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. Human evaluators are recruited to verify and correct GPT-4o's score.

Model	Method	Harry Potter			RoleBench-zh			CharacterLLM ‡		
		KE ↑	KH ↓	UQR ↑	KE ↑	KH ↓	UQR ↑	KE ↑	KH ↓	UQR ↑
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										
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
	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}.
‡ 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
Doubao 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
Doubao 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

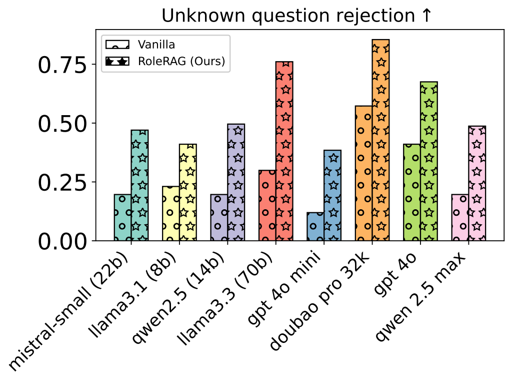

Figure 4: Experiments of out-of-scope questions in RoleBench-zh dataset.

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 Doubao 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*.