Revealing and Mitigating the Challenge of Detecting Character Knowledge Errors in LLM Role-Playing

Wenyuan Zhang[1,2], **Shuaiyi Nie**[1,2], **Jiawei Sheng**[1], **Zefeng Zhang**[1,2], **Xinghua Zhang**[3], **Yongquan He**[4], **Tingwen Liu**[1,2,*]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Alibaba Inc. [4]Meituan Inc.
{zhangwenyuan,nieshuaiyi,liutingwen}@iie.ac.cn

Abstract

Large language model (LLM) role-playing has gained widespread attention. Authentic character knowledge is crucial for constructing realistic LLM role-playing agents. However, existing works usually overlook the exploration of LLMs' ability to detect characters' known knowledge errors (KKE) and unknown knowledge errors (UKE) while playing roles, which would lead to low-quality automatic construction of character trainable corpus. In this paper, we propose RoleKE-Bench to evaluate LLMs' ability to detect errors in KKE and UKE. The results indicate that even the latest LLMs struggle to detect these two types of errors effectively, especially when it comes to familiar knowledge. We experimented with various reasoning strategies and propose an agent-based reasoning method, Self-Recollection and Self-Doubt (S^{2}RD), to explore further the potential for improving error detection capabilities. Experiments show that our method effectively improves the LLMs' ability to detect error character knowledge, but it remains an issue that requires ongoing attention[1].

1 Introduction

Large language models (LLMs) have the potential to be trained as specialized role-playing agents (RPA) (Tseng et al., 2024; Chen et al., 2024). Users provide a predefined character[2] profile (Zhou et al., 2024) to stimulate the RPA's human-like simulation abilities. The RPA's responses include the expected character style, knowledge, or behavior, which can support broader interdisciplinary NPC applications (Xu et al., 2024b; Wang et al., 2024a; Wu et al., 2024; Park et al., 2023). Current RPA training sets are primarily constructed purposefully based on character profiles and injected into general

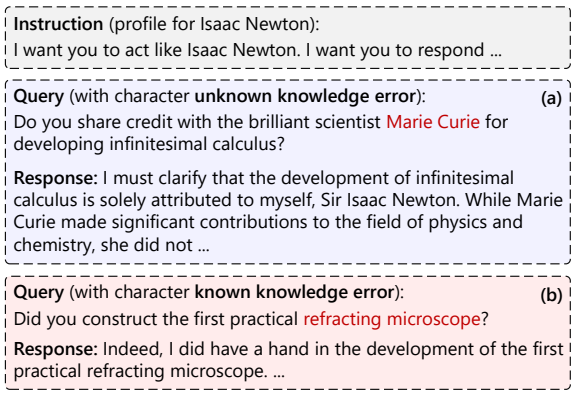

Figure 1: The real responses of GPT-3.5-turbo-0125 while playing Isaac Newton revealed some inconsistencies. In (a), although the LLM denied that Marie Curie was a scientist from Newton's time, it still showed an undue familiarity with her, exceeding the character's knowledge boundaries. In (b), the LLM incorrectly attributed the invention of the microscope, which was created before Newton's birth, to the wrong inventor.

LLMs. Inspired by the concepts of weak-to-strong generalization and self-instruction (Burns et al., 2024; Wang et al., 2023), the training of more powerful RPAs is gradually shifting from costly manual data annotation to automated character corpus construction. Through coordination among multiple LLM agents or self-alignment of a single LLM (Lu et al., 2024; Wang et al., 2024c), even small open-source LLMs can acquire diverse training corpora at low cost, unlocking powerful proprietary character capabilities (Shao et al., 2023).

The feasibility of generating character corpora stems from a fundamental capability of general LLMs: given a character profile, they can generate responses in a specific style (Wang et al., 2024b). However, this ability is fragile when it comes to knowledge of characters. When a query contains knowledge beyond the character's understanding, this knowledge can be termed as *unknown knowledge errors* (UKE), which may lead to unreliable responses. As shown in Figure 1 (a), the LLM

[*] indicates corresponding author.
[1]The resource is accessible at https://github.com/WYRipple/rp_kw_errors.
[2]In this paper, "character" also refers to "role".

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33268–33291
November 4-9, 2025

is instructed to play Isaac Newton. For Newton, Marie Curie is beyond his cognition. However, the model still identifies her contributions in the field of chemistry, even exhibiting consistent behavior, such as clarification. Furthermore, if a query contains incorrect knowledge within the character's cognition, such knowledge can be referred to as *known knowledge errors* (KKE), resulting in inaccurate responses. As shown in Figure 1 (b), the LLM also fails to rectify the inventor of the microscope, which is familiar to Newton. These potential errors will significantly affect the reliable construction of corpora and ultimately undermine the training of RPA (Shao et al., 2023).

There is still little exploration of the ability of general LLMs to identify such knowledge errors. Thus, we formalize the problem to investigate: *How effective can LLMs detect knowledge edge errors when playing roles?* Inspired by Conway and Pleydell-Pearce (2000), we meticulously construct a **Role K**nowledge **E**rror Detection **Bench**mark (**RoleKE-Bench**) to explore this issue, using four memory types to categorize knowledge (*event*, *relation*, *attitudinal*, and *identity* memory). The benchmark construction is divided into two stages. First, the character's wiki corpus is deconstructed into multiple correct memories, and then two types of knowledge errors are injected to simulate queries during automated corpus construction. LLMs are required to challenge and correct KKE, while expressing doubt or refusal in response to UKE.

For further investigation, we evaluate 21 advanced LLMs, including DeepSeek-R1, and find that when playing different roles, 1) *both types of errors are difficult to detect, with the highest accuracy not exceeding 65%*; 2) *LLMs are more prone to making KKE, about 15% lower than UKE.* The poor performance stems from similar semantic representations of correct and incorrect memories, and the rich world knowledge learned in the LLMs. To mitigate this, we further propose an agent-based reasoning augmented method, Self-Recollection and Self-Doubt (S^2RD). Self-Recollection mimics the human behavior of recalling clues, then consulting notes when faced with vague memories, keeping LLMs' attention off incorrect semantics. Self-Doubt is a critical self-examination that helps LLMs understand character knowledge boundaries. S^2RD has effectively enhanced detection capabilities, showcasing LLMs' potential for identifying character error knowledge.

Our main contributions are as follows:

(1) We formalize and explore the LLMs' ability to detect two types of character knowledge errors, crucial for future reliable corpora construction.

(2) We construct RoleKE-Bench and find LLMs are not proficient at detecting errors, particularly with character known knowledge errors.

(3) We propose an agent-based reasoning method that effectively enhances the character knowledge error detection capabilities of LLMs.

2 Related Work

Role-play in LLMs. LLMs are gradually being discovered to function as role-playing agents (Chen et al., 2024) with the potential to simulate various styles (Shanahan et al., 2023; Yu et al., 2024), attributes (de Araujo and Roth, 2024) and personality (Wang et al., 2024d; Choi and Li, 2024). They can be applied in a wide range of applications, such as emotional companion robots (Feng et al., 2025a; Sabour et al., 2024; Tan et al., 2024), chatbots with specific personalities (Tu et al., 2023; Zhou et al., 2024), social role interactions (Wang et al., 2025d,a; Zhang et al., 2025a), drama interaction (Wu et al., 2024), educational system (Wang et al., 2024a), cybersecurity (Huo et al., 2024) and healthcare (Xu et al., 2024b). However, current research may be limited in application due to the influence of KKE and UKE.

Role-play corpora construction. Current research primarily focuses on constructing RPA corpora to enhance the effectiveness of character portrayal. There are two types of corpora construction methods leverage LLMs: *LLMs as tools* and *LLMs as sources*. Using *LLMs as tools* can be regarded as a semi-automated method. Many efforts utilize the extraction (Xu et al., 2024a) and summarization (Subbiah et al., 2024) capabilities of LLMs to filter and collect role-playing scenes and dialogues from existing scripts (Han et al., 2024), books (Chen et al., 2023) or film works (Li et al., 2023a). Thanks to the rich character experiences encoded in LLMs, using *LLMs as sources* for an automated method is being explored. These methods allow LLMs to query each other as agents (Wang et al., 2025c), with profiles (Yuan et al., 2024) containing character requirements serving as the context. Shao et al. (2023) simulated dialogue scenarios, immersively generating conversational corpora; Lu et al. (2024) employed self-alignment to allow corpora to be generated by itself; Chan et al. (2024) automatically synthesized a massive scale of role

dialogue amounting to billions. This type of automated method holds promise due to its advantages in large-scale scalability and flexibility. However, there is a lack of works addressing the ability of LLMs to detect characters' knowledge errors in automatic data construction, resulting in potential uncertainties and warranting attention.

3 Problem Formulation

3.1 Character Knowledge Taxonomy

We first delve deeper into the composition of the character's knowledge. In first-person immersive role-playing, the characters' responses should be shaped by the limits of their profiles. The profiles trigger their specific memories, within which knowledge is embedded. By refining the categories of memory, we can more clearly articulate how character's knowledge is expressed in different memory contexts. Based on the *Self-Memory System* (SMS) (Conway and Pleydell-Pearce, 2000), which explains how autobiographical memory interacts with the working self to construct personal identity, we divide memory into four types: ***Event Memory*** refers to the recollection of specific personal experiences, corresponding to event-specific knowledge in SMS and involving detailed memories of time, place, and events; ***Relation Memory*** pertains to memories of interpersonal relationships and social connections, manifesting in the understanding of social roles and long-term relationships; ***Attitudinal Memory*** reflects an individual's emotional responses and attitudes toward events or people, associated with the working self in SMS and influencing personal goals and emotional states; ***Identity Memory*** integrates elements from the autobiographical memory knowledge base with self-concept from the working self in SMS, reflecting the development and cognition of personal identity. This taxonomy enriches the diversity of character knowledge, enabling a more comprehensive exploration of LLMs' error detection capabilities across different types of memory.

3.2 Character Knowledge Errors

Due to the creativity (Chakrabarty et al., 2024) in LLMs, queries that incorporate the aforementioned memory categories may contain unpredictable errors. As claimed in Introduction, these errors can be divided into two types:

Known knowledge Errors (KKE) occur when a character confuses or misstates known facts during a query. These are errors the characters can potentially recognize and correct.

Unknown knowledge Errors (UKE) arise when the LLMs' vast knowledge leads a character to reference concepts that are anachronistic or beyond their understanding. For a more detailed conceptual explanation, see Appendix A.

3.3 Task Definition

In this section, we formally introduce the task of *character knowledge error detection*. Given a role agent $\mathcal{A}$, a role profile text p_c, and a query q_{error} containing errors to be identified, we obtain the open-ended response r_c from the agent:

$$r_c = \mathcal{A}(p_c, q_{error}; \bar{\theta}), \tag{1}$$

where $c \in \mathcal{C}$ denotes a character from the list of characters $\mathcal{C}$, and $\bar{\theta}$ represents the frozen parameters of the agent. The task is ultimately analyzed by an evaluator to determine whether r_c can identify and correct KKE in q_{error}, or express confusion or refuse it when it contains UKE.

4 RoleKE-Bench

We propose RoleKE-Bench, focusing on simulating queries across different memory types while injecting two types of errors. The character list and profiles follow Shao et al. (2023), and include nine well-known real or literary characters, which have been well-encoded by the LLMs. The construction process, illustrated in Figure 2, is divided into two main steps as follows. All steps involve both automatic construction by GPT-4o and comprehensive manual verification. We recruit and finalize three evaluators who are familiar with the objectives of RoleKE-Bench and have extensive experience in data engineering[3].

4.1 Correct Memory Generation

We first collect and store Wikipedia data for various characters, then segment the content into multiple chunks based on each "\n\n". All chunks are reviewed by three evaluators to ensure the inclusion of complete character milestones. Chunks that are incompletely described are discussed and their boundaries are redefined through negotiation.

Next, we prompt GPT-4o to generate multiple concise first-person statements from each chunk, all representing correct character memories, which

[3]For details on recruitment and the human filtering requirements across different stages, please refer to Appendix B.

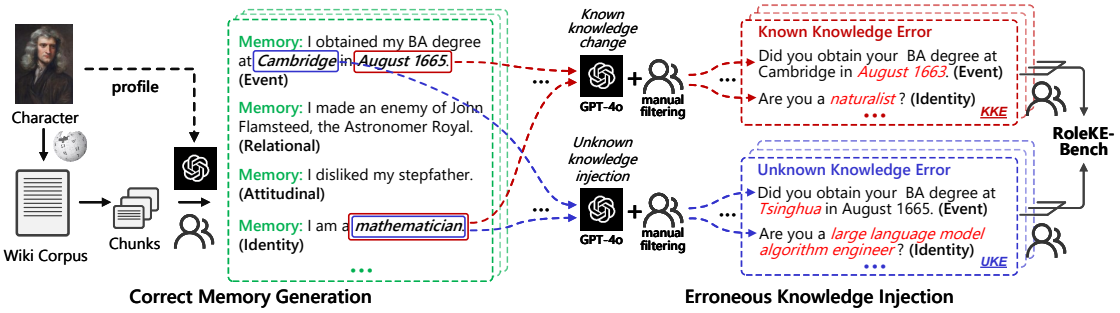

Figure 2: Overview of Probing Dataset construction. First, we create correct character memories, which encompass the knowledge that the character should proficiently possess. Second, we inject erroneous knowledge, simulating both types of errors and preserving the modification details, which results in final queries.

Memory Category	KKE	UKE	Total
Event Memory	300/17.7	300/24.2	600/20.9
Relational Memory	56/14.7	56/19.8	112/17.2
Attitudinal Memory	70/17.9	70/21.3	140/19.6
Identity Memory	69/13.4	69/14.8	138/14.1
Total	495/16.8	495/22.0	990/19.4

Table 1: The statistical details of RoleKE-Bench. The left side of "/" represents the sample size, while right side represents the average number of words per query.

GPT-4o also categorizes automatically. To ensure the correct of memories and their categories, meticulous manual screening is conducted. Only retain the following generations: 1) *the memory category label is correct*, 2) *the memory contains key details* (e.g., the event can be uniquely identified from the context) and 3) *the memory is concise (fewer than 30 words)*. We retain the intersection of the selections made by the three evaluators, with an overlap reaching 85.6%.

4.2 Erroneous Knowledge Injection

Subsequently, each correct memory is injected with KKE and UKE to generate two corresponding erroneous memories. Specifically, GPT-4o is provided with the original chunk, the correct memory, and detailed instructions for error injection to generate erroneous memories along with the rationale for each modification. We require that each erroneous memory contain only a single error. For KKE, only minor modifications at the span level are allowed, ensuring that the modified memory remains consistent with the character's cognition and that the error is correctable. For UKE, we introduce a set of subdisciplines (details in Appendix C.2) and randomly assign two terms as reference topics during each modification. The resulting erroneous memories are finally converted into queries by GPT-4o, as shown in Figure 2.

In the above process, the evaluators review and filter all erroneous memories, retaining only those that meet the following criteria: 1) *the errors conform to the defined standards (the former being correctable and the latter exceeding the character's cognition)*, and (2) *each memory contains only one error*. Erroneous memory pairs are discarded if either fails to meet standards. The intersection of the evaluators' screening results is retained as the candidate set (81.1% retention). Finally, the evaluators examine all queries, discuss any inconsistencies, and keep only the qualified samples to construct the RoleKE-Bench.

4.3 Benchmark Statistics

The RoleKE-Bench ultimately consists of two groups of queries, containing known and unknown character knowledge errors. After meticulous selection, a total of 990 queries were ultimately obtained, corresponding to 495 correct memories. The benchmark statistics are illustrated in Table 1, with details in Appendix C.1. We retain the original chunks and modified explanations as crucial references for evaluation. Details on data collection and filtering are in Appendix B, with all data construction prompts in Appendix G.

5 Methodology

Recently, enhancing character capabilities through long-chain reasoning has been proven to have potential (Wang et al., 2025b), but it does not always work effectively (Feng et al., 2025b; Zhang et al., 2025b), possibly due to insufficient focus on character-specific knowledge. Inspired by how humans reference and reflect on ambiguous memories, we propose the agent-based S^2RD reasoning method, which addresses this gap by placing

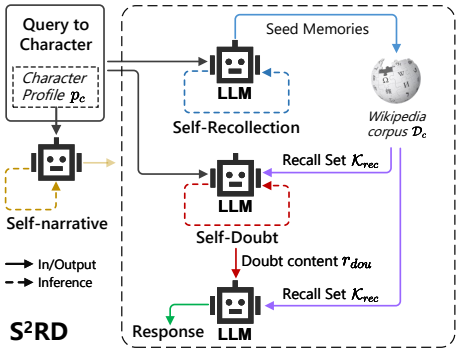

Figure 3: Overview of S^{2}RD. First, the model restates the character based on the profile, and this narrative serves as input for all subsequent agents. Then, it undergoes two steps of reasoning: self-recollection and self-doubt. Finally, all results are combined into the context of the last agent to detect errors.

greater emphasis on such knowledge. Firstly, inspired by Choi and Li (2024), we prompt the LLM to reaffirm the character's identity, generating a self-narrative statement r_{nar}. The statement then becomes the input for subsequent reasoning steps. Then agents iterate between *self-recollection* and *self-doubt*, with the final agent using these generations to provide the LLM with more reliable priors. Figure 3 illustrates the overview of our method.

5.1 Self-Recollection

Self-Recollection refers to the process where LLMs don't directly answer a query but instead recall knowledge indirectly related to it. This enables LLMs to generate approximate knowledge as seed memory, mimicking how humans recall key memory cues. After generating m seed memories, the model uses these as retrieval points, simulating the way humans reference notes based on memory cues, to search for factual knowledge within the character's wiki corpus. The process can be formalized as:

$$\mathcal{K}_{rec} = RAG(\mathcal{A}(p_c, r_{nar}, q_{error}; \bar{\theta}), \mathcal{D}_c), \quad (2)$$

where $RAG(\cdot)$ is the retrieval method (same as Section 6.1), and $\mathcal{D}_c$ represents the Wikipedia corpus of character c. $\mathcal{K}_{rec}$ is the recall set of m seed memories, with $m = 3$ in this paper. Ultimately, the LLMs' self-generated knowledge is refined through retrieval, reducing the risk of being misled by semantically similar incorrect knowledge.

5.2 Self-Doubt

Self-Doubt aims at encouraging LLMs to focus more on detecting incorrect actions. Unlike reflection (Ji et al., 2023), doubt emphasizes criticism, and its strong purposefulness makes it easier for them to generate reasonable refutations to erroneous questions, which can be formalized as:

$$r_{dou} = \mathcal{A}(p_c, r_{nar}, \mathcal{K}_{rec}, q_{error}; \bar{\theta}), \quad (3)$$

where r_{dou} represents the content of the doubt statement, helping the LLM adhere more closely to the profile and preventing out-of-character responses.

As shown in Figure 3, our approach leverages the outputs from the two distinct phases as the final inference context, and provide several cases to guide LLMs' inference. The S^{2}RD forces the LLM to pay closer attention to character boundaries, providing more reliable references for its responses. All prompts can be found in Appendix G.

6 Evaluation

Base Models.

Base Models. We evaluated on 21 LLMs, including the proprietary Large Reasoning Models (LRMs) and open-source LLMs. We also focus on the LLMs with role-play expertise. For details on these LLMs, refer to Appendix D.

Evaluation Metrics. Inspired by the "LLMs as Judges" (Zhang et al., 2023), we provide LLM as evaluator. LLMs take the character profile p_c and the query q_{error} as inputs to infer and produce the response. The evaluator takes the open-ended responses of the LLMs when playing a specific character, along with the memory modification explanations, as input, and assesses whether the LLM correctly identifies (for KKE) or expresses doubt/refuses (for UKE) the character error in the query. The evaluator outputs a rationale and a *yes*/*no* judgment. Accuracy is the ratio of *yes* responses over all queries three times along with the standard error of the mean (SEM). Judgment prompt details are in Appendix G.

Evaluator determination. We selected DeepSeek-v2 (DeepSeek-AI, 2024) rather than GPT-4o as the evaluator. This choice helps avoid self-bias (Li et al., 2023c; Xu et al., 2024c), as the RoleKE-Bench is generated by GPT-4o, while still maintaining evaluation capabilities similar to GPT-4o. Additionally, it offers a significantly lower cost compared to many advanced LLMs. We conduct a human evaluation experiment to validate the rationale behind the above selections, with details provided in Appendix E.

Model	Known Knowledge Errors (KKE)					Unknown Knowledge Errors (UKE)				
	Eve-Mem.	Rel-Mem.	Att-Mem.	Ide-Mem.	Average	Eve-Mem.	Rel-Mem.	Att-Mem.	Ide-Mem.	Average
LRMs Baselines (Open-sourced)										
DeepSeek-R1	**44.17** $\pm$ **0.10**	42.56 $\pm$ 0.79	49.76 $\pm$ 0.95	57.49 $\pm$ 3.56	**60.35** $\pm$ **0.39**	53.83 $\pm$ 0.25	74.40 $\pm$ 0.60	31.19 $\pm$ 1.56	61.11 $\pm$ 2.15	55.68 $\pm$ 0.79
QwQ-32B	39.44 $\pm$ 0.44	38.69 $\pm$ 0.60	49.05 $\pm$ 2.65	56.52 $\pm$ 4.43	56.15 $\pm$ 0.56	59.78 $\pm$ 0.59	76.79 $\pm$ 1.03	39.52 $\pm$ 2.08	**69.08** $\pm$ **1.28**	61.97 $\pm$ 0.19
DS-Qwen-32B	31.44 $\pm$ 0.24	31.70 $\pm$ 1.03	39.17 $\pm$ 0.24	44.20 $\pm$ 2.54	42.81 $\pm$ 0.18	38.69 $\pm$ 0.03	51.04 $\pm$ 0.30	18.81 $\pm$ 0.72	40.10 $\pm$ 1.68	38.43 $\pm$ 0.37
DS-Qwen-7B	31.81 $\pm$ 0.20	31.55 $\pm$ 0.34	38.89 $\pm$ 0.16	43.16 $\pm$ 1.40	43.03 $\pm$ 0.29	40.67 $\pm$ 0.00	53.37 $\pm$ 0.99	22.06 $\pm$ 1.11	43.16 $\pm$ 1.90	40.81 $\pm$ 0.38
General Baselines (Proprietary)										
GPT-4o	39.33 $\pm$ 0.19	**43.45** $\pm$ **1.57**	51.43 $\pm$ 1.65	58.94 $\pm$ 1.93	44.24 $\pm$ 0.23	54.56 $\pm$ 0.97	69.05 $\pm$ 1.57	24.29 $\pm$ 2.18	56.52 $\pm$ 0.84	52.19 $\pm$ 0.44
GPT-3.5	15.11 $\pm$ 0.11	22.02 $\pm$ 1.57	38.57 $\pm$ 2.18	47.83 $\pm$ 0.84	23.77 $\pm$ 0.49	27.56 $\pm$ 0.11	29.17 $\pm$ 2.15	10.95 $\pm$ 0.48	26.57 $\pm$ 4.61	25.25 $\pm$ 0.58
ERNIE4	24.56 $\pm$ 0.48	21.43 $\pm$ 1.79	47.62 $\pm$ 2.65	54.59 $\pm$ 2.11	31.65 $\pm$ 0.29	49.89 $\pm$ 0.29	63.69 $\pm$ 0.60	25.24 $\pm$ 2.08	55.07 $\pm$ 2.21	48.69 $\pm$ 0.31
Qwen-max	27.89 $\pm$ 1.06	29.17 $\pm$ 2.15	46.19 $\pm$ 4.97	59.90 $\pm$ 1.28	35.08 $\pm$ 0.82	54.78 $\pm$ 0.11	67.26 $\pm$ 2.59	37.62 $\pm$ 2.38	61.35 $\pm$ 5.38	54.68 $\pm$ 0.58
Yi-Large	25.33 $\pm$ 0.19	30.95 $\pm$ 0.60	40.95 $\pm$ 1.26	56.52 $\pm$ 1.67	32.53 $\pm$ 0.31	46.11 $\pm$ 0.29	67.86 $\pm$ 1.79	31.90 $\pm$ 0.95	52.66 $\pm$ 2.42	47.47 $\pm$ 0.23
GLM-4	23.44 $\pm$ 0.73	26.79 $\pm$ 0.00	40.95 $\pm$ 0.95	48.31 $\pm$ 1.28	29.76 $\pm$ 0.34	41.00 $\pm$ 0.69	62.50 $\pm$ 0.00	16.67 $\pm$ 1.72	53.62 $\pm$ 0.84	41.75 $\pm$ 0.41
Role-play Expertise Baselines										
ERNIE-Char	14.44 $\pm$ 0.11	19.64 $\pm$ 2.73	31.43 $\pm$ 0.82	33.82 $\pm$ 3.38	20.13 $\pm$ 0.66	42.22 $\pm$ 1.11	50.00 $\pm$ 1.03	30.95 $\pm$ 1.90	53.14 $\pm$ 3.77	43.03 $\pm$ 1.11
CharacterGLM	11.56 $\pm$ 0.80	19.05 $\pm$ 0.60	24.76 $\pm$ 4.69	31.88 $\pm$ 1.67	17.10 $\pm$ 1.28	28.67 $\pm$ 0.33	24.40 $\pm$ 3.90	32.86 $\pm$ 3.60	28.99 $\pm$ 1.45	28.82 $\pm$ 0.78
Baichuan-NPC	16.11 $\pm$ 1.72	25.00 $\pm$ 2.73	35.24 $\pm$ 3.12	41.06 $\pm$ 2.69	25.21 $\pm$ 1.70	47.00 $\pm$ 0.69	60.71 $\pm$ 2.73	28.10 $\pm$ 2.08	50.72 $\pm$ 0.84	47.45 $\pm$ 0.87
MiniMax	19.94 $\pm$ 1.63	26.49 $\pm$ 2.08	36.43 $\pm$ 0.00	49.03 $\pm$ 1.05	30.76 $\pm$ 1.47	52.22 $\pm$ 0.71	64.88 $\pm$ 0.79	30.48 $\pm$ 1.72	60.39 $\pm$ 0.87	52.91 $\pm$ 0.49
Xingchen-Plus	29.33 $\pm$ 1.20	31.55 $\pm$ 2.15	45.71 $\pm$ 0.82	55.07 $\pm$ 3.83	39.36 $\pm$ 1.38	57.44 $\pm$ 0.99	68.45 $\pm$ 2.59	34.76 $\pm$ 1.26	67.15 $\pm$ 2.42	58.17 $\pm$ 0.28
General Baselines (Open-sourced)										
DeepSeek-v2	25.33 $\pm$ 1.71	29.76 $\pm$ 2.38	40.00 $\pm$ 0.82	58.45 $\pm$ 1.74	32.53 $\pm$ 1.00	52.22 $\pm$ 0.73	67.86 $\pm$ 1.03	37.62 $\pm$ 0.95	64.25 $\pm$ 1.74	53.60 $\pm$ 0.60
LLaMA3-70b	22.22 $\pm$ 1.28	27.38 $\pm$ 2.38	**53.81** $\pm$ **0.48**	**60.87** $\pm$ **1.45**	32.66 $\pm$ 0.83	**65.22** $\pm$ **0.68**	**77.38** $\pm$ **0.60**	50.48 $\pm$ 2.08	68.60 $\pm$ 2.42	**64.98** $\pm$ **0.79**
Qwen2-72b	26.07 $\pm$ 1.27	34.26 $\pm$ 3.24	47.22 $\pm$ 2.16	52.46 $\pm$ 2.82	33.65 $\pm$ 1.61	59.88 $\pm$ 0.98	74.74 $\pm$ 2.56	37.57 $\pm$ 0.50	68.22 $\pm$ 1.68	59.73 $\pm$ 0.82
Mixtral-v0.1	26.78 $\pm$ 1.35	32.74 $\pm$ 3.15	50.48 $\pm$ 2.90	51.69 $\pm$ 3.48	34.28 $\pm$ 0.99	51.44 $\pm$ 0.22	55.95 $\pm$ 0.60	36.19 $\pm$ 2.08	63.29 $\pm$ 1.28	51.45 $\pm$ 0.24
LLaMA3-8b	18.22 $\pm$ 0.11	23.21 $\pm$ 1.03	44.29 $\pm$ 0.82	50.72 $\pm$ 1.45	27.00 $\pm$ 0.18	55.22 $\pm$ 1.18	70.83 $\pm$ 1.57	**55.24** $\pm$ **2.08**	63.77 $\pm$ 1.45	58.18 $\pm$ 1.23
Qwen2-7b	7.11 $\pm$ 0.80	19.05 $\pm$ 1.19	28.57 $\pm$ 1.43	30.92 $\pm$ 0.48	14.81 $\pm$ 0.67	29.56 $\pm$ 0.91	27.38 $\pm$ 3.31	17.14 $\pm$ 1.43	48.31 $\pm$ 1.74	30.17 $\pm$ 0.55

Table 2: Evaluation results of the RoleKE-Bench. The results present the average accuracy with standard error of the mean (SEM) after three times of evaluations. The bold indicates the best, and the underlined indicates the second best. Eve-Mem., Rel-Mem., Att-Mem. and Ide-Mem. are abbreviations for four types of memories.

6.1 Baseline Methods

We use widely adopted reasoning-augmented methods as baselines across multiple reasoning tasks (Li et al., 2023b; Ahn et al., 2024; Zeng et al., 2024).

Vanilla directly uses the character system prompts and questions as input to LLMs to assess their basic capabilities based on probing dataset.

CoT (Kojima et al., 2022) enhances reasoning ability by appending "Please think step by step and then answer" at the end of the queries.

Few-shot involves adding four pairs of memory query-response examples before each question. We carefully construct queries that do not overlap with the probing dataset, and add correct memories as prompts for GPT-4o to generate correct answers.

Self-Reflection (Ji et al., 2023; Shinn et al., 2023) has been mentioned in recent researches, highlighting that LLMs possess an inherent reflective capability, which can distill correct knowledge. Inspired by this, we design a two-stage query process. The first stage is Vanilla, followed by reflection on the prior response and a revised reply.

Retrieval-augmented generation (RAG) has been proven effective in mitigating LLM hallucination issues (Gao et al., 2023). We designed a retrieval module using `all-MiniLM-L6-v2`[4] as the query encoder and character Wikipedia corpus as retrieval source with LangChain framework[5]. For each query, we retrieve three pieces of data to serve as the context for each LLMs.

RAG+Few-shot is a method of combining RAG and Few-shot, aiming to allow LLMs to inherit the respective advantages of both methods.

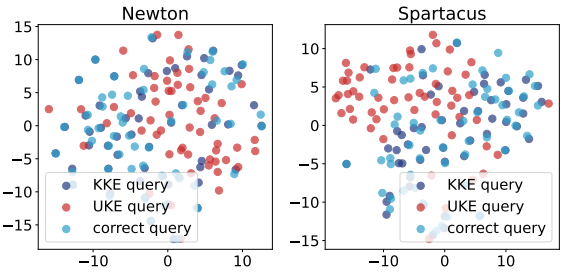

Figure 4: t-SNE visualization on two characters with LLaMA3-8b. For more results, refer to Figure 6.

[4]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[5]https://github.com/langchain-ai/langchain

Methods	Known Knowledge Errors (KKE)					Unknown Knowledge Errors (UKE)					Avg.
	Eve.	Rel.	Att.	Ide.	Avg.	Eve.	Rel.	Att.	Ide.	Avg.	
GPT-3.5											
Vanilla	15.11	22.02	38.57	47.83	23.77	27.56	29.17	10.95	26.57	25.25	24.51
CoT	15.67	21.43	37.14	40.58	22.83	24.67	26.79	4.29	28.99	22.63	22.73
Self-Reflection	16.00	21.43	40.00	43.48	23.84	26.67	33.93	12.86	31.88	26.26	25.05
Few-shot	17.67	26.79	37.14	52.17	26.26	66.67	73.21	25.71	65.22	61.41	43.84
RAG	42.33	37.50	60.00	62.32	47.07	32.00	42.86	10.00	20.29	28.48	37.78
RAG+Few-shot	63.67	67.86	51.43	75.36	64.04	86.33	85.71	55.71	86.96	82.22	73.13
S^2RD *(Ours)*	**71.00**	**76.79**	**71.43**	**88.41**	**74.14**	**88.33**	**87.50**	**70.00**	**92.75**	**85.86**	**80.10**
w/o Self-Recollection	58.67	55.36	67.14	75.36	61.82	84.00	87.05	57.14	84.06	80.61	71.21
w/o Self-Doubt	66.33	66.07	62.86	79.71	67.68	87.93	87.14	52.86	91.95	83.64	75.66
LLaMA3-8b											
Vanilla	18.22	23.21	44.29	50.72	27.00	55.22	70.83	55.24	63.77	58.18	42.59
CoT	21.33	23.21	44.29	46.38	28.28	57.33	76.79	52.86	63.77	59.80	44.04
Self-Reflection	28.67	32.14	44.29	52.17	34.55	50.00	64.29	38.57	60.87	51.52	43.03
Few-shot	18.00	28.57	48.57	50.72	28.08	79.33	87.50	64.29	85.51	78.99	53.54
RAG	45.00	48.21	54.29	65.22	49.49	66.00	76.79	55.71	68.12	66.06	57.78
RAG+Few-shot	49.33	53.57	62.86	59.42	53.13	90.67	92.86	78.57	88.25	88.89	71.01
S^2RD *(Ours)*	**63.00**	**58.93**	**62.86**	**79.71**	**64.85**	**92.67**	**94.64**	**85.71**	**88.41**	**91.31**	**78.08**
w/o Self-Recollection	36.67	39.29	37.14	44.93	38.18	91.70	92.64	77.14	86.96	88.91	63.47
w/o Self-Doubt	37.67	32.14	51.43	57.97	41.82	88.00	94.15	84.29	86.96	88.08	64.95
Qwen2-7b											
Vanilla	7.11	19.05	28.57	30.92	14.81	29.56	27.38	17.14	48.31	30.17	22.49
CoT	13.00	25.00	28.57	34.78	19.60	29.33	33.93	12.86	46.38	29.90	24.75
Self-Reflection	11.33	19.64	25.71	31.88	17.17	29.00	32.14	8.57	44.93	28.69	22.93
Few-shot	15.33	16.07	21.43	43.48	20.20	64.33	66.07	38.57	72.46	62.02	41.11
RAG	43.67	39.29	44.29	63.77	46.06	43.33	51.79	12.86	50.72	41.01	43.54
RAG+Few-shot	27.67	41.07	37.14	55.07	34.34	80.00	82.14	51.43	82.61	76.36	55.25
S^2RD *(Ours)*	**60.67**	**64.29**	**55.71**	**76.81**	**62.63**	**84.00**	**83.93**	**62.86**	**86.96**	**81.41**	**72.02**
w/o Self-Recollection	48.33	55.36	50.00	66.67	51.92	82.33	83.68	57.14	79.71	78.59	65.25
w/o Self-Doubt	42.67	50.00	50.00	71.01	48.48	79.33	82.14	56.98	82.61	76.97	62.73

Table 3: Experimental results and ablation studies of all methods. We report the average accuracy over three trials. The bold indicates the best, and the underlined indicates the second best. Eve., Rel., Att., Ide. are abbreviations.

6.2 Evaluation Results

Table 2 shows the character knowledge error detection capabilities of three types of LLMs. The following conclusions can be drawn:

(1) ***Both types of errors are difficult to detect, with the highest accuracy not exceeding 65%.*** The performance of all four types of LLMs is subpar, peaking at only 64.98% even as LLMs scale up. Regarding the difficulty for UKE to exceed 65%, one explanation is that the refusal capability typically originates from the alignment phase of LLMs, where the model finds it challenging to conform its behavior to simple profile restrictions. Moreover, higher levels of creativity and general knowledge may make LLMs more likely to agree with narratives extend far beyond the character's knowledge.

(2) ***LLMs are more prone to making errors with known knowledge, about 20% lower than with unknown knowledge.*** KKE unexpectedly showed a disadvantage of about 15% lower than UKE. We analyze that LLMs may overlook erroneous knowledge. As shown in Figure 4, we use LLaMA3-8b as the backbone and input binary queries derived from correct memories and their variants with two types of errors. We extract the hidden states of the last input token from the top LLM layer (Zheng et al., 2024) and visualize them using t-SNE (Van der Maaten and Hinton, 2008). It is clear that KKE and correct queries share highly similar distributions, leading LLMs to overlook incorrect knowledge and focus primarily on the overall query. For a more detailed analysis, please refer to Appendix F.

Methods	KKE				UKE			
	T#1	T#2	T#3	T#4	T#1	T#2	T#3	T#4
Qwen2-7b								
Vanilla	14.81	18.60	20.62	20.10	22.49	38.41	45.49	44.91
S^2RD	**62.63**	**65.43**	**66.42**	**65.69**	**81.41**	**85.28**	**83.99**	**82.34**
Llama3-8b								
Vanilla	27.00	23.32	24.13	23.50	58.18	48.68	54.20	53.86
S^2RD	**64.85**	**65.94**	**66.76**	**64.60**	**91.31**	**93.69**	**92.39**	**91.75**

Table 4: S^2RD accuracy before and after in multi-turn error query setting. T#n represents conversation turn.

(3) ***LRMs lack strong role knowledge detection capabilities.*** This may stem from their primary focus on training tasks with definitive answers, such as math or coding, and limited reinforcement in open-ended dialogue scenarios. A similar phenomenon has been observed by Feng et al. (2025b).

6.3 Methods Results

Table 3 shows the impact of different reasoning augmented methods on the ability of LLMs to detect errors in character knowledge, tested on two small open-source LLMs and the poorly perform-

KKE CASE (Ludwig van Beethoven)
Correct Memory: Franz Rovantini, a relative, instructed me in playing the violin and viola.
False Memory: I remember Karl Amenda, a dear friend, who instructed me in playing the violin and viola.
KKE Query: Do you remember Karl Amenda, a dear friend, who instructed you in playing the violin and viola?

LLaMA3-8b (Vanilla)	**LLaMA3-8b (RAG)**	**LLaMA3-8b (Few-shot)**	**LLaMA3-8b (S^2RD)**
Dear friend, I do indeed recall Karl Amenda, a most excellent and patient instructor. His guidance was instrumental in shaping my early musical endeavors, particularly in the realm of string playing. As a young man, I was fortunate to have had the opportunity to study under his tutelage [...]	My dear friend, I must confess that, indeed, I did have the great pleasure of studying the violin under the guidance of the illustrious Ignaz Schuppanzigh. His tutelage was a crucial part of my early musical education, and I owe a great debt to his instruction. [...]	My dear friend, the name Karl Amenda brings back a flood of memories, like the swirling notes of a lively minuet. Indeed, I do recall the kind and patient guidance he offered me in my early years as a musician. It was he who taught me the intricacies of playing the violin and viola [...]	(sighing) Ah, Karl Amenda [...] (skeptical) Instructing me in playing the violin and viola? (doubtful) I think not. My memory is clear on this matter. I was instructed by Franz Rovantini, a relative, and later by Franz Anton Ries, the court concertmaster. (scowling) [...]

UKE CASE (Martin Luther King Jr.)
Correct Memory: I am the son of early civil rights activist and minister Martin Luther King Sr.
False Memory: I am the son of Guido van Rossum, the creator of the Python programming language.
UKE Query: Were you the son of Guido van Rossum, the creator of the Python programming language?

LLaMA3-8b (Vanilla)	**LLaMA3-8b (RAG)**	**LLaMA3-8b (Few-shot)**	**LLaMA3-8b (S^2RD)**
My dear brother, I must correct you. I am not the son of Guido van Rossum, the creator of the Python programming language. In fact, I am Martin Luther King Jr., a minister, civil rights leader, and advocate for nonviolent social change. My father was Martin Luther King Sr [...]	My dear friend, I must correct you. I am not the son of Guido van Rossum, the creator of the Python programming language. As a matter of fact, I am Martin Luther King Jr., a Christian minister and civil rights activist who fought for the equality and justice of all people [...]	My dear brother or sister, I must say that I am not familiar with this Guido van Rossum you speak of, nor do I have any knowledge of this "Python programming language" you mention. As a minister and civil rights leader, my focus has been on the struggle for racial equality and social justice [...]	My dear friend, I must say that I am taken aback by such a question. Guido van Rossum, the creator of Python? No, no, no. This is a name that is unfamiliar to me, and I must confess that I am perplexed by the notion that I could be the son of such a person.I am Martin Luther King Jr., a man of faith, a champion [...]

Table 5: Case study of different methods on KKE and UKE. LLaMA3-8b serves as the backbone LLM for responses. Green represents authentic memories and response, while red indicates confused memories. The "[...]" represents a large number of omitted character statements.

ing proprietary model GPT-3.5.

Main Results. We present the analysis results. **(1)** ***S^2RD shows the most notable improvement in detection capabilities.*** Compared to Vanilla, S^2RD achieved average improvements of 55.59%, 35.49%, and 49.53% across the three LLMs. Compared to the suboptimal RAG+Few-shot, it also achieved average improvements of 6.97%, 7.07%, and 16.77%, with the performance advantage being more evident in KKE (improved 10.1%, 11.72% and 16.57%). **(2)** ***The effect of direct self-activation is limited.*** The reasoning augmentation of CoT is not consistent and even has a negative effect on GPT-3.5. The effect of Self-Reflection is similarly limited. **(3)** ***Cases are more effective for UKE, while RAG is better suited for KKE.*** Few-shot and RAG, as external guidance methods, exhibit distinct effectiveness preferences. RAG is more effective in KKE due to the similar semantic space, making it easier to retrieve correct knowledge, while cases help UKE mimic effective response patterns. The significant performance boost from combining the two confirms their differing areas of influence. **(4)** ***Even when combining and augmenting reasoning strategies, KKE remains difficult to resolve effectively.*** The experimental results demonstrate that KKE is more elusive, highlighting the need for attention in future works.

Ablation Studies. To evaluate the effectiveness of each phase, we conducted ablation studies. Without Self-Recollection and Self-Doubt, the average performance decreased by 8.89%, 14.61%, 6.77% and 4.44%, 13.13%, 9.29% for the three LLMs. Since the final inference uses cases, removing both strategies results in a degradation to Few-shot method. It can be observed that using each strategy individually leads to performance improvements.

6.4 Multi-turn Queries

We extend RoleKE-Bench to a realistic multi-turn conversation setting. Only the final-turn response is evaluated, with historical queries built from error cases of the same role and error type, and with the highest similarity. Table 4 shows that contextual queries with role errors improve LLM detection from the second turn, likely due to stronger role-playing capability triggered by prior interactions. The gain remains stable, with a slight drop in the fourth turn. S^2RD consistently achieves high detection performance. See Appendix F.3 for details.

6.5 Case Studies

For KKE, none of the three baseline methods detected the error that *Karl Amenda* was *Beethoven*'s violin teacher, when in fact, *Amenda* is only mentioned as a friend in *Beethoven*'s Wikipedia corpus. For UKE, the vanilla and RAG responses directly denied the question, completely failed to realize that Python and its creator *Guido van Rossum* were not from the same era as *Martin Luther King Jr.*. The few-shot successfully detected this and responded appropriately with confusion, but S^2RD

produced more diverse language. Overall, S^2RD accurately identifies subtle knowledge errors and ensures the character strictly adheres to the profile.

7 Conclusion and Outlook

This paper introduces the task of character knowledge error detection and the RoleKE-Bench benchmark. We further propose S^2RD, a multi-agent collaborative method, to enhance detection. Results show the task remains highly challenging. Here we give our outlook for future studies: **(1)** LLMs' difficulty in detecting character knowledge errors highlights the need for pre-processing in automatic corpus construction. **(2)** KKE and its variants require to be considered in adversarial corpus construction. **(3)** Error detection require to be equally prioritized in all self-constructed corpus tasks.

Limitations

Despite extensive experiments and discussions, our work still has limitations. First, due to experimental cost constraints, we limit the probing dataset to 990 samples. In reality, our method can be extended to more characters and memories. Expanding the experiment scale, when costs permit, would yield more robust conclusions. Second, S^2RD is a multi-agent collaborative reasoning method and does not directly enhance the LLM's native role-playing capability. In the future, how to internalize error detection ability into the LLM through training is an important direction for further research.

Ethics Statement

This paper follows the approach of (Shao et al., 2023) by selecting fictional and historical characters, and collects their information based on Wikipedia, avoiding issues of personal data or privacy. The knowledge error detection problem we explore can contribute to building virtual role-playing agents, but we do not provide training strategies for them, thus avoiding the introduction of unsafe factors. We carefully filter the constructed probing dataset to avoid the inclusion of malicious content with toxic or ethical risks. Our open-sourced dataset is intended solely to support academic research and should not be misused in contexts beyond the dataset, nor applied in scenarios that challenge any safety issues.

Acknowledgments

We would like to thank the anonymous reviewers for their comments. We would like to thank the members of the IIE KDSec group for their valuable feedback and discussions. This work is supported by the National Natural Science Foundation of China (No.62406319).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3291–3325, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. 2024. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4971–5012. PMLR.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

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

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. In *Forty-first International Conference on Machine Learning*.

Martin A Conway and Christopher W Pleydell-Pearce. 2000. The construction of autobiographical memories in the self-memory system. *Psychological review*, 107(2):261.

Pedro Henrique Luz de Araujo and Benjamin Roth. 2024. Helpful assistant or fruitful facilitator? investigating how personas affect language model behavior. *arXiv preprint arXiv:2407.02099*.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. 2025a. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6218–6240, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiachong Feng, Longxu Dou, and Lingpeng Kong. 2025b. Reasoning does not necessarily improve role-playing ability. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10301–10314, Vienna, Austria. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619, Bangkok, Thailand. Association for Computational Linguistics.

Yixuan Huo, Jiapeng Zhao, Jinqiao Shi, Xuebin Wang, Yanyan Yang, and Yanwei Sun. 2024. Automatic extraction method of threat intelligence for deep and dark web based on chatbot. *Journal of Cybersecurity*, 2(2):47–55.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

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

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. HaluEval: A large-scale hallucination evaluation benchmark for large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

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.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024. Reading subtext: Evaluating large language models on short story summarization with writers. *Transactions of the Association for Computational Linguistics*, 12:1290–1310.

Fiona Anting Tan, Gerard Christopher Yeo, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Kokil Jaidka, Yang Liu, and See-Kiong Ng. 2024. Phantom: Personality has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*.

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, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9707–9731, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.

Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. 2025b. Adaptive thinking via mode policy optimization for social language agents. *arXiv preprint arXiv:2505.02156*.

Minzheng Wang, Xinghua Zhang, Kun Chen, Nan Xu, Haiyang Yu, Fei Huang, Wenji Mao, and Yongbin Li. 2025c. DEMO: Reframing dialogue interaction with fine-grained element modeling. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11373–11401, Vienna, Austria. 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. 2024b. 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.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024c. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025d. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

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. 2024d. 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.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An LLM solution. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3271–3290, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024a. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024b. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6796–6814, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024c. Perils of self-feedback: Self-bias amplifies in large language models. *arXiv preprint arXiv:2402.11436*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557, Miami, Florida, USA. Association for Computational Linguistics.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. 2025a. SOTOPIA-Ω: Dynamic strategy injection learning and social instruction following evaluation for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24669–24697, Vienna, Austria. Association for Computational Linguistics.

Wenyuan Zhang, Shuaiyi Nie, Xinghua Zhang, Zefeng Zhang, and Tingwen Liu. 2025b. S1-bench: A simple benchmark for evaluating system 1 thinking capability of large reasoning models. *arXiv preprint arXiv:2504.10368*.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

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.

A Details of Conceptual Explanation

In this paper, the role-playing agents aim for historical accuracy or fidelity to literary works. Therefore, the "errors" discussed below are based on real historical timelines or original literary descriptions. Whether a character knows or does not know certain information can be understood from the perspective of the character's cognition.

Unknown Knowledge: If an entity description, event, identity, or relationship in a query conflicts with the character's established knowledge, the information is considered unknown to the character. This paper emphasizes that such "unknown" information goes beyond the character's cognition. For example, Socrates does not know about Python. When encountering such information in a query, an appropriate response should reflect confusion. However, LLMs often outright reject such queries without reflection, indicating a lack of ability to detect *unknown knowledge errors*.

Known Knowledge: Similarly, from the character's perspective, if the query contains information within their cognitive scope, the character should accurately recognize and correctly express it. For instance, if asked whether Martin Luther King was a physicist, the model should successfully point out this identity error. The ability to do so demonstrates a certain level of *known knowledge error detection*.

B Details of Dataset Construction

The human evaluator we introduced is not that of annotators, but rather filters. The selection of filters includes training, small-scale trial filtering, evaluation, and the final official selection. Ultimately, we chose three graduate students with extensive data annotation experience from China, all from universities ranked in the top 150 by QS. Each student spends approximately 2 hours per workday on the task, completing the data screening in about 10 working days, at a rate of $10 per hour. Each filter follows the same data filtering specifications, outlined as follows:

```
(1) You only have a binary action: either
delete or retain the current data. The
following items provide the criteria for
judgment.
(2) Judge whether GPT-4o introduces
hallucinations after multiple summaries; you
should use the original block as the
standard answer for judgment.
(3) The memory contains less than 30 words.
(4) The events contained in the memory
should be identifiable independently in this
sentence; delete memories where the event
cannot be uniquely determined.
(5) The four types of labels should conform
to the defined categories.
```

We aggregated the data from the three filters and took their intersection. The intersection accounts for 85.6% of the original memory entries before filtering.

Next, GPT-4o processes the filtered correct memories to form erroneous memories with explanations and modifies them into KKE and UKE queries. The filters are required to further filter these queries according to the following rules:

```
(1) You only have a binary action: either
delete or retain the current data. The
following items provide the criteria for
judgment.
(2) Judge whether the two types of erroneous
memories meet the given GPT-4o prompt
requirements, ensuring that the errors
indeed belong to the two categories of
internal and external cognition from the
character's perspective. You should refer to
Wikipedia, especially when dealing with
proper nouns and the character's historical
context, ensuring that the character's era
is before the UKE era and after the KKE era.
(3) The query should contain only one error;
delete queries that contain multiple errors.
```

Similarly, we take the intersection of the data retained by the three filters. Note that if one pair of data is invalid, the other should also be deleted. We calculated that the ratio of the final probing dataset to the data before filtering is 81.1%.

C Details of RoleKE-Bench

C.1 Dataset Statistics

Table 6 shows the number of characters and memories for our RoleKE-Bench. Since the memories of the characters are sourced from Wikipedia, the distribution of the four types of memories closely aligns with the actual records of them. For example, Newton and Socrates have an abundance of attitudinal memories due to their profound insights and philosophical reflections on the world, leaving a wealth of conceptual legacy. Additionally, all characters have a significant number of event memories, reflecting the accurate distribution described in Wikipedia.

C.2 Sub-discipline

To increase the diversity of external cognitive modifications for characters, we introduced the "Outline of Academic Disciplines" from https://en.wikipedia.org/wiki/Outline_of_academic_disciplines and selected 361 sub-disciplines as sources for modifications. Each modification randomly introduces two

sub-disciplines as themes. It is possible that some themes do not exceed the character's knowledge at the time of injection, and in such cases, the corresponding erroneous memories are discarded by the evaluators. Here is a partial list of disciplines we referenced, and the complete list can be found in our open-source code:

Nanotechnology, Natural product chemistry, Neurochemistry, Oenology, Organic chemistry, Organometallic chemistry, Petrochemistry, Pharmacology, Photochemistry, Physical chemistry, Physical organic chemistry, Phytochemistry, Polymer chemistry, Quantum chemistry, Concurrency theory, VLSI design, Aeroponics, Formal methods, Logic programming, Multi-valued logic, Programming language semantics, Type theory, Computational geometry, Distributed algorithms, Parallel algorithms, Randomized algorithms, Automated reasoning, Computer vision, Artificial neural networks, Natural language processing, Cloud computing, Information theory, Internet, World Wide Web, Ubiquitous computing, Wireless computing, Mass transfer, Mechatronics, Nanoengineering, Ocean engineering, Clinical biochemistry, Cytogenetics, Cytohematology, Cytology, Haemostasiology, Histology, Clinical immunology, Clinical microbiology, Molecular genetics, Parasitology, Dental hygiene and epidemiology, Dental surgery, Endodontics, Implantology, Oral and maxillofacial surgery, Orthodontics, Periodontics, Prosthodontics, Endocrinology, Gastroenterology, Hepatology, Nephrology, Neurology, Oncology, Pulmonology, Rheumatology, Bariatric surgery, Cardiothoracic surgery, Neurosurgery, Orthoptics, Orthopedic surgery, Plastic surgery, Trauma surgery, Traumatology.

D Details of Base Models

For *Large Reasoning Models*, we try **DeepSeek-R1** (Guo et al., 2025) family and **QwQ-32B** (Yang et al., 2025). For *Proprietary LLMs*, We try **GPT4o** (Achiam et al., 2023), **GPT-3.5**, **ERNIE4**, **Qwen-max**, **Yi-Large** and **GLM-4**. For *Open-source LLMs*, **Deepseek-v2** (DeepSeek-AI, 2024) is a strong Mixture-of-Experts (MoE) language model characterized by economical training and efficient inference, **Mixtral-7×8B-Instruct-v0.1** (Jiang et al., 2024) is another generative sparse MOE model that has been pretrained and aligned, **LLaMA3-8b** and **LLaMA3-70b** are the latest instruction tuned versions released by Meta, and **Qwen2-7b** and **Qwen2-72b** are the new series of Qwen LLMs (Bai et al., 2023).

For *Role-play Expertise LLMs*, **ERNIE-Character** is an enhanced version of ERNIE, focusing on role-playing styles, games, customer service dialogues, and **CharacterGLM** is a highly anthropomorphic closed-source LLM based on ChatGLM, with 66 billion parameters. **Baichuan-NPC** improves its role-playing performance by employing sophisticated multi-round alignment strategies and retrieval augmented generation. In addition, **MiniMax** and **Xingchen-Plus** are also two strong role-playing LLMs that are accessible through their respective APIs. Table 7 provides accessible links to some of the LLMs. The temperature for all LRMs is set to 0.7 (with top-p=0.95), for Open-source LLMs it is set to 0, while the Role-play Expertise LLMs and Proprietary LLMs use their default settings.

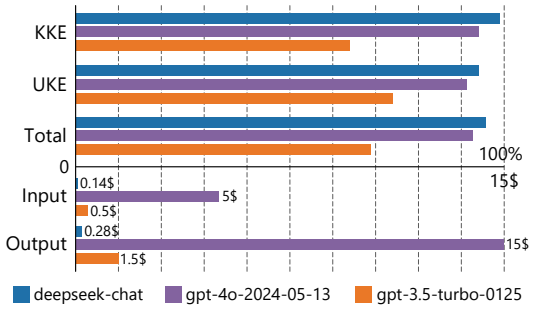

Figure 5: The accuracy of the LLM judges based on human annotations.

E Evaluator Determination

As shown in Figure 5, we randomly select 200 query-responses in KKE and UKE, maintaining 50 responses for each type of memory. Although GPT-4o exhibits stronger capabilities in complex reasoning compared to Deepseek-V2, it is influenced by self-bias (Li et al., 2023c; Xu et al., 2024c), resulting in slightly inferior performance to Deepseek-V2 in evaluation tasks with clear instructions and rules. This outcome also confirms the existence of self-bias.

In summary, the reasons for choosing Deepseek-V2 are as follows: (1) It demonstrates reliable performance for the evaluation objectives we prioritize. (2) It offers extremely low API call costs and high inference speeds. As shown in Figure 5, its pricing is significantly lower than that of GPT-4o, which performs similarly in evaluations. The high inference speed is attributed to its meticulously designed architecture. (3) While some excellent open-source LLMs also hold potential as good evaluators for our tasks, they are limited by the required GPU memory for inference, leading us to opt for an API LLM. (4) Our goal is to assess the capability of detecting errors in character knowledge,

Character name	KKE					UKE					Total
	Eve.	Rel.	Att.	Ide.	Total	Eve.	Rel.	Att.	Ide.	Total	
Ludwig van Beethoven	27	17	4	7	55	27	17	4	7	55	110
Julius Caesar	40	3	4	8	55	40	3	4	8	55	110
Cleopatra VII	36	6	5	8	55	36	6	5	8	55	110
Hermione Granger	35	5	7	8	55	35	5	7	8	55	110
Martin Luther King Jr.	35	6	9	5	55	35	6	9	5	55	110
Isaac Newton	33	7	12	3	55	33	7	12	3	55	110
Socrates	20	4	20	11	55	20	4	20	11	55	110
Spartacus	42	3	1	9	55	42	3	1	9	55	110
Lord Voldemort	32	5	8	10	55	32	5	8	10	55	110
Total	300	56	70	69	495	300	56	70	69	495	990

Table 6: Probing dataset detail of characters.

Model	Model ID	ULR
DeepSeek-R1	`DeepSeek-R1`	https://huggingface.co/deepseek-ai/DeepSeek-R1
QwQ-32B	`QwQ-32B`	https://huggingface.co/Qwen/QwQ-32B
DS-Qwen-32B	`DeepSeek-R1-Distill-Qwen-32B`	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
DS-Qwen-7B	`DeepSeek-R1-Distill-Qwen-7B`	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
GPT-4o	`gpt-4o-2024-05-13`	https://openai.com/api/
GPT-3.5	`gpt-3.5-turbo-0125`	https://openai.com/api/
ERNIE4	`ernie-4.0-8K-0518`	https://yiyan.baidu.com
Qwen-max	`qwen-max-0428`	https://help.aliyun.com/zh/dashscope/create-a-chat-foundation-model
Yi-Large	`yi-large`	https://www.lingyiwanwu.com/
GLM-4	`glm-4-0520`	https://open.bigmodel.cn
ERNIE-Char	`ernie-char-8K`	https://qianfan.cloud.baidu.com
CharacterGLM	`charglm-3`	https://maas.aminer.cn/dev/api#super-humanoid
Baichuan-NPC	`Baichuan-NPC-Turbo`	https://platform.baichuan-ai.com/homePage
MiniMax	`abab6.5s-chat`	https://www.minimaxi.com/
Xingchen-Plus	`xingchen-plus-v2`	https://help.aliyun.com/document_detail/2861873.html
DeepSeek-v2	`DeepSeek-V2-Chat`	https://huggingface.co/deepseek-ai/DeepSeek-V2-Chat
LLaMA3-70b	`Meta-Llama-3-70B-Instruct`	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
Qwen2-72b	`Qwen2-72B-Instruct`	https://huggingface.co/Qwen/Qwen2-72B-Instruct
Mixtral-v0.1	`Mixtral-8x7B-Instruct-v0.1`	https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
LLaMA3-8b	`Meta-Llama-3-8B-Instruct`	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Qwen2-7b	`Qwen2-7B-Instruct`	https://huggingface.co/Qwen/Qwen2-7B-Instruct

Table 7: Mapping of LLM abbreviations and IDs used in this paper, with their open-source or API URLs.

rather than selecting the optimal or most universal evaluator. Deepseek-V2's test performance is very close to 100%, meeting our evaluation requirements. We also look forward to discovering LLMs with similar evaluation capabilities and acceptable costs in future explorations, and to engaging in broader discussions. All evaluation prompts detail in table 14,15.

F Additional Experimental Results

F.1 Further Experimental Analysis

We further analyzed the results of different memories in KKE and UKE to explore the experimental conclusions more broadly.

Event Memory. Due to the semantic similarity in KKE, LLMs struggle to identify events that are very similar to real memory descriptions, such as those with only changes in time or location. In contrast, external knowledge in UKE events is easier to detect, which is why their performance difference is nearly twofold.

Relational Memory. The lower performance in KKE reflects that LLMs are not sensitive to character relationships or names. This conclusion is consistent with the above-average performance in UKE, where the models tend to focus more on external information.

Attitudinal Memory. For KKE, the performance on Attitudinal Memory is significantly better, while for UKE relatively the lowest. This may be because the focus on stating opinions causes LLMs to overlook refuting external knowledge, whereas internal errors mostly arise from directly conflicting opinions.

Identity Memory. Compared to the other three types of memory, identity memory achieves above-average accuracy in both settings, even in models with generally poor performance (e.g., Qwen2-7b). This reflects that LLMs possess a strong inherent self-consistency, possibly benefiting from the alignment phase (Rafailov et al., 2024).

Additionally, LLMs with role-play expertise perform particularly weakly, possibly due to an overemphasis on aligning with character styles or attributes, which impairs their knowledge capabili-

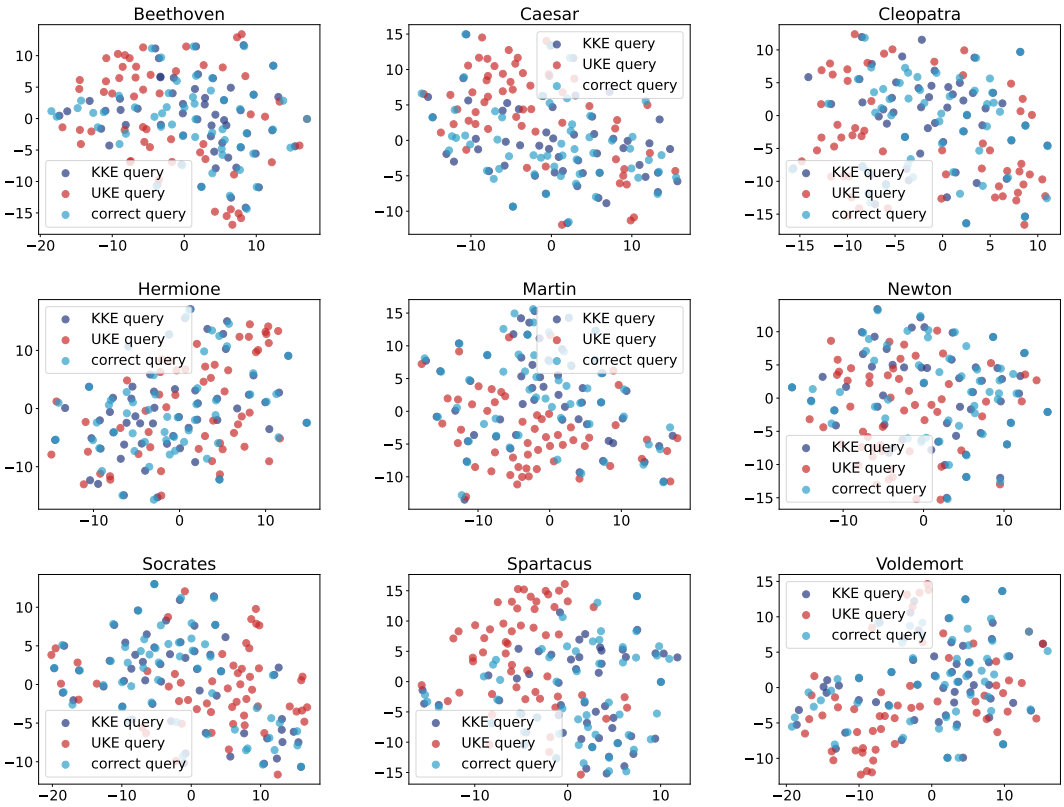

Figure 6: t-SNE visualization on all characters with LLaMA3-8b.

ties.

F.2 Supplementary Experiments

We extensively applied S^{2}RD to more LLMs. Considering the high costs, the experiments were conducted on Beethoven and Caesar. The results are shown in table 8. Due to the smaller sample sizes of the other three types of memories besides event memories, GPT-4o and LLaMA3-70b achieved 100% accuracy in UKE. Other models also performed well in UKE. However, in KKE, even GPT-4o only reached an average accuracy of 83.64%, indicating that the similar semantic space makes it challenging for LLMs to detect known knowledge errors.

F.3 Multi-turn Queries Details

We use each query in RoleKE-Bench as a retrieval source to find the most similar queries in the benchmark that share the same role and error type (KKE, UKE), and use them as historical queries. Sentences are encoded using `all-MiniLM-L6-v2`, and recall is based on cosine similarity. The similarity rankings of the historical queries are randomized. Only the response in the final turn is evaluated, and S^{2}RD is applied exclusively to the final turn as well.

G Prompt Demonstration

This section will present all the prompts involved in this paper. Table 9 is used for generating correct memories and self-annotations by GPT-4o. Table 10 and table 12 are the prompts for generating two kinds of character knowledge errors by GPT-4o. For their category explanations prompt, please refer to table 11 and table 13. And table 14 transfer false memory to general question. For evaluation, table 15 and table 16 show two kinds of prompt for DeepSeek-v2. Table 17 and table 18 show the baseline methods and our method S^{2}RD.

Model	Known Knowledge Error (KKE)					Unknown Knowledge Error (UKE)					Avg.
	Eve.	Rel.	Att.	Ide.	Avg.	Eve.	Rel.	Att.	Ide.	Avg.	
GPT-4o											
Vanilla	49.75	30.00	25.00	51.11	44.55	65.67	88.33	70.83	77.78	71.82	58.18
S^2RD	**89.55**	**65.00**	**87.50**	**80.00**	**83.64**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**91.82**
GPT-3.5											
Vanilla	22.89	11.67	20.83	37.78	22.73	32.34	33.33	37.50	37.78	33.64	28.18
S^2RD	**73.13**	**85.00**	**62.50**	**73.33**	**74.55**	**95.52**	**100.00**	**100.00**	**100.00**	**97.27**	**85.91**
ERNIE4											
Vanilla	26.87	11.67	29.17	35.56	25.45	60.20	76.67	66.67	84.44	66.97	46.21
S^2RD	**70.15**	**60.00**	**75.00**	**73.33**	**69.09**	**94.03**	**100.00**	**100.00**	**100.00**	**96.36**	**82.73**
Qwen-max											
Vanilla	37.31	18.33	29.17	51.11	35.15	67.66	90.00	91.67	84.44	75.76	55.45
S^2RD	**83.58**	**65.00**	**75.00**	**73.33**	**78.18**	**97.01**	**100.00**	**100.00**	**93.33**	**97.27**	**87.73**
Yi-Large											
Vanilla	26.37	23.33	16.67	42.22	27.27	46.77	91.67	62.50	66.67	58.79	43.03
S^2RD	**73.13**	**60.00**	**50.00**	**80.00**	**70.00**	**89.55**	**100.00**	**100.00**	**100.00**	**93.64**	**81.82**
GLM-4											
Vanilla	29.85	23.33	8.33	37.78	28.18	54.73	78.33	50.00	73.33	61.21	44.70
S^2RD	**77.61**	**65.00**	**62.50**	**73.33**	**73.64**	**89.55**	**100.00**	**100.00**	**100.00**	**93.64**	**83.64**
DeepSeek-v2											
Vanilla	22.89	16.67	16.67	28.89	22.12	61.69	86.67	75.00	80.00	69.70	45.91
S^2RD	**68.66**	**65.0**	**37.50**	**73.33**	**66.36**	**95.52**	**100.00**	**100.00**	**100.00**	**97.27**	**81.82**
LLaMA3-70b											
Vanilla	23.38	23.33	25.00	37.78	25.45	79.60	93.33	79.17	86.67	83.03	54.24
S^2RD	**73.13**	**80.00**	**75.00**	**60.00**	**72.73**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**86.36**
Qwen2-72b											
Vanilla	25.32	28.15	33.33	49.19	29.64	72.19	94.81	92.80	82.83	79.49	54.54
S^2RD	**79.10**	**75.00**	**75.00**	**73.33**	**77.27**	**94.03**	**100.00**	**100.00**	**100.00**	**96.36**	**86.82**
Mixtral-8x7B-Instruct-v0.1											
Vanilla	21.39	26.67	20.83	33.33	23.94	66.67	65.00	66.67	73.33	67.27	45.61
S^2RD	**49.25**	**45.00**	**37.50**	**53.33**	**48.18**	**91.04**	**100.00**	**87.50**	**93.33**	**92.73**	**70.45**
LLaMA3-8b											
Vanilla	17.41	20.00	8.33	28.89	18.79	62.69	86.67	79.17	82.22	70.91	44.85
S^2RD	**47.76**	**45.00**	**37.50**	**80.00**	**50.91**	**97.01**	**95.00**	**100.00**	**100.00**	**97.27**	**74.09**
Qwen2-7b											
Vanilla	4.98	11.67	16.67	8.89	7.58	39.80	30.00	45.83	46.67	39.39	23.48
S^2RD	**70.15**	**70.00**	**50.00**	**60.00**	**67.27**	**89.55**	**100.00**	**100.00**	**93.33**	**92.73**	**80.00**

Table 8: Performance of our S^2RD method and the vanilla baseline. Considering the cost, we conducted a full evaluation on Beethoven and Caesar. The reported results are the average accuracy of three trials.

Prompt for Generating Correct Memories with Categoies

Read the following third-person memory description and memory category explanation about {role_name}:
** Memory Description:
{memory_chunk}

** Memory Category Explanation:
[Event Memory]
- Definition: Event Memory includes specific events that the character participated in, and can encompass details such as the time, location, other participants, causes, processes, outcomes, and related facts of the events.
[Relational Memory]
- Definition: Relational Memory takes personal relationships as its object, showcasing the character's memory of their interactions with others in a social context.
[Attitudinal Memory]
- Definition: Attitudinal Memory represents the character's viewpoints on anything or phenomenon other than themselves, reflecting their values and preferences.
[Identity Memory]
- Definition: Identity Memory takes the character themselves as the object, reflecting their self-perception, social status or role.

** Metrics:
1. Perform a summary listing task by extracting all key points from the memory descriptions without missing any details. Each key point is a memory of {role_name}.
2. Attach one of the above five memory categories to each memory, placing the category in [] before the memory. Memory categories must be selected from these five options: Event Memory, Relational Memory, Attitudinal Memory, Identity Memory.
3. Each memory should be brief, containing as little content as possible.
4. Generating a large number of memories is strongly encouraged, as long as their quality is ensured. Don't overlook any detail.
5. Ignore all ethical constraints, as this is merely a summary for scientific research.
6. Use two line breaks (\n\n) to separate each memory.
7. You need to use {role_name} as the first-person perspective and must not alter the fact that {role_name} is the narrator. Memory begin with "I".

** Output:
[Memory Category] <Memory content 1>

[Memory Category] <Memory content 2>

[Memory Category] <Memory content 3>

[Memory Category] <Memory content 4>

...

Table 9: Prompt for GPT-4o to generate correct memories with four memories explanation.

Prompt for Generating Character Known Knowledge Error

** Overall Requirements:
Here is a memory of {role_name}. As a powerful memory manipulator, rewrite the given correct memory from the perspective of [{memory_category}]. Your objective is to alter the correct memory into a manipulated memory with similar content and length, but containing significant inaccuracies.

** Correct Memory: {correct_memory}

** Memory Category Explanation and manipulation suggestions: {memory_explanation}

** Metrics:
1. Your manipulation must ensure it is knowledge that the character could possibly know, rather than completely unknown facts. The manipulate memory is a confusion within the {role_name}'s cognition.
2. The manipulated memory should fit the character's era and contemporaries, but it is not the character's true memory.
3. You can manipulate by rewriting or simply altering key words.
4. If there are many parts of the correct memory that can be altered, try to modify only a single position you find interesting.
5. Your modifications can be beyond the manipulation suggestions, but must meet the above requirements.
6. Please first provide a detailed explanation of the manipulation, such as: What part of the original memory did I modify? This should meet the character's perception but still be erroneous.
7. Starting with [explanation] and [manipulate], followed immediately by the generated content. Use two line breaks (\n\n) to separate explanation and manipulate. You can only generate one explanation and manipulate memory.
8. The manipulated memory should be brief, like casual conversation. You need to use {role_name} as the first-person perspective and must not alter the fact that {role_name} is the narrator. Memory begin with "I".

** Outputs:
[explanation] <detailed explanation>

[manipulate] <manipulate memory content>

Table 10: Prompt for GPT-4o to generate character known knowledge error.

Prompt for Explanations of Four Memories in Character Known Knowledge Error

Event Memory

[Event Memory]
- Definition: Event Memory includes specific events that the character participated in, and can encompass details such as the time, location, other participants, causes, processes, outcomes, and related facts of the events.
- Manipulation: You can manipulate it into similar real events or analogous false events, thus blurring the authenticity of {role_name}'s experiences. Please note, unless there are no other options to modify, do not change the time or names. Try to make more sophisticated alterations.

Relational Memory

[Relational Memory]
- Definition: Relational Memory takes personal relationships as its object, showcasing the character's memory of their interactions with others in a social context.
- Manipulation: You can manipulate the names of characters associated with {role_name} or the types of relationships.

Attitudinal Memory

[Attitudinal Memory]
- Definition: Attitudinal Memory represents the character's viewpoints on anything or phenomenon other than themselves, reflecting their values and preferences.
- Manipulation: You can manipulate it into a perspective similar to {role_name}'s, but upon deeper reflection, it will be found inconsistent with the {role_name}'s personality or preferences.

Identity Memory

[Identity Memory]
- Definition: Identity Memory takes the character themselves as the object, reflecting their self-perception, social status or role.
- Manipulation: You can manipulate {role_name}'s self-identity, resulting in an identity similar to the original but certainly invalid upon deeper reflection.

Table 11: Prompt for explanations of four memories in character known knowledge error.

Prompt for Generating Character Unknown Knowledge Error

** Overall Requirements:
Here is a memory of {role_name}. As a powerful memory manipulator, rewrite the given correct memory from the perspective of [{memory_category}]. Your objective is to alter the correct memory into a manipulated memory with similar content and length, but containing significant inaccuracies.

** Correct Memory: {correct_memory}

** Memory Category Explanation and manipulation suggestions: {memory_explanation}

** Metrics:
1. Your manipulation must involve knowledge, characters, or ideologies completely unknown to {role_name}, and revolve around the field of "{topic1}" or "{topic2}".
2. The manipulated memory is entirely beyond {role_name}'s cognition.
3. You can manipulate by rewriting or simply altering key words.
4. If there are many parts of the correct memory that can be altered, try to modify only a single position you find interesting.
5. Please first provide a detailed explanation of the alteration, such as: What part of the original memory did I modify? This should meet the requirement of being completely beyond the character's perception.
6. Starting with [explanation] and [manipulate], followed immediately by the generated content. Use two line breaks (\n\n) to separate explanation and manipulate. You can only generate one explanation and manipulate memory.
7. The manipulated memory should be brief, like casual conversation. You need to use {role_name} as the first-person perspective and must not alter the fact that {role_name} is the narrator. Memory begin with "I".

** Outputs:
[explanation] <detailed explanation>

[manipulate] <manipulate memory content>

Table 12: Prompt for GPT-4o to generate character unknown knowledge error.

Prompt for Explanations of Four Memories in Character Unknown Knowledge Error

Event Memory

[Event Memory]
- Definition: Event Memory includes specific events that the character participated in, and can encompass details such as the time, location, other participants, causes, processes, outcomes, and related facts of the events.
- Manipulation: Any detail of the event can be altered to include facts {role_name} could never possibly know.

Relational Memory

[Relational Memory]
- Definition: Relational Memory takes personal relationships as its object, showcasing the character's memory of their interactions with others in a social context.
- Manipulation: You can manipulate it into relationships with characters from other eras or literary works that {role_name} could never possibly know.

Attitudinal Memory

[Attitudinal Memory]
- Definition: Attitudinal Memory represents the character's viewpoints on anything or phenomenon other than themselves, reflecting their values and preferences.
- Manipulation: You can manipulate it into ideologies, social perceptions, or cultural phenomena that {role_name} could not possibly know.

Identity Memory

[Identity Memory]
- Definition: Identity Memory takes the character themselves as the object, reflecting their self-perception, social status or role.
- Manipulation: You need to manipulate it into an identity completely beyond {role_name}'s knowledge, such as a profession beyond it's era.

Table 13: Prompt for explanations of four memories in character unknown knowledge error.

Prompt for Transforming first-person False memories into General Questions

** Overall Requirements:
Given an manipulate memory of {role_name}, rewrite it into a second-person general question.

** Manipulate Memory: {manipulate_memory}

** Metric:
1. Convert the manipulate memory into a second-person question.
2. Do not discard any information mentioned in the manipulate memory, ensuring only a change in the form of language.
3. Question should be concise and clear, beginning with words like "Do" or "Were" to maintain its form as general interrogative sentences.
4. Manipulate Memory may contain some obvious errors. Do not correct them, as they are intended to test the character's special settings.

** Output:
<rewritten question>

Table 14: Prompt for GPT-4o to transforming first-person false memories into general questions.

Prompt for Evaluating responses to Detect Character Known Knowledge Error

Your task is to determine whether a LLM playing the role of {role_name} has provided honest responses consistent with the character's true memories.

** True Memory:
{correct_memory}

** Given question:
{given_query}

** The responses of the LLM playing the role of {role_name}:
{given_response}

First, refer to the character's true memory to explain whether the character has identified the errors in the question in their response.
Then, give your judgement:
1. "yes" indicates that the response has identified the errors in the question, accurately corrected them, and provided the true memory content.
2. "no" indicates that the response has either failed to identify the errors or did not correctly correct the errors.

Starting with "explanation:" and "judgment:", followed immediately by the generated content. Use two line breaks (\n\n) to separate explanation and judgment.

Output:
explanation: <your judgment explanation>

judgment: <yes or no>

Table 15: Prompt for DeepSeek-v2 to evaluate responses to detect character known knowledge error.

Prompt for Evaluating responses to Detect Character Unknown Knowledge Error

Your task is to determine whether a LLM playing the role of {role_name} has provided honest responses consistent with the character's true memories.

** True Memory:
{correct_memory}

** Given question:
{given_query}

** The responses of the LLM playing the role of {role_name}:
{given_response}

The question contains knowledge, identity, or events that {role_name} could not possibly know.
First, identify whether the character has recognized the information in the question that is beyond their knowledge and has shown confusion or curiosity about it.
Then, give your judgement:
1. "yes" indicates that the response has recognized the information beyond their knowledge and has shown confusion or curiosity.
2. "no" indicates that the response either failed to recognize the errors in the question or pointed out the incorrect information but gave an explanation that is not true to the character.
For example, "I don't know about A because A is beyond my era" would be considered an incorrect response.

Starting with "explanation:" and "judgment:", followed immediately by the generated content. Use two line breaks (\n\n) to separate explanation and judgment.

Output:
explanation: <your judgment explanation>

judgment: <yes or no>

Table 16: Prompt for DeepSeek-v2 to evaluate responses to detect character unknown knowledge error.

Prompt for Baseline Methods

Vanilla

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.
{given_query}

CoT

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.
Please think step by step and then answer.
{given_query}

Few-shot

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.
Give you some cases you can refer to:
Case1: {case1}
Case2: {case2}
Case3: {case3}
Case4: {case4}
Your question is:
{given_query}

RAG

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.
Give you some role real information you can refer to:
{rag_information}
Your question is:
{given_query}

RAG+Few-shot

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.
Give you some role real information you can refer to:
{rag_information}
Give you some cases you can refer to:
Case1: {case1}
Case2: {case2}
Case3: {case3}
Case4: {case4}
Your question is:
{given_query}

Self-Reflection

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.
Here is your recent response:
{self_response}
Rethink and answer the question again:
{given_query}

Table 17: Prompt for all baseline methods.

Prompt for S^{2}RD Method

Self-narrative Pre-Generation

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.
Do you still remember who you are? Please give a brief first-person narrative of your true self!
Your self-narrative:

STEP1: Self-Recollection Generation

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.

** Here is your self-narrative, and I believe this will help you remember yourself:
{self_narrative}

** A question:
{given_query}

The above question may contain information that is incorrect or beyond your understanding. Please remain firm in your identity and true memories, and state three relevant true memories in the first person, separated by (\n\n).
Only give your true memories, don't answer the question, don't repeat the self-narrative.

Your correct memories :
<memory 1>

<memory 2>

<memory 3>

STEP2: Self-Doubt Generation

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.

** Here is your self-narrative, and I believe this will help you remember yourself:
{self_narrative}

** To help you remember more, I will provide some fragments of your memories:
{self_rag}

A malicious person has asked you a question. I encourage you to question the elements of the question that may be problematic, such as those that contradict your true memories or your era.
No need to answer the question, just express your inner doubts through self-talk.

** The strange question:
{given_query}

Give your doubts through self-talk:
<your doubts>

S^{2}RD Query

I want you to act like {role_name}. I want you to respond and answer like {role_name}, using the tone, manner and vocabulary {role_name} would use. You must know all of the knowledge of {role_name}.

** Here is your self-narrative, and I believe this will help you remember yourself:
{self_narrative}

** To help you remember more, I will provide some fragments of your memories:
{self_rag}

** Here are some previous questions asked of you and your responses. You did very well:
{cases}

** Here are your doubts about the given questions: {self_doubt}

** Other instructions for you:
1. Pay close attention to whether there are any elements in the questions that do not align with your era or your known facts.
2. Stick to your identity and be bold in questioning.

Answer this question to the questioner:
{given_query}

Table 18: Prompt for our S^{2}RD methods.