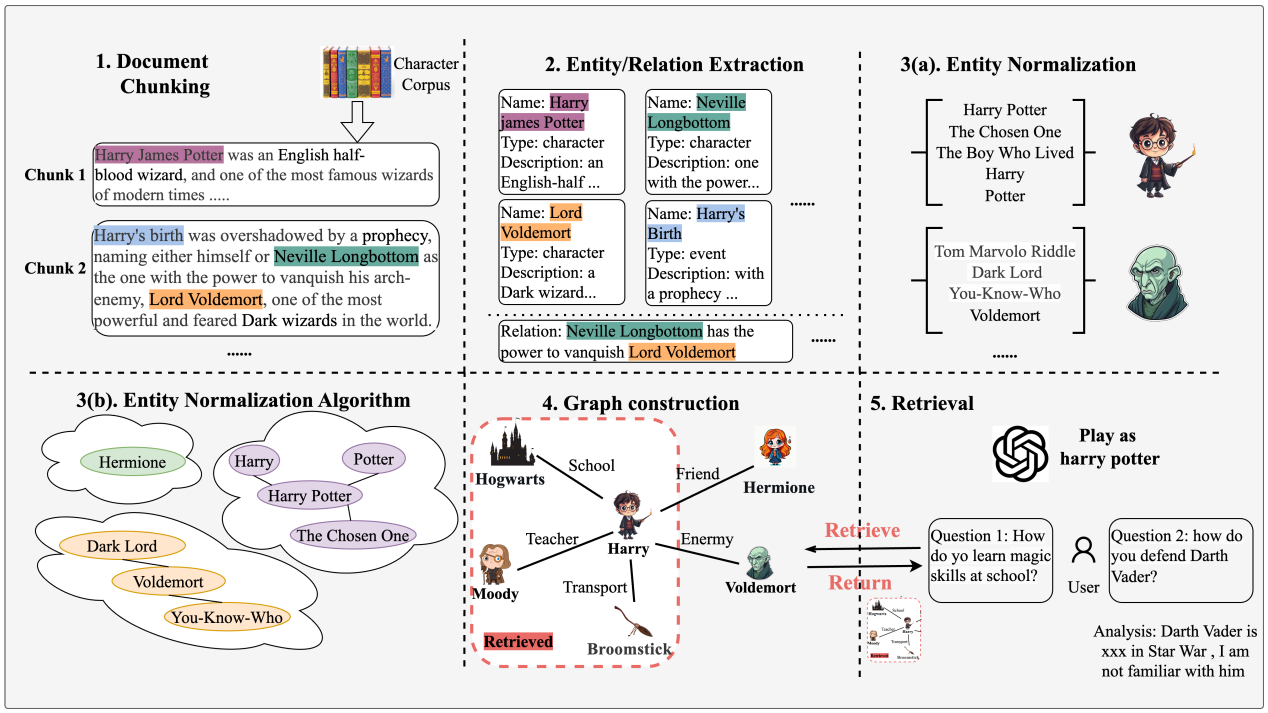


Figure 2: Workflow of our proposed RoleRAG.

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,

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

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-

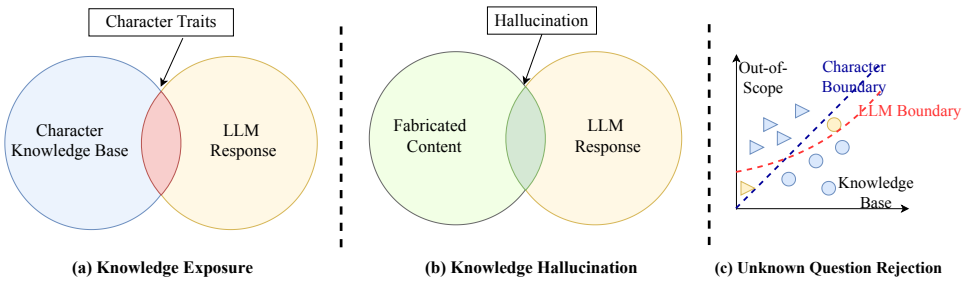


Figure 3: Illustration of evaluation metrics. We encourage LLMs to exhibit more personal traits, minimize fabricated content, and align more closely with the boundaries of character cognition.

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 32k[1]).

## 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.