted in Figure 2, which comprises two major modules: (1) constructing the knowledge graph from a character document; (2) given a question pertaining to the designed character, retrieving relevant content using the graph indexing system.

3.1 Entity and Relation Extraction

In role-playing scenarios, characters often originate from historical novels, television series, and celebrities, and their extensive background information typically surpasses the token limits of LLMs. To address this issue, we segment detailed character descriptions into manageable chunks $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$ in accordance with established practices (Edge et al., 2024; Wu et al., 2024; Guo et al., 2024). Each chunk is processed independently by LLMs and subsequently aggregated to form a cohesive output.

For each chunk $\mathcal{D}_i$, we employ LLMs to meticulously extract all possible entities, adhering to a predefined data structure: {*name*, *type*, *description*}, denoted by $\mathbf{n}_i$. We limit the entity types to character, location, time, event, organization, and object to streamline the extraction process. Furthermore, we prompt LLMs to identify structural relations between two entities, specifically, {*source*, *target*, *description*, *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 facilitate semantic retrieval, for each entity $\mathbf{n}_i$, we utilize a text embedding model that encodes both the entity name and its description into a high-dimensional vector $\mathbf{v}_i$. Subsequently, the node and vector pair $\{\mathbf{n}_i, \mathbf{v}_i\}$ is stored in the vector database $\mathcal{V}$, which allows for rapid retrieval based on semantic similarity. We represent the retrieval interface by $f_k(\mathcal{V}, \mathbf{n})$, which retrieves the top k entities most similar to a query entity from the vector database.

3.2 Entity Normalization

A character may have different names in various contexts, complicating the retrieval of relevant information. For example, critical story snippets may be missing due to the use of different names in user queries and story episodes, even though both names refer to the same individual. A straightforward example is found in the Harry Potter series, where both ‘Voldemort’ and ‘Tom Marvolo Riddle’ refer to the same character at different stages. Simply matching by ‘Voldemort’ would inevitably miss important story details concerning his early life. More challenging cases may arise when a char-

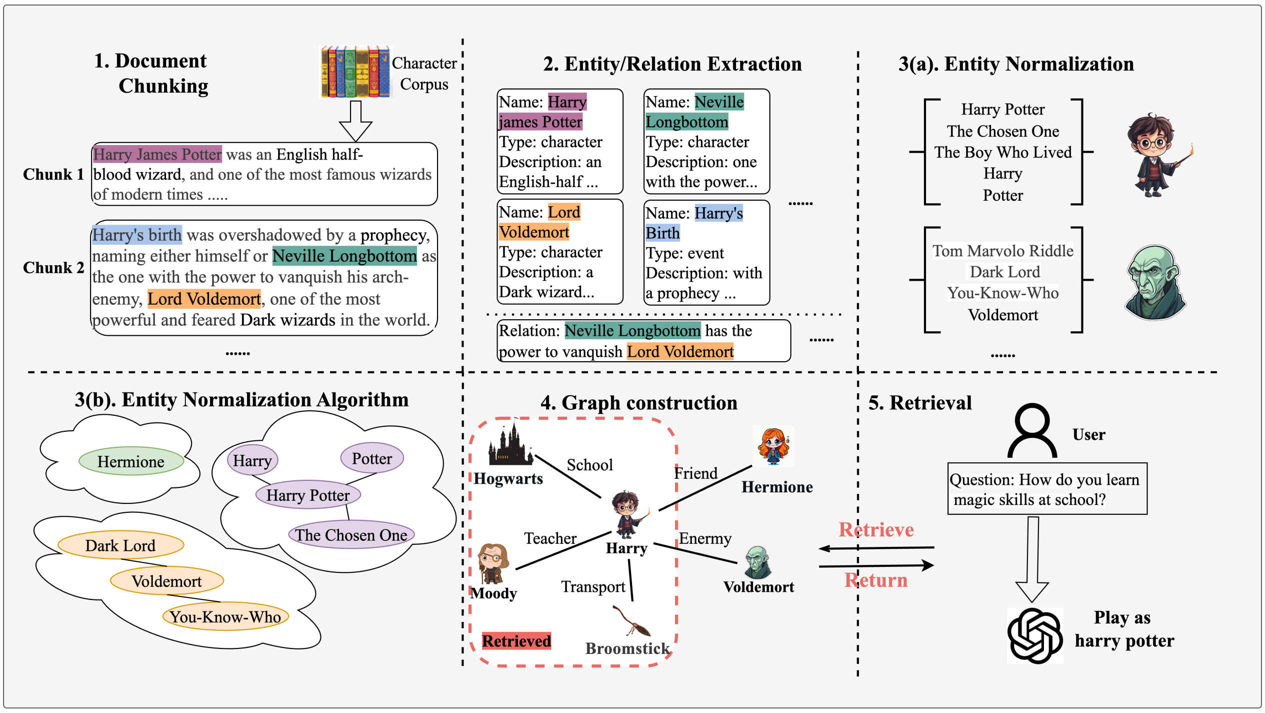

1. Document Chunking
Character Corpus
Chunk 1
Harry James Potter was an English half-blood wizard, and one of the most famous wizards of modern times .....
Chunk 2
Harry's birth was overshadowed by a prophecy, naming either himself or Neville Longbottom as the one with the power to vanquish his arch-enemy, Lord Voldemort, one of the most powerful and feared Dark wizards in the world.
......
2. Entity/Relation Extraction
Name: Harry james Potter
Type: character
Description: an English-half ...
Name: Neville Longbottom
Type: character
Description: one with the power...
Name: Lord Voldemort
Type: character
Description: a Dark wizard...
Name: Harry's Birth
Type: event
Description: with a prophecy ...
......
Relation: Neville Longbottom has the power to vanquish Lord Voldemort
......
3(a). Entity Normalization
Harry Potter
The Chosen One
The Boy Who Lived
Harry
Potter
Tom Marvolo Riddle
Dark Lord
You-Know-Who
Voldemort
......
3(b). Entity Normalization Algorithm
Hermione
Harry
Potter
Harry Potter
The Chosen One
Dark Lord
Voldemort
You-Know-Who
4. Graph construction
Hogwarts
School
Friend
Hermione
Teacher
Harry
Enermy
Moody
Transport
Voldemort
Retrieved
Broomstick
5. Retrieval
User
Question: How do you learn magic skills at school?
Retrieve
Return
Play as harry potter

Figure 2: Workflow of our proposed RoleRAG.

acter's short name, alias, and title are introduced in a particular episode.

To reduce entity ambiguity, we introduce a semantic entity normalization algorithm, which can be found in Appendix B. Given an entity database, we iterate over each entity to identify the most semantically similar k entities from the entity vector database. Then, we provide both the names and descriptions of entities to LLMs and prompt the LLM to determine whether two entities refer to the same character. If the entities are identified as same individual, we connect the nodes with an edge in the entity graph $\mathcal{G}$. After processing all entities, we partition the entity graph into several connected sub-graphs, each representing the same individual, as depicted in sub-figure 3(b) of Figure 2. Finally, we prompt the LLM once more to identify a unified name for each connected subgraph.

3.3 Graph Construction

After we identify groups of entities referring to the same character and establish a unified name within each group, we create a name mapping table that links source names to the unified name. Subsequently, we normalize all names in both the entity and relationship databases to ensure consistency and enhance search efficiency.

Entity normalization inevitably leads to identical entities and relations extracted from multiple data chunks $\mathcal{D}_i$, therefore we merge duplicate nodes and edges by summarizing their descriptions using LLMs. Finally, we formally construct the knowledge graph from character database as follows,

$$\hat{\mathcal{G}} = \{\hat{\mathcal{N}}, \hat{\mathcal{R}}\} \tag{1}$$

where $\hat{\mathcal{N}}, \hat{\mathcal{R}}$ denote nodes and relations after deduplication.

3.4 Retrieval Step

Upon receiving a user query, we initially employ an LLM to infer the hypothetical contexts relevant to the answer sought, inspired by HyDE (Gao et al., 2023). We then identify and extract entities present in both the original question and the hypothetical content. During the entity extraction process, we collect the following information: the entity's name, its relevance to the character for LLM role-playing along with the underlying reasons, and the entity's specificity level, either specific or general. This information induces three different retrieval strategies to obtain information to give to an LLM along with the character summary:

- For entities outside a character's knowledge scope (e.g., asking an ancient figure about Apollo 11), we elucidate the reasons for their irrelevance into LLM prompts.
- For specific entities, we first retrieve the top-k relevant entities from the vector database $\mathcal{V}$ whose

cosine similarity exceeds a predefined threshold. If we fail to retrieve any entities, it may indicate that the query entity does not exist in the character's knowledge base. Otherwise, we proceed to retrieve detailed information about these entities and their relationships with the designated character from the knowledge graph.

- For general entities such as interests, hobbies, the entity type is leveraged to retrieve entities with the same type from 1-hop neighborhood nodes of the target character from the graph.

These retrieval strategies not only improve the identification of related entities and relationships through keyword matching but also enrich the detail provided for general questions. More importantly, understanding the reasons for irrelevance can facilitate the dismissal of out-of-scope queries.

4 Experimental Setup

4.1 Baselines

We compare RoleRAG against the following set of baselines: **Vanilla**, which prompts an LLM to role-play as a given character when answering queries; **RAG** (Lewis et al., 2020), where relative information from chunked source material is retrieved based on the user's query. Following standard RAG procedures, the chunked texts are converted into text embeddings, and then text-embedding pairs are stored in a vector database. The text chunk that has the closest embedding to the query, will be provided to the LLM to use to generate its response; **Character profile** (Zhou et al., 2024), which provides the LLM with a profile of the character that the LLM is portraying.

For each character, we collect source materials from Wikipedia or Baidu Baike, which are used as the retrieval database for RAG and RoleRAG. For the character profile, we prompt GPT-4 to summarize the Wikipedia or Baidu Baike page into a short paragraph, which is then inserted before user questions to provide character background.

4.2 Evaluation Metrics

To perform our evaluation, we select metrics that were proposed in previous work (Tu et al., 2024; Lu et al., 2024). Role-play LLMs should seamlessly embody the designed role within a conversation, providing accurate and reliable answers to queries, maintaining character integrity throughout interactions. As shown in Figure 4 in Appendix C, we include (1) **Knowledge Exposure** refers to the extent to which personalized traits are recalled. This involves utilizing background, behavior, knowledge and experiences from the established characters. We use a scale that ranges from 1 to 10 where a higher score is better and represents a response that demonstrates deep knowledge about a character. (2) **Knowledge Hallucination** measures the precision of responses in conversation. It checks for the model's ability to avoid generating misleading, incorrect, or out-of-scope information. This criterion is crucial for maintaining the credibility of the LLM and ensuring that the information provided during role-play is precise and trustworthy from the viewpoint of the designed role. We use a scale that ranges from 1 to 10, where a lower score represents a response that is free from misinformation regarding a character and their background. (3) **Unknown Questions Rejection** measures the self-awareness in role-playing. This involves recognizing the limits of the character's knowledge and communicating these boundaries clearly to maintain realism and coherence in the role-play scenario. If a question is outside the knowledge base specific to the role, LLM models should clearly reject the question. Responses are scored on a binary scale of 0 or 1, where 1 is a better score given when the LLM correctly answers or rejects a given question.

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).

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 F. Following Ditto (Lu et al., 2024), we set the temperature of GPT-4o to 0.2 to penalize creativity during evaluation.

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 D.

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 tokens. 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. For each character, we construct a single graph and use it as a fixed retrieval database for all questions related to that character.

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 32k[1]). For further details, please refer to our public codebase.

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. In Vanilla methods, LLMs must rely on its own knowledge to answer character-related questions, so they perform worst. Although user profiles provide additional context for the LLM, it is not sufficient enough to ensure that the LLM can accurately answer a given question. RAG also provides more information to an LLM, the information given is dependent on the retrieval process, failing to answer questions that scatter across a large amount of text. RoleRAG organizes information so that it is easily accessible, allowing for the retrieval of relevant information regarding a character and their relationships to other characters, events, and objects to more accurately answer questions.

Fine-tuning models for role-play tasks can result in improved performance, as demonstrated by the results from Doubao Pro on RoleBench-zh dataset. However, there is an almost endless variety and abundance of possible characters that one could ask an LLM to portray, making it difficult to fine-tune a model to inject the vast amount of knowledge. The inferior performance of Doubao pro on Harry Potter and CharacterLLM datasets also verified this. RoleRAG allows a character's knowledge to be easily accessed by any LLM, as shown by our experimental results. Additionally, the knowledge graph in RoleRAG can be easily modified to incorporate new information, whereas a fine-tuned model would need to be retrained.

The results in Table 1 may appear to show only a marginal improvement. However, from our observations, this is due to the tendency of the judge LLMs to assign high scores for knowledge exposure and low scores for knowledge hallucination, as long as the response does not contain significant errors. For example, in the case of knowledge exposure, the judge LLM may give a score of 7 or

[1] https://www.volcengine.com/product/doubao

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 ↑ / ↓ means higher/lower results are better.

Model	Method	Harry Potter			RoleBench-zh			CharacterLLM ‡		
		KE ↑	KH ↓	UQJ ↑	KE ↑	KH ↓	UQJ ↑	KE ↑	KH ↓	UQJ ↑
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
	+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
	+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
	+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
	+Profile	8.021	2.050	—	6.457	3.014	0.754	9.258	1.272	0.964
	+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
	+Profile	8.221	2.021	—	6.232	3.754	0.733	9.009	1.317	0.945
	+RoleRAG	**8.821**	**1.571**	—	**6.994**	**2.697**	**0.857**	**9.138**	**1.211**	**0.978**
Close-source Role-playing Model										
Doubao 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
	+Profile	7.450	2.179	—	7.207	2.429	0.905	8.927	1.351	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], UQJ: Unknown question rejection {0, 1}.
‡ Human evaluation takes extremely longer on this dataset, we average scores from two trials of GPT4o.

8 as long as the response answers the question appropriately, with the human evaluator providing an additional 1 or 2 points for responses that include detailed information. Since the initial knowledge exposure score from the LLM is high, there is limited potential for improvement.

5.2 RoleRAG for general questions

Table 2 compares knowledge exposure and hallucination scores for general questions from the Harry Potter dataset, as evaluated by humans. These general questions pertain to personal interests, attitudes, and viewpoints. The results indicate that LLMs have lower hallucination score but exhibit few character traits. We hypothesize that LLMs have internalized knowledge pertinent to these high-level questions due to the vast datasets used to train them. From the character neighborhood, we retrieve 1-hop nodes that have the same type of general keywords. This provides richer detail, thereby significantly enhancing the exposure of character-related knowledge and reducing fabricated content.

5.3 RoleRAG for specific questions

Table 3 demonstrates knowledge exposure and hallucination scores for specific questions from the Harry Potter dataset, evaluated by humans. Compared with responses to general questions, when asked about details, LLMs tend to fabricate stories or are reluctant to provide specific information. As expected, 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.

5.4 RoleRAG for minority group

In Table 4, we report the performance across characters with varying frequencies in the Harry Potter series, with characters sorted by their frequency

Table 2: 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
Doubao Pro 32K	7.300	8.414	1.586	1.057

Table 3: 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
Doubao Pro 32K	7.085	8.029	2.929	2.071

Table 4: 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	$1.83_{-0.25}$

of appearance in the series. The results demonstrate that for popular 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.

5.5 RoleRAG for Out-of-scope questions

From Figure 3, we can see when LLMs are tasked to role-play, they have a tendency to answer all questions posed to them, even if the question is out of the character's scope of knowledge. This indicates that LLMs fail to fully assume the perspective of the target character and simply answer questions based on their internalized knowledge. This behaviour is shown even for more larger models like GPT-4o and Qwen2.5-Max. The high performance of Doubao Pro demonstrates that fine-tuning can improve an LLM's awareness of a character's cognition boundary. However, it cannot adapt to a new character without a fine-tuning dataset. Overall, however, regardless of an LLM's size or whether it is fine-tuned, the results show that RoleRAG provides LLMs the information to correctly reject out-of-scope questions, making an LLM's cognition boundary more aligned with a given character.

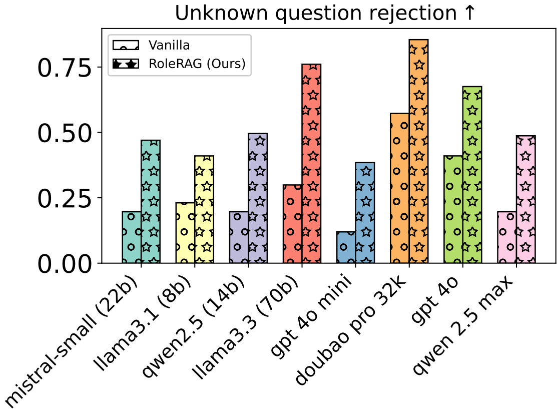


Figure 3: Experiments of out-of-scope questions in RoleBench-zh dataset.

6 Conclusion

When tasked with role-playing a specific character, LLMs often generate responses that are prone to various issues, such as hallucinations, insufficient depth of character knowledge, and the inclusion of information that falls outside the character's known universe. To address these issues, we introduced RoleRAG, a novel approach that constructs knowledge graphs from the source material associated with the character. These graphs allow for the retrieval of relevant character-specific information, which is then fed to the LLM during inference, ensuring that the model can draw from an accurate and consistent knowledge base. Through rigorous experimentation, we demonstrated that RoleRAG consistently outperforms relevant baselines by providing accurate and contextually appropriate responses. 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 Limitation and Future Work

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

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Can knowledge graphs reduce hallucinations in LLMs? : A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico. Association for Computational Linguistics.

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 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. 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 Kantarcioglu, 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.

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, Yuhan 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, Guanqun 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 5 shows a comparison of different methods used for using LLMs in role-playing tasks. Retrieval-based methods can avoid training a model, which means that data does not need to be labelled. Additionally, it means that existing LLMs can be used, whereas a fine-tuned LLM would require the LLM to be deployed, thereby increasing computational costs. To adapt to new roles or modify an LLM's knowledge, retrieval-based methods can extend or modify the knowledge base so that updated information can be retrieved for the LLM.

B Entity Normalization Algorithm

In this section, we provide the detailed Entity Normalization algorithm, as shown in Algorithm 1. Compared to the brute-force approach that compares every arbitrary pair of entities using LLMs, our proposed algorithm reduces the number of LLM invocations by $|\mathcal{N}|/k$, where $|\mathcal{N}|$ represents the total number of entities in $\mathcal{N}$ and k refers to the number of entities returned from the vector database. Modern efficient vector databases make this approach more efficient and cheaper compared to $|\mathcal{N}|^2$ LLM invocations.

C Evaluation Metrics

In this section, we present a figure that illustrates the purpose of each evaluation metric.

D Dataset Statistics

The statistics of our experimental datasets are illustrated in Table 6. 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 6: 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

Algorithm 1 Entity Normalization Algorithm

```
Require: Entity Database $\mathcal{N}$.
Ensure: a unified name for each name group.
Initialize empty entity graph $\mathcal{G}$.
Initialize empty vector database $\mathcal{V}$.
for $\mathbf{n}_i \in \mathcal{N}$ do
    if $\mathbf{n}_i \in \mathcal{V}$ then
        continue;                          ▷ node exists
    else
        $\mathcal{N}_k = f_k(\mathbf{n}_i, \mathcal{V})$
        Insert $\mathbf{n}_i$ to $\mathcal{V}$
        Insert $\mathbf{n}_i$ to $\mathcal{G}$
    end if
    for $\mathbf{n}_j \in \mathcal{N}_k$ do
        if $\mathbf{n}_i == \mathbf{n}_j$ then          ▷ LLM prompt
            Insert $\mathbf{n}_j$ to $\mathcal{G}$
            Connect $\mathbf{n}_i$ and $\mathbf{n}_j$ in $\mathcal{G}$
        else
            continue
        end if
    end for
end for
Count the number of connected components in $\mathcal{G}$
for each connected components $G$ in $\mathcal{G}$ do
    Select the unified name in $G$          ▷ LLM prompt
end for
```

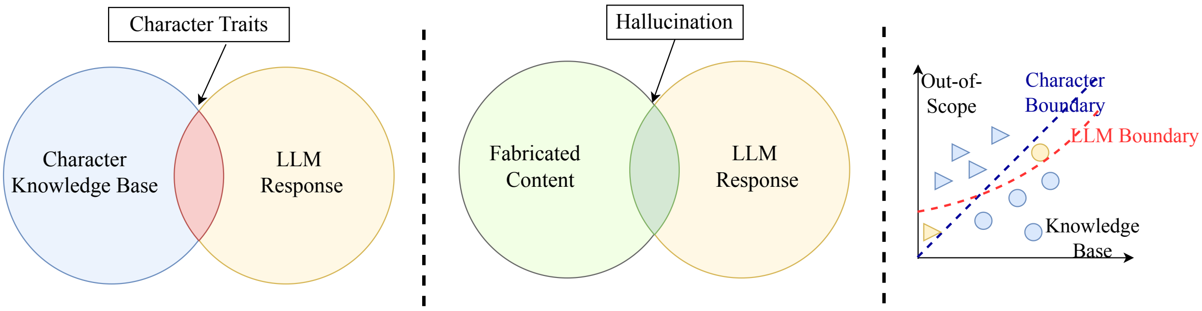
E Additional Experiments

E.1 Demonstration of RoleRAG retrieval

In Figure 5, we demonstrate the types of information retrieved from our RoleRAG system from an interview to LLM-played Ludwig Beethoven in CharacterLLM dataset. The question concerns the relationship between Beethoven, Haydn, and Mozart. Our system first identifies the entities within the question, their familiarity with Beethoven and the level of specificity of each entity. Since all three entities are specific and closely associated with Beethoven, our system directly provides information on these entities and their relationships from the knowledge database.

E.2 Word Clouds

In this section, we provide some word clouds in order to illustrate how LLMs' word usage changes to adapt to the role they are given. Figure 6 shows a word cloud for responses generated by GPT-4o mini when acting as Harry Potter, and Figure 7



(a) Knowledge Exposure **(b) Knowledge Hallucination** **(c) Unknown Question Rejection**

Figure 4: 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.

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 5: Use case of retrieval step in our RoleRAG.

Table 5: 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

shows a word cloud for Voldemort.

F 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.

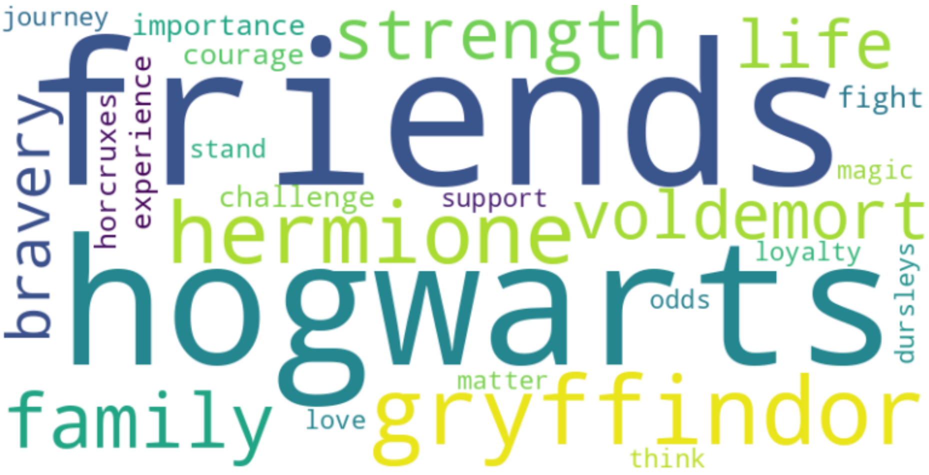


Figure 6: Word cloud for responses generated by GPT-4o mini when role-playing as Harry Potter.

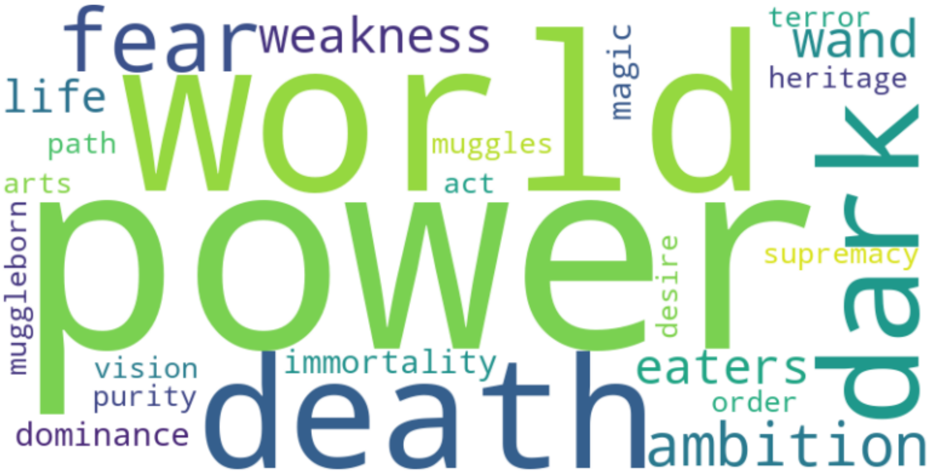


Figure 7: Word cloud for responses generated by GPT-4o mini when role-playing as Voldemort.

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.

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}.