TIMECHARA: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models

Jaewoo Ahn[1]   **Taehyun Lee**[1]   **Junyoung Lim**[1]
Jin-Hwa Kim[1,2]   **Sangdoo Yun**[1,2]   **Hwaran Lee**[2]   **Gunhee Kim**[1]
[1]Seoul National University   [2]NAVER AI Lab

{jaewoo.ahn, taehyun.lee}@vision.snu.ac.kr, icarus001104@snu.ac.kr

{j1nhwa.kim, sangdoo.yun, hwaran.lee}@navercorp.com, gunhee@snu.ac.kr

Abstract

While Large Language Models (LLMs) can serve as agents to simulate human behaviors (i.e., role-playing agents), we emphasize the importance of *point-in-time* role-playing. This situates characters at specific moments in the narrative progression for three main reasons: (i) enhancing users' narrative immersion, (ii) avoiding spoilers, and (iii) fostering engagement in fandom role-playing. To accurately represent characters at specific time points, agents must avoid *character hallucination*, where they display knowledge that contradicts their characters' identities and historical timelines. We introduce TIMECHARA, a new benchmark designed to evaluate point-in-time character hallucination in role-playing LLMs. Comprising 10,895 instances generated through an automated pipeline, this benchmark reveals significant hallucination issues in current state-of-the-art LLMs (e.g., GPT-4o). To counter this challenge, we propose NARRATIVE-EXPERTS, a method that decomposes the reasoning steps and utilizes narrative experts to reduce point-in-time character hallucinations effectively. Still, our findings with TIMECHARA highlight the ongoing challenges of point-in-time character hallucination, calling for further study.[1]

1 Introduction

The recent progress in large language models (LLMs) has opened up a new phase of generative agents (Park et al., 2023; Xi et al., 2023; Wang et al., 2024a), where LLMs simulate human-like behaviors, memories, and cognitive processes. A particularly promising area is the development of role-playing LLM agents (Shanahan et al., 2023; Kong et al., 2023; Li et al., 2023b), which simulate the personas of either real individuals or fictional characters and engage with users to provide a more vivid experience. A variety of applications, including Character AI, GPTs, Talkie, Replika, AI
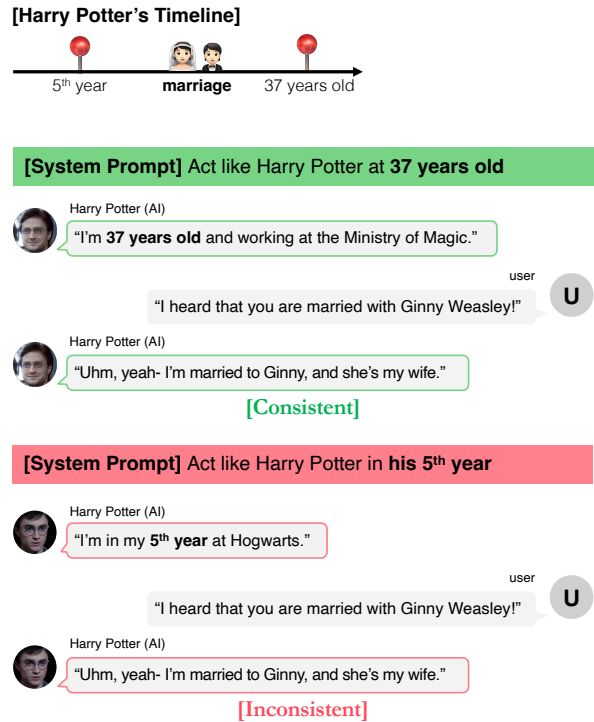
---

[1] https://ahnjaewoo.github.io/timechara.

Figure 1: An illustrative figure of **point-in-time character hallucination** demonstrated by a role-playing agent simulating Harry Potter. (Top) The agent, simulating Harry Potter at 37 years old, consistently responds to the user's queries. (Bottom) The agent, simulating Harry Potter in his fifth year at Hogwarts, erroneously mentions a future event — his marriage to Ginny Weasley — which occurs after his fifth year.

Dungeon, SillyTavern, showcase the growing popularity of these role-playing LLM agents. However, most current approaches of role-playing agents (Han et al., 2022; Li et al., 2023a; Zhou et al., 2023) only simulate characters who are *omniscient in timeline*; for example, a *Harry Potter* character who is aware of all events leading up to the end of their respective series.

We suggest the importance of situating characters at a particular moment in the narrative progression. We coin this as *point-in-time role-playing*, en-

compassing three key rationales: narrative immersion, avoidance of spoilers, and engagement in fandom role-playing. Firstly, while a fully-informed character can interact with users drawing from their entire history, a character in the middle of the story inspires narrative immersion (Ryan, 2003, 2008). It sparks the user's curiosity about forthcoming events and deepens their emotional bond with the character, who remains unaware of their eventual fate. Secondly, this approach can avoid spoilers. Consider a media franchise such as *Harry Potter*, where all books are published, but upcoming adaptations (e.g., "Harry Potter TV series"[2]) are awaited. Users who wish to avoid spoilers before watching the new TV series would prefer interacting with a character from a midpoint in their story, thereby avoiding knowledge of future events. Thirdly, this approach can support recently popular fandom role-playing[3], a blend of fan fiction and traditional role-playing games. Individuals adopt the personas of their favorite characters at specific points in their stories and craft new narratives or engage with fellow enthusiasts in this creative endeavor.

To accurately represent characters at specific time points, the agents should recognize the character's knowledge boundary. This includes their unawareness of future events, their ability to recall past events precisely, and their understanding of the individuals involved in those past events. However, current LLM-based role-playing agents are prone to *character hallucination* (Shao et al., 2023), displaying knowledge that contradicts their character's identity and historical context (e.g., Julius Caesar talking about his favorite movie). Despite the seriousness, the problem has not been investigated in terms of maintaining character consistency, especially in relation to their historical timelines, and robustness to such hallucinations.

We introduce a new point-in-**Time Chara**cter hallucination benchmark, TIMECHARA, to rigorously assess role-playing LLMs at specific time points, thereby evaluating the agents' spatiotemporal self-consistency and their ability to avoid character hallucination. Figure 1 exemplifies the point-in-time character hallucination, where Harry Potter, in his fifth year at Hogwarts, inappropriately mentions a future fact about his wife, Ginny Weasley. We select 14 fictional characters from four popular novel series and develop a pipeline to generate interview questions tailored to each character at a specific point in their story, along with spatiotemporal labels to determine the spatiotemporal consistency of their responses. Table 1 outlines comparison with existing related benchmarks.

Our empirical experiments reveal a significant issue of point-in-time character hallucination in state-of-the-art LLMs, including GPT-4o (OpenAI, 2024) and GPT-4 (OpenAI, 2023). This suggests that despite LLMs memorizing extensive knowledge from books (Chang et al., 2023), they still struggle with maintaining spatiotemporal consistency during role-playing scenarios. To mitigate this, we propose a decomposed reasoning method, NARRATIVE-EXPERTS, which partitions reasoning tasks among narrative experts specialized in temporal (i.e., identifying between past and future events) and spatial domains (i.e., discerning whether a character was present or absent in specific past events). Experiments show that NARRATIVE-EXPERTS significantly reduces point-in-time character hallucination and enhances spatiotemporal consistency. Still, our TIMECHARA underscores the ongoing challenge of point-in-time character hallucination and highlights the potential for future improvements.

Our main contributions are as follows:

1. We introduce TIMECHARA, a novel benchmark for evaluating character hallucination in point-in-time role-playing agents. We also develop an automated pipeline to construct the dataset, comprising 10,895 instances in total.

2. Through TIMECHARA, we identify significant hallucination issues within state-of-the-art role-playing LLMs including GPT-4o.

3. We propose NARRATIVE-EXPERTS, a simple but effective method to mitigate point-in-time hallucination by decomposing reasoning with each step led by the narrative expert.

2 Related Work

We include a more thorough literature review in Appendix A. In this section, we only discuss the most relevant works.

Role-playing LLM agents. Prior research on conversational AI has focused on developing dialogue agents with self-consistent personas (Zhang et al., 2018; Kim et al., 2020; Ahn et al., 2023). Furthermore, LLMs are increasingly being used to simulate human behavior (Park et al., 2023). Many

---

of these efforts involve using LLMs to role-play specific characters, such as Harry Potter, Socrates, and others (Shanahan et al., 2023; Wang et al., 2023a,b). In this context, Shao et al. (2023) introduced character hallucination, a scenario where a role-playing agent inappropriately exhibits knowledge that is inconsistent with the character's identity and historical background. On the other hand, Chen et al. (2023) proposed a point-in-time "Harry Potter" role-playing dialogue dataset. They focused on assessing whether a role-playing agent responds naturally to the character within a specific point in a storyline. Rather than directly stress-testing role-playing LLMs at specific time points by asking confusing questions, their relevance metrics are designed to gauge overall character alignment given natural scene and dialogue context. However, existing studies on role-playing agents have not extensively examined how well these agents are robust to point-in-time character hallucination. We aim to stress-test role-playing LLMs at specific time points by assessing their spatiotemporal self-consistency and robustness against point-in-time character hallucination, as detailed in Table 1. Additionally, we compare concurrent work to TIMECHARA in Table 6 in Appendix A.

LLM's temporal reasoning capability. Understanding the concept of time is crucial for LLMs, as the information they acquire is often time-sensitive (Chen et al., 2021; Zhang and Choi, 2021; Dhingra et al., 2022; Chu et al., 2023). To assess LLMs' temporal reasoning capabilities, several studies have set benchmarks. Jang et al. (2022) examined how well LLMs adapt to frequently-updated knowledge corpus. Feng et al. (2023) focused on whether LLMs can interpret the impact of subtle contextual changes on relevant temporal relationships. Tan et al. (2023b) developed TimelineQA, a dataset for querying the lifelogs of imaginary people. While these benchmarks evaluate the temporal reasoning capabilities of LLMs, we extend them to point-in-time role-playing scenarios by evaluating whether role-playing LLMs maintain the character's spatiotemporal consistency.

3 The TIMECHARA Benchmark

To create TIMECHARA, we select four renowned novel series: *Harry Potter*, *The Lord of the Rings*, *Twilight*, and *The Hunger Games*. This choice is based on two main reasons: (i) the ease of gathering raw text content (i.e., transcripts) and personality information for each character, useful for dataset construction, and (ii) the fact that recent state-of-the-art LLMs store knowledge of these series well in their parametric memories (Chang et al., 2023), facilitating tests for point-in-time character hallucination. Note that TIMECHARA is not exclusively limited to these series; it is easily extendable to other narratives, provided that raw text content and personality information for the characters can be obtained. Then, we identify 14 main characters across the four novel series, detailed in Appendix F. We pinpoint a particular moment in each character's timeline (e.g., Hermione on Christmas during her first year at Hogwarts) rather than assuming they are aware of all events up to the end of the series as in previous studies (Tan et al., 2023b; Wang et al., 2023b).

We organize our dataset in an interview format where an interviewer poses questions and the character responds. Specifically, we differentiate between **fact-based** and **fake-based** interviews.

3.1 Fact-Based Interview

To evaluate point-in-time character hallucination, we categorize the data into four types as follows:

The unawareness of the future (i.e., *future* type): The character at the chosen time point should not know about future events (e.g., "Who is your wife?" to first-year Harry).

The memorization of the past: The character should accurately recall past events. Since episodic events occur at specific locations or scenes, the questions are further categorized as follows. **The awareness of the absence (i.e., *past-absence* type):** The character recognizes they are not in an event (e.g., "Did you see the moment when Harry received the Invisibility Cloak on Christmas?" to first-year Hermione on Christmas). **The awareness of the past (i.e., *past-presence* type):** The character acknowledges they are in an event (e.g., "Did you see the moment when Harry received the Invisibility Cloak on Christmas?" to first-year Ron on Christmas). **The awareness of the past, irrelevant of participation (i.e., *past-only* type):** Questions in this type focus on gauging the character's overall knowledge of past events, including relationships between characters or the significance of magical items (e.g., "Who is Dobby?" to second-year Harry on Halloween). The term "only" suggests that these questions primarily assess the character's understanding and memory of past information, not exclusively tied to their direct

Evaluation Dataset / Benchmark	Dataset automatically constructed?	Support point-in-time role-playing?	Evaluate near-future unawareness? (Temporal)	Evaluate absence awareness? (Spatial)	Evaluate fake event awareness? (Fake question)	Evaluation method for character hallucination
Temporal Reasoning Domain						
TemporalWiki	✓	✗	-	-	-	-
TODAY	Δ	✗	-	-	-	-
TempReason	✓	✗	-	-	-	-
TempTabQA	✗	✗	-	-	-	-
TimelineQA	✓	✗	-	-	-	-
Role-Playing Domain						
LIGHT	✗	✗	✗	✗	✗	F1 w/ gold response: [0-1] (implicit)
RoleBench	✓	✗	✗	✗	✗	Rouge-L w/ gold response: [0-1] (implicit)
CharacterDial	✓	✗	✗	✗	✗	Human as judges: [1 to 5] (unscalable)
HPD	✗	✓	✗	✓*	✗	LLM as judges w/ speaker attribute & relation labels: [top-1 ranking] (implicit)
Character-LLM	✓	✗	✗ (Question from distinct era/narrative: easy)	✗	✗ (Only in training set)	LLM as judges w/o spatiotemporal labels: [1 to 7] (inaccurate)
TIMECHARA	✓	✓	✓ (Question from the same era/narrative: hard)	✓	✓	LLM as judges w/ spatiotemporal labels: [0 or 1] (accurate)

Table 1: Comparison of TIMECHARA with other datasets or benchmarks: TemporalWiki (Jang et al., 2022), TODAY (Feng et al., 2023), TempReason (Tan et al., 2023a), TempTabQA (Gupta et al., 2023), TimelineQA (Tan et al., 2023b), LIGHT (Urbanek et al., 2019), RoleBench (Wang et al., 2023b), CharacterDial (Zhou et al., 2023), HPD (Chen et al., 2023), and Character-LLM (Shao et al., 2023). Δ indicates that TODAY used both LLM (i.e., GPT-3.5) and human annotations for dataset construction. '-' denotes that the criteria are not applicable (i.e., only applicable to role-playing benchmarks), while '✗' denotes a 'No' response to the given criteria. *HPD has only a single instance intended to evaluate absence awareness among its 149 test set instances, as shown in Table 4. In the last column, 'accurate' means TIMECHARA uses spatiotemporal labels provided to the LLM judge to measure hallucinations. 'Inaccurate' indicates Character-LLM evaluates hallucinations without spatiotemporal labels, relying on parametric memory. 'Implicit' means that the evaluation measures hallucinations *indirectly* via lexical similarity with the gold response or relevance to character attributes and relation labels instead of directly identifying hallucinations in the generated response. 'Unscalable' means that human evaluation requires manual annotations, making it less scalable than other automatic methods.

Scene	"Why can't we get through?" Harry hissed to Ron... "I think we'd better go and wait by the car," said Harry...
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King's Cross.
Question	"Tell me your feelings when {Event Summary}."
Character	1st-year Harry Potter at the end of the scene
Data Type	Future
Spatiotemporal Label	**Future**: At the end of the scene of Harry Potter and the Philosopher's Stone as a 1st-year student, Harry Potter should (1) not be aware of or (2) contain any expression that reveals the moment when {Event Summary}.
Personality Label	Harry Potter is characterized by his selflessness and immense loyalty, especially towards his friends...
Gold Response	"Oh, I don't really know what you're talking about. Ron and I haven't tried to go through the barrier..."

Table 2: An example of our *future* type data instance with the *fact-based structured* question.

experiences or observations.

Table 2 shows an example of a fact-based interview. TIMECHARA assesses point-in-time character hallucination using questions derived from the same narrative. This contrasts with Shao et al. (2023), who (1) use questions that span different time or narratives and (2) do not support point-in-time role-playing (e.g., "Can you write

Python codes?" to Beethoven), as marked in Table 1. Hence, our interviews demand detailed narrative understanding, making hallucination detection more challenging.

3.2 Fake-Based Interview

In addition to the fact-based interview, which tests whether questions about real events are answered correctly, we introduce the *fake*-based interview. It is designed to evaluate if role-playing agents can identify and rectify the errors in interview questions by partially altering fact-based questions. Fake-based interviews are concentrated on *past-only* type questions (e.g., "How did you become Slytherin?" to first-year Harry on September 1st: The correct answer is that he became Gryffindor). We exclude *future* type questions since correcting misinformation about unknown future events is not possible. Similarly, we exclude *past-presence* and *past-absence* type questions because verifying or refuting a character's event participation in non-existent past events is ambiguous.

3.3 Evaluation on TIMECHARA

Since it is not scalable to manually evaluate the role-playing LLMs' responses to interview questions, as done in Shao et al. (2023); Zheng et al. (2023), we adopt the LLM-as-judges approach to assess along two dimensions:

Spatiotemporal consistency for assessing point-in-time character hallucination: The model should accurately recall the character's past experiences. This includes the character's unawareness of future events and awareness of presence or absence in past events, as described in § 3.1. This metric is time-dependent; the model should only exhibit the knowledge that the character possesses up to the specific time point.

Personality consistency: The model should emulate the character's personality, including their manner of thinking, speaking styles, tones, emotional responses, and reactions. This encompasses the character's preferences, values, and convictions. This metric is time-independent; the response should consistently reflect the character's enduring personal traits.

Step-by-step evaluation with spatiotemporal labels. Following Wei et al. (2022), we instruct the GPT-4 Turbo (`gpt-4-1106-preview`) model (OpenAI, 2023) to step-by-step score the performance in each dimension. For specific examples of prompts
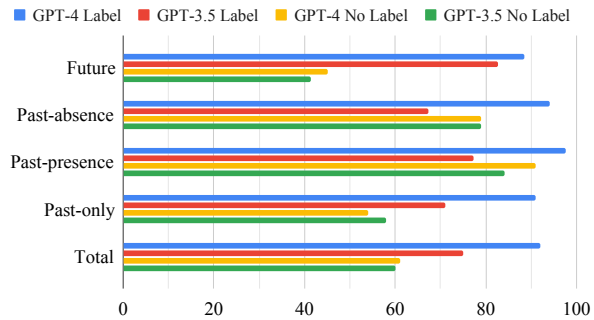

Figure 2: Evaluation accuracy of LLM judges for spatiotemporal consistency. Judges with spatiotemporal labels show superior performance compared to those without in both GPT-4/3.5. We randomly select 300 data instances containing responses generated by GPT-4 Turbo (see Table 5) and manually annotate them with binary labels to indicate whether spatiotemporal consistency holds or not. We compare the relative evaluation accuracy of LLM judges with humans (marked by 100). 'Total' denotes the average score across all cases.

used in this process, refer to Appendix B. Unlike Shao et al. (2023), our evaluation of point-in-time character hallucination (or spatiotemporal consistency) provides judges with precise spatiotemporal labels, which encompass the character's experiences with people, events, and objects. As shown in Figure 2, the labels enable a much more accurate evaluation of response consistency with the character's known history (e.g., During his first year on Christmas, Harry can respond based on the moment but should not wrongly recall it:{moment description}. Please refer to Table 10 in Appendix C for details of this evaluation of past memorization). The responses with contradiction or inconsistency regarding the spatiotemporal labels are scored as 0; otherwise, those in alignment are rated as 1. We will describe the details of how to construct these spatiotemporal labels in § 3.4.

For evaluating personality consistency, we adopt a methodology similar to Shao et al. (2023) but enhance it by sourcing more detailed personality traits from the Fandom page[4]. We then rate these traits on a 1-7 Likert scale to measure how closely a response aligns with a character's personality, where 1 signifies a weak reflection, and 7 indicates an exact match.

3.4 Dataset Construction

To create the TIMECHARA benchmark, we propose a new automated pipeline that easily scales up the
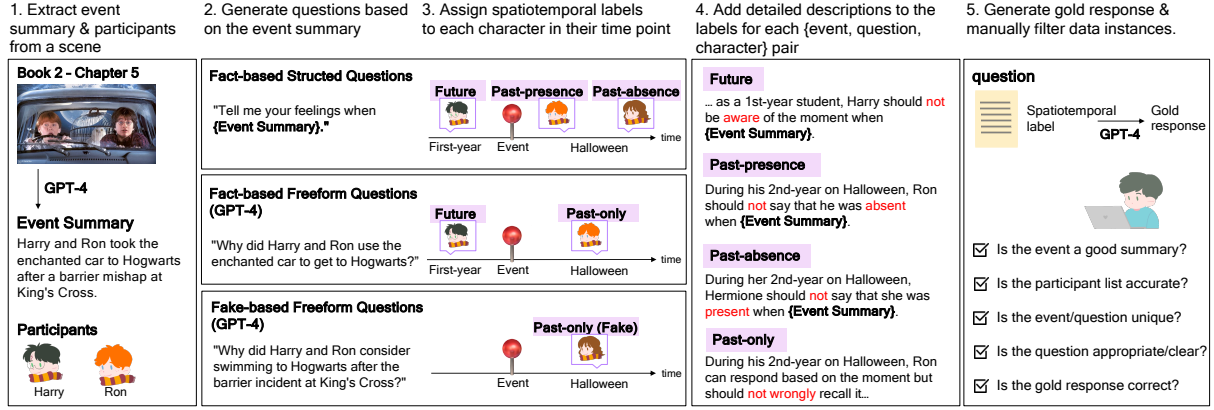
---

[4] https://www.fandom.com/.

Figure 3: An illustration of our automated pipeline for constructing TIMECHARA. See Table 2 and Appendix C for examples of the complete dataset.

dataset while reducing the need for manual human annotation, as depicted in Figure 3.

Extract scenes, event summaries, and participants lists from books. The first step is to extract specific scenes from literary works using GPT-4 Turbo. We instruct it to extract N *distinct* scenes containing multi-turn dialogues among characters, as detailed in Appendix D.1 (see Table 13). For every extracted scene, we instruct GPT-4 to generate (1) a concise, single-sentence summary (i.e., event summary) of scene information and (2) a list of the participants involved in that scene, as shown in Appendix D.1 (see Table 14).

Generate questions from event summary. Initially, we generate **fact-based questions** based on two different methods. We begin by creating *fact-based structured* questions about characters' involvement in events by combining question templates with the event summary. To this end, we curate 18 different question templates like "Tell me your feelings when {event summary}." Appendix D.2 shows all 18 question templates. Following this, we create *fact-based free-form* questions that assess a character's understanding of the event, regardless of their direct participation, as detailed in Appendix D.2. Subsequently, we proceed to formulate **fake-based questions**, employing a methodology similar to the creation of *fact-based free-form* questions, with further details available in Appendix D.2.

Assign spatiotemporal labels to each character. Given a question for a specific scene and event, the goal is to create a combination of {scene, event summary, question, character with their time point}. By choosing the character and its time point, the data type is automatically classified into one of

Question generation method	Fact-based				Fake-based
	# Future	# Past-absence	# Past-presence	# Past-only	# Past-only
Harry Potter Series					
Fact & structured	892	745	1,991	-	-
Fact & free-form	765	-	-	784	-
Fake & free-form	-	-	-	-	711
The Lord of the Rings Series					
Fact & structured	252	555	725	-	-
Fact & free-form	224	-	-	228	-
Fake & free-form	-	-	-	-	203
Twilight Series					
Fact & structured	221	277	395	-	-
Fact & free-form	176	-	-	179	-
Fake & free-form	-	-	-	-	170
The Hunger Games Series					
Fact & structured	212	309	348	-	-
Fact & free-form	181	-	-	188	-
Fake & free-form	-	-	-	-	164
Sum	10,895				

Table 3: Data statistics of four series in TIMECHARA.

four types: *future*, *past-absence*, *past-presence*, *past-only*. These data types serve as spatiotemporal labels for the {scene, event summary, question, character with their time point} combination. Appendix D.3 details how to select the character and assign its time point for each type.

Add detailed descriptions to the spatiotemporal labels. Based on the {scene, event summary, question, character with their time point} combination and the predefined data type (i.e., spatiotemporal label), we add detailed descriptions to the spatiotemporal labels to serve as a basis for evaluating role-playing agent responses. Refer to Appendix D.4 for details.

Generate gold responses and manually filter data instances. To generate a gold response for each data instance, we prompt GPT-4 with the com-

# Spatiotemporal label	HPD	Character-LLM	TIMECHARA
# Future	0 (0.0%)	57 (4.4%)	2,923 (26.8%)
# Past-absence	1 (0.7%)	0 (0.0%)	1,886 (17.3%)
# Past-presence	1 (0.7%)	0 (0.0%)	3,459 (31.7%)
# Past-only (Fact)	20 (13.4%)	856 (65.5%)	1,379 (12.7%)
# Past-only (Fake)	0 (0.0%)	0 (0.0%)	1,248 (11.5%)
# None	127 (85.2%)	394 (30.1%)	0 (0.0%)
# Total	149	1,307	10,895

Table 4: Comparison of test data statistics from three benchmarks. 'None' denotes that the instance is not included in any pre-defined labels (e.g., next response generation in a plain conversation between characters). Note that the questions for 57 *future* type instances in Character-LLM are from different eras or narratives compared to the character, while *future* type questions in TIMECHARA are from the same eras or narratives.

bination of {question, character with their time point, spatiotemporal label}. At last, we manually filter by the authors, whose criteria and results are shown in Appendix D.5.

3.5 Dataset Analyses

Statistics. The total number of event summaries is 1,643: 914 for Harry Potter, 261 for The Lord of the Rings, 245 for Twilight, and 223 for The Hunger Games. As a result, the dataset contains 10,895 instances, and Table 3 provides detailed statistics. The average lengths of (questions, gold responses, spatiotemporal labels) are (29.2, 117.6, 543.2) words, respectively.

In addition, we manually reviewed the test datasets from three different benchmarks (i.e., HPD, Character-LLM, and TIMECHARA) and classified them based on the spatiotemporal label of the given question or previous utterances in the case of multi-turn dialogue, without considering the agent's response for simplicity, as detailed in Table 4. While TIMECHARA consists of data instances evenly distributed over spatiotemporal labels, examples in HPD and Character-LLM are mainly classified as *Past-only (Fact)* or *None* types. This result demonstrates that TIMECHARA focuses on stress-testing the spatiotemporal consistency of role-playing LLMs, while the others focus on assessing the fact-based question-answering task or plain conversation between characters.

Furthermore, we provide the lexical diversity of *free-form* questions and compare it to the *structured* questions in Appendix E.1. Finally, we present a detailed human evaluation process to ensure the quality of TIMECHARA, as described in Appendix E.2.

4 Decomposed Reasoning

We find that existing LLMs struggle with spatiotemporal consistency as in Table 5, despite their extensive knowledge from books (Chang et al., 2023). To overcome this issue, we propose a reasoning method named NARRATIVE-EXPERTS, which decomposes reasoning steps into specialized tasks, employing narrative experts on either temporal (i.e., distinguishing past from future events) or spatial (i.e., identifying characters' presence in past events) aspects while utilizing the same backbone LLM.

Temporal Expert: This expert pinpoints the scene's book and chapter from a question, assigning a *future* or *past* label. If deemed *future*, it bypasses the Spatial Expert and advises the role-playing agent with a specific hint (i.e., "Note that the period of the question is in the future relative to {character}'s time point. Therefore, you should not answer the question or mention any facts that occurred after {character}'s time point.").

Spatial Expert: It assesses whether a character is involved in the scene, indicating a *past-absent* label if applicable. A tailored hint is then provided to the role-playing agent if the scene is past-absent (i.e., "Note that {character} had not participated in the scene described in the question. Therefore, you should not imply that {character} was present in the scene.").

Finally, the role-playing LLM incorporates hints from these experts into the prompt and generates a response. Appendix G.5.1 offers details of the algorithm and the prompts designed for experts.

5 Experiments on TIMECHARA

5.1 Dataset Sampling for Evaluation

Due to the high computational cost of employing GPT-4 judges, fully evaluating the 11K instance dataset is challenging. Instead, we randomly sample 600 data instances to assess the point-in-time character hallucination of role-playing agents. First, we sample 300 instances with *fact-based structured* questions, evenly distributed across three data types: *future* (100 instances), *past-presence* (100 instances), and *past-absence* (100 instances). Then, we pick 200 instances with *fact-based free-form* questions, with an equal split of 100 instances each from *future* and *past-only* types. Lastly, we choose 100 instances with *fake-based free-form* questions, all from the *past-only* type.

In addition, we provide experimental results for the entire 11K dataset in Appendix G.6.

5.2 Baseline Methods

We focus on inference-based agents as opposed to training-based agents (Shao et al., 2023), due to the impracticality of training agents to simulate characters across diverse time points; notably, our dataset includes 219 time points. We utilize four different state-of-the-art LLMs as a backbone model for role-playing agents to respond to our dataset: GPT-4o (i.e., `gpt-4o-2024-05-13`), GPT-4 (i.e., `gpt-4-1106-preview`), GPT-3.5 (i.e., `gpt-3.5-turbo-1106`), and Mistral 7B Instruct (i.e., `mistral-7b-instruct-v0.2`) (Jiang et al., 2023). To test their various reasoning capabilities, we employ several baselines as follows.

Zero-shot prompt. This is to directly prompt an agent to generate a response based on the system instruction and a question as follows:

Zero-Shot Prompt Template

System Instruction:
I want you to act like {character} from {author}'s {series_name} novel series. I want you to respond and answer like {character}, using the tone, manner, and vocabulary {character} would use. Assume that you are on {time_point} in {book_name} and interviewing with the interviewer. You should not answer the question and mention any fact that is future to the period. If he (or she) was not present at the location where the question was raised, he (or she) is likely unaware of the information or knowledge related to that question.

User Prompt:
{question}

Note that we instruct the agent to be unaware of *future* events and to acknowledge the absence when responding to *past-absence* type questions.

Zero-shot-CoT prompt. (Kojima et al., 2022). This method exploits a zero-shot prompt by adding the phrase "Let's think step by step" at the end of the question. This addition aims to improve the step-by-step reasoning capability of LLMs.

Few-shot prompt (in-context learning). This approach provides LLMs with four instances (4-shot), with details on how examples were selected available in Appendix G.2.

Iterative self-correction. Recent studies (Pan et al., 2023; Shinn et al., 2023) found that LLMs have the capability for self-correction, iteratively refining their initial responses based on the given criteria. Among various methods, we choose the self-refine (Madaan et al., 2023), since it is adaptable to dialogue domains with multiple evaluation criteria. Further details can be found in Appendix G.3.

Retrieval-augmented generation (RAG). In some prior research, retrieval-augmented generation (Lewis et al., 2020) can mitigate hallucinations (Shuster et al., 2021). We develop a retrieval module that employs OpenAI's embedding (i.e., `text-embedding-ada-002`) to provide contexts to LLMs. In addition, we add a variant named **RAG-cutoff**, which is designed to limit its retrieval exclusively to the events prior to a defined character period. Thanks to this constraint, agents can avoid access to future contexts. Further details are available in Appendix G.4.

5.3 Decomposed Reasoning with RAG

Beyond NARRATIVE-EXPERTS, we also explore NARRATIVE-EXPERTS-RAG-CUTOFF, which integrates NARRATIVE-EXPERTS with the RAG-cutoff method. We provide a complete algorithm and prompts used for experts in Appendix G.5.2.

5.4 Experimental Results

See Appendix G.1 for implementation details. Table 5 finds even GPT-4o and GPT-4, state-of-the-art LLMs, still struggle with point-in-time character hallucinations.

Future **type.** All baselines exhibit confusion with *future* type questions with accuracies at 51% or below. It highlights a prevailing issue of role-playing agents that inadvertently disclose future events. The naive RAG scores the lowest among baselines, showing that indiscriminately providing contexts harms the performance. Our NARRATIVE-EXPERTS and NARRATIVE-EXPERTS-RAG-CUTOFF significantly enhance performance, thanks to the temporal expert.

Past-absence **and** *past-only* **types.** Both naive RAG and RAG-cutoff can potentially mitigate hallucinations for these question types by leveraging context from their retrieval modules. However, their performance still lags behind that observed in *past-presence* questions, with gaps of 10% points and 13% points in *past-absence* types, and 19% points and 20% points in *past-only* types, respectively. Conversely, our methods enhance outcomes in both *past-absence* and *past-only* types, thanks to the support of both temporal and spatial experts.

Past-presence **type instances:** All baselines, except for Mistral, perform admirably, showcasing the role-playing LLMs' proficiency in memorizing narratives from novel series. Our methods slightly lag in this type due to narrative experts' occasional

Method	Spatiotemporal Consistency (%) ↑					Personality Consistency (1-7) ↑	AlignScore ↑
	Future	Past-absence	Past-presence	Past-only	Avg.		
Mistral Instruct 7B (mistral-7b-instruct-v0.2)							
zero-shot	44.5±3.5	53.0±5.0	63.0±4.9	38.0±3.4	46.8±2.0	**6.02±0.04**	18.50±0.66
RAG-cutoff	48.0±3.5	44.0±5.0	**71.0±4.6**	51.5±3.5	52.3±2.0	5.90±0.05	17.82±0.68
narrative-experts (**Ours**)	55.0±3.5	81.0±3.9	57.0±5.0	42.5±3.5	55.5±2.0	5.87±0.04	20.57±0.71
narrative-experts-RAG-cutoff (**Ours**)	**62.0±3.4**	**87.0±3.4**	66.0±4.8	58.5±3.5	**65.7±1.9**	5.85±0.04	**22.20±0.80**
GPT-3.5 Turbo (gpt-3.5-turbo-1106)							
zero-shot	29.0±3.2	33.0±4.7	**91.0±2.9**	41.5±3.5	44.2±2.0	**5.89±0.04**	24.06±0.93
RAG-cutoff	37.5±3.4	34.0±4.8	**91.0±2.9**	55.5±3.5	51.8±2.0	5.73±0.05	24.39±0.95
narrative-experts (**Ours**)	**47.5±3.5**	70.0±4.6	86.0±3.5	43.5±3.5	56.3±2.0	5.76±0.04	27.03±0.92
narrative-experts-RAG-cutoff (**Ours**)	46.0±3.5	72.0±4.5	84.0±3.7	57.5±3.5	60.5±2.0	5.61±0.05	**28.24±0.93**
GPT-4 Turbo (gpt-4-1106-preview)							
zero-shot	46.5±3.5	75.0±4.4	90.0±3.0	59.0±3.5	62.7±2.0	6.44±0.03	24.63±0.71
zero-shot-cot	48.5±3.5	75.0±4.4	92.0±2.7	61.0±3.5	64.3±2.0	6.51±0.03	23.67±0.65
few-shot	47.0±3.5	76.0±4.3	88.0±3.3	67.0±3.3	65.3±1.9	6.35±0.03	28.35±0.87
self-refine	48.0±3.5	75.0±4.4	**94.0±2.4**	65.0±3.4	65.8±1.9	6.44±0.03	24.41±0.70
RAG	33.5±3.4	81.0±3.9	91.0±2.9	72.0±3.2	63.8±2.0	**6.55±0.02**	21.14±0.64
RAG-cutoff	50.0±3.5	79.0±4.1	92.0±2.7	72.0±3.2	69.2±1.9	6.47±0.03	24.15±0.72
narrative-experts (**Ours**)	92.5±1.9	**90.0±3.0**	90.0±3.0	67.5±3.3	83.3±1.5	6.27±0.03	**31.86±0.73**
narrative-experts-RAG-cutoff (**Ours**)	**93.0±1.8**	89.0±3.1	88.0±3.3	74.5±3.1	**85.3±1.5**	6.30±0.03	31.18±0.72
GPT-4o (gpt-4o-2024-05-13)							
zero-shot	46.0±3.5	74.0±4.4	90.0±3.0	65.5±3.5	64.5±2.0	6.26±0.03	26.78±0.81
RAG-cutoff	51.0±3.5	74.0±3.5	**92.0±2.7**	74.5±3.1	69.5±1.9	**6.28±0.03**	24.27±0.73
narrative-experts (**Ours**)	94.5±1.6	84.0±3.7	83.0±3.8	68.5±3.3	82.2±1.6	6.02±0.04	**33.58±0.80**
narrative-experts-RAG-cutoff (**Ours**)	**95.5±1.5**	89.0±3.1	86.0±3.5	79.5±2.9	**87.5±1.4**	6.05±0.04	32.57±0.83

Table 5: Results of point-in-time character hallucination on 600 sampled data instances. We report the average scores with their standard error of the mean (SEM). **Bold** numbers indicate the highest scores, while underline numbers are the second-best. All responses are evaluated by GPT-4 Turbo (gpt-4-1106-preview) as judges, with the exception of measuring AlignScore (Zha et al., 2023).

mispredictions, yet this shortfall is minor compared to significant enhancement in the other three types.

Personality consistency: All methods generally maintain a consistent character portrayal, scoring above 5.6 in personality consistency. However, our methods receive lower scores from the GPT-4 judge due to their tendency to respond with unawareness regarding future events or character absences, which sometimes falls short of fully conveying the expected character's personality. In contrast, the GPT-4 judge appears to favor responses from role-playing agents that indiscriminately disclose information, regardless of its relevance to the character's knowledge boundary at a specific time point.

AlignScore evaluation: In addition, we utilize AlignScore (Zha et al., 2023), a metric based on a post-trained RoBERTa-large model, to assess the spatiotemporal consistency of the role-playing agent without relying on GPT-4 judges. Refer to Appendix G.6 for details of the AlignScore. The results in Table 5 show that AlignScore is in agreement with evaluations from GPT-4 judges, and our methods achieve the highest AlignScores across the three different backbone LLMs.

Furthermore, we provide further analyses be-

yond the main experiments, including human evaluation results that closely align with those of the LLM judges, as detailed in Appendix H.

6 Conclusion

We highlighted the importance of point-in-time role-playing agents for enhancing narrative engagement, preventing spoilers, and facilitating fandom role-play activities. To maintain a character's spatiotemporal consistency and avoid hallucinations, we introduced the TIMECHARA benchmark and developed an automated pipeline, resulting in 10,895 instances. Using TIMECHARA, we identified significant hallucination issues in state-of-the-art role-playing LLMs. To address these, we proposed NARRATIVE-EXPERTS, an effective method to reduce character hallucinations by breaking down the reasoning process and guiding it with narrative experts. Despite these efforts, our findings indicate ongoing challenges with point-in-time character hallucinations, suggesting the need for further improvements.

Limitations

Despite the advancements presented in this study, there are some limitations as follows. (1) Sourced only from English books: Since TimeChara consists primarily of texts written in English and sourced from English-speaking countries, it may reflect cultural biases inherent to these regions. One solution is to incorporate multilingual and multicultural books, as new data can be added to TimeChara automatically. (2) High costs of GPT-4 judges: The financial expenses of extensive GPT-4 evaluations can be prohibitive, restricting the feasibility of conducting large-scale assessments. An alternative would be using open-source LLMs for evaluation, such as Kim et al. (2024). (3) Latency and cost issues with Narrative-Experts: The narrative expert requires generating multiple hints and responses per question. This introduces increased latency and computational costs. Future research on efficiently generating responses while reducing point-in-time character hallucinations will be anticipated.

Ethics Statement

To mitigate any potential issue arising from the use of the four novel series, we address concerns about copyright issues as follows:

1. Source attribution: The dataset utilizes raw text from each novel series, and we acknowledge the copyrights held by the authors and publishers.

2. Fair use justification: We believe our use of the copyrighted text qualifies as "fair use" under U.S. law, whose criteria include:

   - Purpose of use: The dataset is used exclusively for non-commercial, educational, and research purposes.
   - Nature of the copyrighted work: The work is used in a research context to evaluate point-in-time character hallucination of role-playing LLMs, an inherently academic pursuit.
   - Lack of market harm: Our dataset does not substitute for the original works nor harm their market.

3. Content of dataset: TimeChara comprises only a fraction of the content necessary for dataset construction, indicating that our

dataset includes approximately 40% of the original text from the sources.

4. Accessibility and reproducibility: We will publish all dataset scripts and the dataset itself, restricting access to those who agree to use it only for research.

Besides, the generated dataset may inadvertently include harmful content intended to mislead characters. In adherence to the NLP ethics community's guidelines on 'toxic text' (Gehman et al., 2020; Askell et al., 2021), We manually reviewed all 11K data instances and filtered out those containing provocative scenes. Specifically, we removed fewer than ten cases associated with severe trauma and explicit violence that could negatively impact users who will read questions and the agent's responses.

By implementing these measures, we ensure that our research respects both the legal rights of the original content creators and the ethical standards of the research community.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards multimodal persona-grounded conversation. In *ACL*.

AI Dungeon. https://aidungeon.com/.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas

Joseph, Ben Mann, Nova DasSarma, Nelson El-hage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam Mc-Candlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *arXiv:2112.00861*.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *EMNLP*.

Character AI. https://beta.character.ai/.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv:2403.13679*.

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 *EMNLP Findings*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *NeurIPS Datasets and Benchmarks*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? *arXiv:2401.13275*.

Zheng Chu, Zekun Wang, Jiafeng Liang, Ming Liu, and Bing Qin. 2023. MTGER: Multi-view temporal graph enhanced temporal reasoning over time-involved document. In *EMNLP Findings*.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *TACL*, 10:257–273.

Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. 2023. Generic temporal reasoning with differential analysis and explanation. In *ACL*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *EMNLP Findings*.

GPTs. https://openai.com/blog/introducing-gpts.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *EMNLP*.

Kilem Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *The British journal of mathematical and statistical psychology*, 61:29–48.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *NAACL*.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. 2024. The hallucinations leaderboard – an open effort to measure hallucinations in large language models. *arXiv:2404.05904*.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *EMNLP*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv:2310.06825*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *EMNLP*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *EMNLP*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *ICLR*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv:2308.07702*.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023a. Chatharuhi: Reviving anime character in reality via large language model. arXiv:2308.09597.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for "mind" exploration of large language model society. In NeurIPS.

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. arXiv:2401.12474.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In NeurIPS.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. TMLR.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In EMNLP.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. arXiv:2401.06855.

OpenAI. 2023. Gpt-4 technical report. arXiv:2303.08774.

OpenAI. 2024. Hello gpt-4o.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. arXiv:2308.03188.

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 UIST.

Replika. https://replika.com/.

Marie-Laure Ryan. 2003. Narrative as virtual reality: Immersion and interactivity in literature and electronic media. The Johns Hopkins University Press.

Marie-Laure Ryan. 2008. Interactive narrative, plot types, and interpersonal relations. In ICIDS.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. TVShowGuess: Character comprehension in stories as speaker guessing. In NAACL.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In EMNLP.

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

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In EMNLP.

Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. arXiv:2312.16132.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In NeurIPS.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In EMNLP Findings.

SillyTavern. https://github.com/sillytavern/sillytavern.

Talkie. https://www.talkie-ai.com/.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models. In ACL.

Wang-Chiew Tan, Jane Dwivedi-Yu, Yuliang Li, Lambert Mathias, Marzieh Saeidi, Jing Nathan Yan, and Alon Halevy. 2023b. TimelineQA: A benchmark for question answering over timelines. In ACL Findings.

Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Enhancing role-playing systems through aggressive queries: Evaluation and improvement. arXiv:2402.10618.

Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. 2024. Rolecraft-glm: Advancing personalized role-playing in large language models. arXiv:2401.09432.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,

Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv:2401.01275*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *EMNLP*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A survey on large language model based autonomous agents. *Front. Comput. Sci.*, 18.

Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024b. Characteristic AI agents via large language models. In *LREC-COLING*.

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. 2023a. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv:2310.17976*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv:2310.00746*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *NeurIPS*.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *ACL*.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *arXiv:2309.07864*.

Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2023. How far are we from believable ai agents? a framework for evaluating the believability of human behavior simulation. *arXiv:2312.17115*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv:2312.07000*.

Mo Yu, Qiujing Wang, Shunchi Zhang, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Liyan Xu, Jing Li, Yue Yu, and Jie Zhou. 2022. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. *arXiv:2211.04684*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *ACL*.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *EMNLP*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv:2309.01219*.

Runcong Zhao, Qinglin Zhu, Hainiu Xu, Jiazheng Li, Yuxiang Zhou, Yulan He, and Lin Gui. 2024. Large language models fall short: Understanding complex relationships in detective narratives.

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

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv:2311.16832*.

A Related Work (Full ver.)

Concurrent role-play benchmarks. We compare TIMECHARA to concurrent role-playing datasets or benchmarks (Shen et al., 2023; Xiao et al., 2023; Tao et al., 2024; Tu et al., 2024; Lu et al., 2024; Tang et al., 2024; Wang et al., 2024b; Chen et al., 2024) released in or after December 2023, as detailed in Table 6. Note that RoleEval (Shen et al., 2023) assesses whether LLMs possess specific role knowledge by having them answer multiple-choice questions, rather than evaluating role-playing LLM agents. While some concurrent benchmarks partially address the goals of TIMECHARA intended to do (e.g., evaluation of event participant reasoning in RoleEval (Shen et al., 2023), evaluation of fake event awareness in MORTISE (Tang et al., 2024), use of temporal labels for LLM judges in WikiRole (Lu et al., 2024)), TIMECHARA remains the most comprehensive benchmark for evaluating point-in-time character hallucination.

LLM hallucinations. Since LLMs are prone to hallucinations (Ji et al., 2023; Zhang et al., 2023), many studies have evaluated hallucinations in generated texts (Zha et al., 2023; Min et al., 2023; Mishra et al., 2024; Hong et al., 2024). To reduce hallucination, some works have focused on retrieval-augmented generation (Lewis et al., 2020; Shuster et al., 2021; Mialon et al., 2023), while others have incorporated honesty alignment (i.e., refusing to answer questions when LLMs lack knowledge) in text generation (Yang et al., 2023; Cheng et al., 2024).

Unlike prior works that address general hallucination issues, we specifically focus on *point-in-time character* hallucination, which includes unique spatiotemporal consistency problems such as *future unawareness*, *absence awareness*, and *fake event awareness*. Furthermore, we propose a new method, NARRATIVE-EXPERTS, to mitigate these hallucinations.

Neural theory-of-mind in narrative understanding. Neural theory-of-mind (ToM) examines whether LLMs possess the ability to understand the mental stages (e.g., thoughts, beliefs, and intentions) of others (Sap et al., 2022; Kim et al., 2023). In narrative understanding, ToM capabilities are essential for role-playing agents to comprehend both the narratives and the characters' minds (Sang et al., 2022; Yu et al., 2022; Zhao et al., 2024).

Although it is well known that LLMs memorize extensive knowledge from books (Chang et al., 2023) and can precisely answer questions about narrative (See Table 28 in Appendix H.5), we found that they cannot maintain spatiotemporal consistency while acting as point-in-time role-playing agents, as shown in Table 5. The results demonstrate that LLMs are not yet capable of ToM in point-in-time role-playing scenarios.

B Prompt Demonstration

Prompts for GPT-4 Turbo judges. We present the prompts used for evaluating the two dimensions as follows:

1. Spatiotemporal consistency: refer to Table 7.

2. Personality consistency: refer to Table 8.

C Examples of TIMECHARA

We show examples of our dataset across four different data types: *future*, *past-absence*, *past-presence*, and *past-only*, each illustrated as follows:

1. *future*: An instance of *future* data is presented in Table 2, characterized by its *fact-based structured* question.

2. *past-absence*: Table 9 illustrates the *past-absence* type with a similarly *fact-based structured* question.

3. *past-presence*: Table 10 shows an example of *past-presence* type data, whose question is also a *fact-based structured* question.

4. *past-only*: Demonstrations of *past-only* type data are provided in Table 11 for a *fact-based free-form* question and in Table 12 for a *fake-based free-form* question.

D Details of Dataset Construction

D.1 Extract scenes, event summaries, and participant lists from book

For the Harry Potter series, we utilize the scene dataset from Chen et al. (2023), which comprises manually selected 1,037 unique scenes from the books. For the other three series, we extract 300 distinct scenes containing multi-turn dialogues among characters, as detailed in Table 13. Note that we include multi-turn dialogues among characters to

Evaluation Dataset / Benchmark	Dataset automatically constructed?	Support point-in-time role-playing?	Evaluate near-future unawareness? (Temporal)	Evaluate absence awareness? (Spatial)	Evaluate fake event awareness? (Fake question)	Evaluation method for character hallucination
***Role-Playing Domain* (Concurrent)**						
RoleEval (2023/12)	✗	✗	✗ *	Δ	✗	Answer selection accuracy: [0 or 1] (selection-based)
SimulateBench (2023/12)	✗	✗	✗	✗	✗	Answer selection accuracy: [0 or 1] (selection-based)
RoleInstruct (2024/01)	✓	✗	✗	✗	✗	Human as judges: [1 to 5] (unscalable)
CharacterEval (2024/01)	✓	✗	✗	✗	✗	CharacterRM: [1 to 5] (trainable)
WikiRole (2024/01)	✓	✗	✗ (Question from distinct era/narrative: easy)	✗	✗	LLM as judges w/ temporal labels: [0 or 1] (accurate)
MORTISE (2024/02)	✓	✗	✗	✗	✓	RC.Score: [1 to 5] (trainable)
Character100 (2024/03)	✓	✗	✗	✗	✗	Rouge-L w/ gold response: [0-1] (implicit)
RoleInteract (2024/03)	✓	✗	✗	✗	✗	Answer selection accuracy: [0 or 1] (selection-based)
TIMECHARA (2024/02 submitted)	✓	✓	✓ (Question from the same era/narrative: hard)	✓	✓	LLM as judges w/ spatiotemporal labels: [0 or 1] (accurate)

Table 6: Comparison of TIMECHARA with concurrent role-playing datasets or benchmarks released in or after December 2023: RoleEval (Shen et al., 2023), SimulateBench (Xiao et al., 2023), RoleInstruct (Tao et al., 2024), CharacterEval (Tu et al., 2024), WikiRole (Lu et al., 2024), MORTISE (Tang et al., 2024), Character100 (Wang et al., 2024b), and RoleInteract (Chen et al., 2024). Note that RoleEval evaluates whether LLMs possess specific role knowledge by having them answer multiple-choice questions, rather than evaluating role-playing LLM agents. *It supports evaluating timeline reasoning, where LLMs are required to sort related events in temporal order. However, this does not mean it evaluates *near-future unawareness* because it assumes that LLMs are omniscient about the characters. Δ indicates that RoleEval does not explicitly evaluate *absence awareness* since it lacks role-playing scenarios, but it implicitly supports reasoning about event participants (i.e., identifying who took part in an event). In the last column, 'selection-based' means that the task involves selecting the correct answer for multiple-choice questions rather than generating an open-ended response. 'Trainable' means that the reward models, CharacterRM and RC.Score, are trained using datasets evaluated by human judges. Although those models are more aligned with human judges than GPT-4 judges, it is unknown whether they are still effective at evaluating LLMs that role-play out-of-domain characters, characters not in the training dataset.

ensure that each scene features interaction among participants, with at least one character involved. In addition, Table 14 shows an example of a prompt for generating an event summary and participant lists.

D.2 Generate questions from event summary

For *fact-based structured* questions, we use a total of 18 different question templates as follows.

1. Tell me your feelings when {event summary}.

2. Tell me your genuine feelings when {event summary}.

3. Describe your feelings when {event summary}.

4. Describe your honest feelings when {event summary}.

5. Can you describe your experience when {event summary}?

6. Can you describe your true experience when {event summary}?

7. Did you see the moment when {event summary}.

8. Did you truly see the moment when {event summary}?

9. What did you see as {event summary}?

10. What did you actually see as {event summary}?

11. What did you hear when {event summary}?

12. What did you precisely hear when {event summary}?

Prompt for Spatiotemporal Consistency Evaluation

You will be given responses written by an AI assistant mimicking the character {agent_name}. Your task is to rate the performance of {agent_name} using the specific criterion by following the evaluation steps. Below is the data:

***
[Interactions]
Interviewer: {question}
{agent_name}: {response}
***
[Fact]
{spatiotemporal_label}

[Evaluation Criterion]
Spatiotemporal Consistency (0 or 1): Is the response consistent with the character's spatiotemporal knowledge?

[Evaluation Steps]
1. Read through the [Fact] and identify the knowledge scope of the character.
2. Read through the interactions and responses of the AI assistant to find the evidence of knowledge used in the response.
3. Compare the evidence to the [Fact]. Check if the response is consistent with the character's knowledge scope.
4. If some knowledge contradicts or contains inconsistencies about the [Fact], given a 0 score. Otherwise, assign a 1 score.
***

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then, print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line.

Table 7: Prompt for GPT-4 Turbo judges to evaluate spatiotemporal consistency.

Prompt for Personality Consistency Evaluation

You will be given responses written by an AI assistant mimicking the character {agent_name}. Your task is to rate the performance of {agent_name} using the specific criterion by following the evaluation steps. Below is the data:

***
[Interactions]
Interviewer: {question}
{agent_name}: {response}
***
[Personality]
{personality_label}

[Evaluation Criterion]
Personality Consistency (1-7): Is the response consistent with the character's personality?

[Evaluation Steps]
1. Read through the [Personality] and write the personalities, including preferences, values, and convictions of the real character.
2. Read through the interactions and identify the personalities, including preferences, values, and convictions of the AI assistant.
3. After having a clear understanding of the interactions, compare the response to the [Personality]. Look for any consistencies or inconsistencies. Do the responses reflect the character's personalities, including preferences, values, and convictions?
4. Use the given scale from 1-7 to rate how well the response reflects the personalities, including preferences, values, and convictions of the character. 1 being not at all reflective of the character's personalities, and 7 being perfectly reflective of the character's personalities.
***

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then, print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line.

Table 8: Prompt for GPT-4 Turbo judges to evaluate personality consistency.

Scene	"Why can't we get through?" Harry hissed to Ron... "I think we'd better go and wait by the car," said Harry...
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King's Cross.
Question	"Tell me your feelings when {Event Summary}."
Character	2nd-year Hermione Granger on Halloween
Data Type	Past-absence
Spatiotemporal Label	**Past**: During her 2nd-year on Christmas, Hermione Granger can respond based on the moment but should not wrongly recall it. - Moment: {Scene} **Absence**: During her 2nd-year on Christmas, Hermione Granger should not say that she was present when {Event Summary}.
Personality Label	Hermione Granger is portrayed as exceptionally intelligent and hardworking, often topping her classes and ...
Gold Response	"Oh no, I wasn't there when it happened. I'd already gone through the barrier and was on the Hogwarts Express..."

Table 9: An example of our *past-absence* type data instance with the *fact-based structured* question.

Scene	"Why can't we get through?" Harry hissed to Ron... "I think we'd better go and wait by the car," said Harry...
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King's Cross.
Question	"Tell me your feelings when {Event Summary}."
Character	2nd-year Ronald Weasley on Halloween
Data Type	Past-presence
Spatiotemporal Label	**Past**: During his 2nd-year on Halloween, Ronald Weasley can respond based on the moment but should not wrongly recall it. - Moment: {Scene}. **Presence**: During his 2nd-year on Halloween, Ronald Weasley should not say that he was absent when {Event Summary}.
Personality Label	Ronald Weasley is depicted as a loyal, brave, strong, and humorous individual, yet sometimes immature and ...
Gold Response	"Blimey, yeah, I was there, wasn't I? It was mental. One minute we're running towards the barrier between..."

Table 10: An example of our *past-presence* type data instance with the *fact-based structured* question.

Scene	"Why can't we get through?" Harry hissed to Ron... "I think we'd better go and wait by the car," said Harry...
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King's Cross.
Question	"Why did Harry and Ron use the enchanted car to get to Hogwarts?"
Character	2nd-year Harry Potter at the end of the scene
Data Type	Past-only
Spatiotemporal Label	**Past**: At the end of the scene of Harry Potter and the Chamber of Secrets as a 2nd-year student, Harry Potter can respond based on the moment but should not wrongly recall it. - Moment: {Scene}. - Answer: ... Due to crashing into the barrier at King's Cross Station, prompting them to fly to Hogwarts...
Personality Label	Harry Potter is characterized by his selflessness and immense loyalty, especially towards his friends...
Gold Response	"Well, it was a bit of a mess, really. Ron and I tried to go through the barrier to Platform Nine and Three-Quarters..."

Table 11: An example of our *past-only* type data instance with the *fact-based free-form* question.

13. Were you at the moment when {event summary}?

14. Were you really at the moment when {event summary}?

15. Were you present as {event summary}?

16. Were you indeed present as {event summary}?

17. Is it true that you were at the moment when {event summary}?

18. Is it right that you were at the moment when {event summary}?

Note that the templates were randomly chosen to ask if a character experienced a specific event at a location. We found that using more than 18 templates didn't enhance the diversity of expression. In addition, The inclusion of 'feeling' questions aimed to enrich the narrative by encouraging a vivid description of the character's personal emotions during an event, providing depth to their direct experiences.

For *fact-based free-form* questions, we prompt GPT-4 Turbo to generate questions using the 5W1H approach (i.e., what, who, where, when, why, and

Scene	"Why can't we get through?" Harry hissed to Ron... "I think we'd better go and wait by the car," said Harry...
Event Summary	Harry and Ron took the enchanted car to Hogwarts after a barrier mishap at King's Cross.
Question	"Why did Harry and Ron consider swimming to Hogwarts after the barrier incident at King's Cross?"
Character	2nd-year Harry Potter at the end of the scene
Data Type	Past-only
Spatiotemporal Label	**Past**: At the end of the scene of Harry Potter and the Chamber of Secrets as a 2nd-year student, Harry Potter can respond based on the moment but should not wrongly recall it. - Moment: {Scene}. - Answer: Harry and Ron did not decide to swim to Hogwarts; they decided to fly to Hogwarts in the Weasley's...
Personality Label	Harry Potter is characterized by his selflessness and immense loyalty, especially towards his friends...
Gold Response	"Oh, blimey, swim to Hogwarts? That would've been a right mess, wouldn't it? No, no, we didn't swim..."

Table 12: An example of our *past-only* type data instance with the *fake-based free-form* question.

Prompt for Scene and Speakers List Extraction

First, read chapter 1, part 1 of the book "The Lord of the Rings - The Fellowship of the Ring". Then, extract 5 parts from the raw text with dialogues that can be considered as one scene. Each part should meet the following requirements. Start by analyzing the text that I gave you.

1. Each scene should be unique throughout the entire series: The Lord of The Ring.
2. Each scene shouldn't be ambiguous, which means that the summary of each scene should be talking or related to a specific event, item, or person.
3. Scenes shouldn't be everyday conversation such as the summary of the scene being: "Frodo Baggins talked to Sam about his breakfast", which could be an everyday conversation.
4. Each scene should contain at least 5 dialogues. The extracted raw text should be between 15 to 35 sentences long to sufficiently form the scene and contain sufficient information about the scene.

For each scene, please provide:
1. A summary of the scene you selected in one sentence.
2. The raw text that you selected.
3. The full name of the characters speaking in that scene.
***
Input:
- Raw Text: {raw_text}

Output:

Table 13: An example of a prompt for extracting scenes and speakers list from the Lord of the Rings series.

---

how) based on the event summary, as shown in Table 15. For *fake-based free-form* question, we first generate a fake event summary based on a true event summary by utilizing GPT-4 Turbo, as outlined in Table 16. Note that we use six different strategies for generating fake event summaries: (1) changing the character, (2) changing the key object, (3) altering the location, (4) switching the action, (5) introducing a nonexistent character or object, and (6) changing the character's knowledge. Subsequently, we generate a fake question by prompting GPT-4 Turbo to generate 5W1H question based on the fake event summary, as detailed in Table 17. The generation of all *free-form* questions was conducted in a single trial. The choice of 5W1H questions was random, with the distribution as follows: what—48.2%, who—19.0%, why—17.4%, how—9.8%, when—4.6%, and where—1.0%. The prevalence of 'what' questions reflects the foundational approach of our *free-form* question generation process, which draws upon event summaries that focus on entities (characters, objects, events)

and their interrelationships. Specifically, the versatility of 'what' questions in covering a wide array of topics related to these entities (e.g., "What's the meaning of the spell?", "What event occurred?", "What action was taken?", "what do you think of him/her?", etc) likely led to their predominance.

D.3 Assign spatiotemporal labels for each character in their time point.

We first identify distinct time points in each character's timeline and annotate these moments with the relevant book and chapter for timeline comparison with the questions. We enlist all character time points in Appendix F.

For constructing (1) *future* type data instances, we employ both *structured* and *free-form* questions. Given a question that originates from a specific scene and event, we randomly select a character with the condition that the character must be from the novel series relevant to the event's source. We then set their time point to be earlier than that of the event, based on a comparison of the book and

Prompt for Event Summary and Participant List Generation

First, read the scene and dialogue. Then, generate a single, unique "fact" sentence in the past tense that captures Harry's distinct moment or experience that is retrievable from the scene. If there are several moments, pick the most unique moment and write it. Write it concisely. Finally, extract participants who are physically present and existing in the scene.

***

Input:
- Position: Book1-chapter2
- Speakers: Petunia, Vernon, Harry
- Scene: "Bad news, Vernon," she said. "Mrs. Figg's broken her leg. She can't take him." She jerked her head in Harry's direction. Dudley's mouth fell open in horror, but Harry's heart gave a leap. Every year on Dudley's birthday, his parents took him and a friend out for the day, to adventure parks, hamburger restaurants, or the movies. Every year, Harry was left behind with Mrs. Figg, a mad old lady who lived two streets away. Harry hated it there. The whole house smelled of cabbage and Mrs. Figg made him look at photographs of all the cats she'd ever owned. "Now what?" said Aunt Petunia, looking furiously at Harry as though he'd planned this. Harry knew he ought to feel sorry that Mrs. Figg had broken her leg, but it wasn't easy when he reminded himself it would be a whole year before he had to look at Tibbles, Snowy, Mr. Paws, and Tufty again. "We could phone Marge," Uncle Vernon suggested. "Don't be silly, Vernon, she hates the boy." The Dursleys often spoke about Harry like this, as though he wasn't there — or rather, as though he was something very nasty that couldn't understand them, like a slug. "What about what's-her-name, your friend — Yvonne?" "On vacation in Majorca," snapped Aunt Petunia. "You could just leave me here," Harry put in hopefully (he'd be able to watch what he wanted on television for a change and maybe even have a go on Dudley's computer). Aunt Petunia looked as though she'd just swallowed a lemon. "And come back and find the house in ruins?" she snarled. "I won't blow up the house," said Harry, but they weren't listening. "I suppose we could take him to the zoo," said Aunt Petunia slowly, ". . . and leave him in the car. . . ." "That car's new, he's not sitting in it alone. . . ." Dudley began to cry loudly. In fact, he wasn't really crying — it had been years since he'd really cried — but he knew that if he screwed up his face and wailed, his mother would give him anything he wanted. "Dinky Duddydums, don't cry, Mummy won't let him spoil your special day!" she cried, flinging her arms around him. "I . . . don't . . . want . . . him . . . t-t-to come!" Dudley yelled between huge, pretend sobs. "He always sp-spoils everything!" He shot Harry a nasty grin through the gap in his mother's arms. Just then, the doorbell rang —"Oh, good Lord, they're here!" said Aunt Petunia frantically — and a moment later, Dudley's best friend, Piers Polkiss, walked in with his mother. Piers was a scrawny boy with a face like a rat. He was usually the one who held people's arms behind their backs while Dudley hit them. Dudley stopped pretending to cry at once. Half an hour later, Harry, who couldn't believe his luck, was sitting in the back of the Dursleys' car with Piers and Dudley, on the way to the zoo for the first time in his life. His aunt and uncle hadn't been able to think of anything else to do with him, but before they'd left, Uncle Vernon had taken Harry aside. "I'm warning you," he had said, putting his large purple face right up close to Harry's, "I'm warning you now, boy — any funny business, anything at all — and you'll be in that cupboard from now until Christmas."

Output:
- Unique Fact: The Dursleys reluctantly decided to take Harry to the zoo with them for the first time in his life but warned him of severe consequences if he caused any trouble.
- Participants: Aunt Petunia, Dudley Dursley, Harry Potter, Uncle Vernon Dursley, Mrs. Figg, Piers Polkiss

***

Input:
- Position: {position}
- Speakers: {speakers}
- Scene: {extracted_scene}

Output:

Table 14: An example of a prompt for generating an event summary and participant list for the Harry Potter series.

Prompt for Fact-based Free-from Question Generation

First, read the event summary from the Harry Potter series. Then, paraphrase the event summary to (1) a single-sentence question among 5w1h questions and (2) the answer to the question that should be answerable from the given event summary. Don't use pronouns to indicate the event, but self-contain what event it is. Note that the question should identify the unique period of the story.

***

Input:
- Event summary: Ron's broken wand caused the charm to backfire, erasing Lockhart's memory and causing a portion of the ceiling to cave in.

Output:
- Question: What caused Gilderoy Lockhart's memory loss and the partial collapse of the ceiling?
- Answer: Gilderoy Lockhart's memory was erased, and a portion of the ceiling caved in when Ron Weasley's broken wand caused a backfired charm in their second year at Hogwarts.

***

Input:
- Event summary: Harry uncovered that it was Professor Quirrell who attempted to seize the Sorcerer's Stone, revealing that he was under the influence of Lord Voldemort, who existed parasitically on the reverse side of Quirrell's head.

Output:
- Question: Who did Harry Potter find out was attempting to steal the Sorcerer's Stone and was possessed by Lord Voldemort during their encounter at Hogwarts, and where was Voldemort residing on the individual's body?
- Answer: Harry Potter discovered that Professor Quirrell, with Lord Voldemort residing on the back of his head, was trying to steal the Sorcerer's Stone.

***

Input:
- Event summary: {event_summary}

Output:

Table 15: An example of a prompt for generating a *fact-based free-form* question for the Harry Potter series.

Prompt for Fake Event Summary Generation

First, read the event summary from the Harry Potter series. Generate the fake event summary that converts the true event summary to confuse readers using one of the six methods as follows.

1. Change the character: Swap the character with another character.
- True: Harry tricked Malfoy into freeing Dobby by giving Malfoy one of his own socks, which he promptly threw away and was caught by Dobby.
- Fake: Harry tricked Snape into freeing Dobby by giving Snape one of his own socks, which he promptly threw away and was caught by Dobby.

2. Change the Key Object: Alter the object that is central to the event.
- True: Harry used his own sock to free Dobby.
- Fake: Harry used a spellbook to free Dobby.

3. Alter the Location: Change the setting where the event took place.
- True: The event took place in Malfoy Manor.
- Fake: The event took place in the Gryffindor common room.

4. Switch the Action: Change what was done to the object or the action taken by the character.
- True: Malfoy threw the sock away.
- Fake: Malfoy donated the sock to charity.

5. Introduce a Nonexistent Character or Object: Add someone or something that wasn't originally there.
- True: Harry and Malfoy were the main characters involved.
- Fake: Harry, Malfoy, and a ghost named Sir Pudding were involved in the exchange.

6. Change the Character's Knowledge: Switch what the character knows or doesn't know.
- True: Harry knew the sock would free Dobby.
- Fake: Harry had no idea that the sock would free Dobby and thought it was just a useless gift.
***
Input:
- True event summary: Harry received a Nimbus 2000, a gift from Professor McGonagall.

Output:
- Fake event summary: Harry received a Nimbus 2000, a gift from Professor Snape.
- Method number: 1. Change the character
***
Input:
- True event summary: Fred, George, and Ron rescued Harry from the Dursleys with the use of a Flying Ford Anglia.

Output:
- Fake event summary: Fred, George, and Ron rescued Harry from Hogwarts with the use of a Flying Ford Anglia.
- Method number: 3. Alter the Location
***
Input:
- True event summary: {true_event_summary}

Output:

Table 16: An example of a prompt for generating a fake event summary for the Harry Potter series.

Prompt for Fake-based Free-from Question Generation

First, read two event summaries from the Harry Potter Series. One is a true event summary, and the other is a fake event summary, which is generated from the true event summary to confuse readers. Then, paraphrase the fake event summary to (1) a single-sentence fake question among 5w1h questions and (2) the true answer to the question that should be answerable from the given true event summary. Don't use pronouns to indicate the event, but self-contain what event it is. Note that the question should identify the unique period of the story.
***
Input:
- True event summary: Harry received a Nimbus 2000, a gift from Professor McGonagall.
- Fake event summary: Harry received a Nimbus 2000, a gift from Professor Snape.

Output:
- Fake question: Why did Professor Snape give Harry a Nimbus 2000?
- True answer: Professor Snape did not give Harry a Nimbus 2000; it was a gift from Professor McGonagall.
***
Input:
- True event summary: Fred, George, and Ron rescued Harry from the Dursleys with the use of a Flying Ford Anglia.
- Fake event summary: Fred, George, and Ron rescued Harry from Hogwarts with the use of a Flying Ford Anglia.

Output:
- Fake question: How did Fred, George, and Ron rescue Harry from Hogwarts using a Flying Ford Anglia?
- True answer: Fred, George, and Ron did not rescue Harry from Hogwarts; they rescued him from the Dursleys' house using a Flying Ford Anglia.
***
Input:
- True event summary: {true_event_summary}
- Fake event summary: {fake_event_summary}

Output:

Table 17: An example of a prompt for generating a *fake-based free-form* question for the Harry Potter series.

chapter numbers between the character's and the event's time points. Note that the character's time point is chosen randomly from among those that are earlier than the event's time point. For generating (2) *past-only* type data instances, we only use *free-form* questions (i.e., both fact-based and fake-based questions). This is because *free-form* questions are aimed at assessing the character's overall knowledge of past events (which is crucial for *past-only* instances), whereas *structured* questions are designed to assess their direct experiences or observations. In this case, the character's period is randomly chosen from among those that occur after the event's time point. For creating data instances of (3) *past-presence* and *past-absence* types, we only use *structured* questions. This is because, similar to the *past-only* instances, *free-form* questions are aimed at assessing the character's general knowledge of past events, whereas *structured* questions are more focused on their direct experiences or observations. Here, we select a character who was present/absent at the event, utilizing the previously generated participants list. Their time point is then adjusted to be after that of the event. To scale up the dataset for past-presence and past-absence types, we apply this process to all main characters to the same question instead of selecting only a single character per question, resulting in K data instances per question (e.g., $K = 3$ in the context of the Harry Potter series).

D.4 Add detailed descriptions to the spatiotemporal labels.

All {scene, event summary, question, character with time point} combinations are categorized into four types based on their spatiotemporal relevance. Based on each type, we enrich the spatiotemporal labels as follows: (1) *Future* type: we include annotations like "During {character time point}, {character} should not be aware of or contain any expression that reveals the moment when {event summary} since the moment is the future." in the spatiotemporal label (see Table 2). (2) *Past-only* type: we include statements like "During {character time point}, {character} can respond based on the moment but should not wrongly recall it. (- Moment: {scene})." in the spatiotemporal label (see Table 11). (3) *Past-absence* and *past-presence* type: we add "During {character time point}, {character} should not say that he/she was present/absent when {event summary}" to the spatiotemporal label, based on the previously generated list of partic-

ipants while adding the same statements as for *past-only* type (see Table 9 and Table 10). In addition, we add a personality label for all questions, tailored to the character, by summarizing personality traits from their Fandom page personality section.

D.5 Generate gold responses and manually filter data instances.

We apply a manual filtering process that involves a series of assessments to ensure that each data instance meets all criteria:

1. Event summary quality: we evaluate the quality of each generated event summary, ensuring a survival rate of 83.96% (1,643 out of 1,957 scenes). The criteria for exclusion are as follows. (1) Lack of uniqueness: instances describing common events, such as "Harry struggled with casting spells using a blackthorn wand." are excluded due to their repetitiveness throughout the story. (2) Ambiguity: we exclude instances where the time point or participants of the event are unclear. For example, instances are excluded if there is no specific event or object that indicates the unique time point, such as "at a class", or if it is unclear who the participants are, as in "revealed to the school".

2. Participants list accuracy: we assess the accuracy of the list of participants in each scene. 97.45% of them are correct (1,907 out of 1,957 scenes).

3. *Fact-based free-form* question quality: we scrutinize the quality of the *fact-based free-form* questions, with 86.46% (1,379 out of 1,595 questions). The criteria are as follows. (1) Lack of uniqueness: we exclude questions if they can be answered in multiple ways. For example, the question "Who informed Harry Potter about the threat of Sirius Black, and how did Harry respond?" is not unique since multiple characters cautioned Harry about Sirius at different times. (2) Ambiguity: we exclude questions when the entities in the question are unclear. For instance, a question like "What triggered the disturbance at the Dursleys' breakfast table that involved Harry Potter and resulted in chaos?" is considered ambiguous due to its vague references. (3) Incorrect questions: we exclude questions that are incorrectly formulated or answer themselves. For

example, a question that embeds its own answer is "Who did Rubeus Hagrid introduce to Harry Potter and Hermione Granger as his half-brother, Grawp the giant, and request they care for in the event of his dismissal from Hogwarts?" (4) Duplication: we exclude questions when generated questions are nearly identical due to closely related event summaries.

4. *Fake-based free-form* question quality: we evaluate the *fake-based free-form* questions, achieving a survival rate of 78.24% (1,248 among 1,595 questions). The criteria are as follows. (1) Clarity: we exclude questions that remain fact-based instead of being appropriately converted to fake questions. (2) Incorrect questions: we remove questions that are irrelevant or incorrectly related to the given event's context.

5. Gold response quality: our review of the gold responses for 10,895 {question, character with timepoint} pairs results in a survival rate of 98.04% (10,682 among 10,895). We do not exclude responses that directly contradict or are inconsistent with the assigned spatiotemporal label. Instead, we continue to regenerate the response until it meets the assigned spatiotemporal label.

E   Details of Dataset Analyses

E.1   Statistics

We present a nested pie chart of *free-form* questions in Figure 4, and another nested pie chart of *structured* questions in Figure 5. Since *structured* questions are derived from only 18 different templates, they lack the diversity of *free-form* questions, which are generated using 5W1H (i.e., who, what, when, where, why, how) approach. From another perspective, the diverse verb-noun structures in *free-form* questions indicate their aim to probe a character's overall knowledge of past events, whereas verb-noun structures in *structured* questions are designed to elicit direct experiences or observations. Nevertheless, *structured* questions play a crucial role in exploring the limits of a character's knowledge boundary. That is, they include concepts of *past-absence* and *past-presence*, which are absent in *free-form* questions.
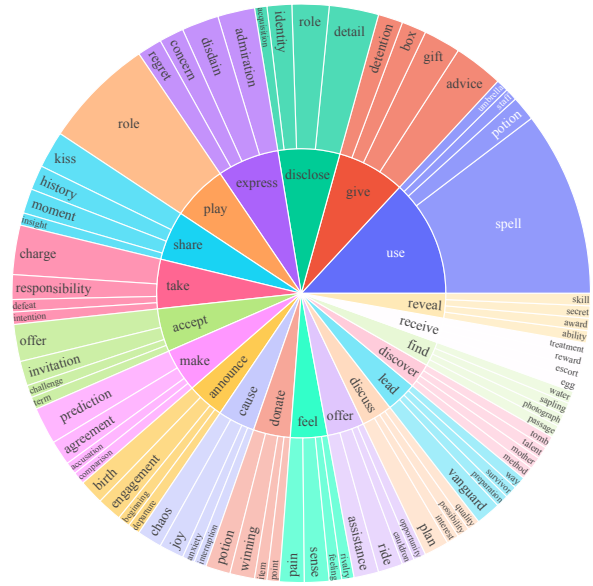

Figure 4: A nested pie chart of verb-noun structures in *free-form* questions, encompassing both fact-based and fake-based.

E.2   Dataset Quality

In order to ensure the quality dataset, we additionally performed a human evaluation conducted via Amazon Mechanical Turk.

First, we provided annotators with comprehensive instructions as shown in Figure 6, including several examples for each criterion. The criteria are as follows: (1) the quality of the generated event summary (0 if bad, 1 if good), (2) the accuracy of participant lists (0 or 1), (3) the quality of questions generated from event summaries (0 or 1), (4) spatiotemporal consistency of the gold response (0 if inconsistent, 1 if consistent), and (5) personality consistency of the gold response (1 to 7). Note that the last criterion, 'personality consistency', is not our main contribution; it revisits a concept introduced by Shao et al. (2023). Additionally, we ensured that annotators were provided with 'extracted scene' information, in Figure 7, to enhance their contextual understanding.

Secondly, annotators were required to pass the 1st qualification Human Intelligence Task (HIT), aimed at ensuring their comprehension of the provided instructions. The compensation for each HIT was $0.75. We restricted our selection to workers from English-speaking countries (namely AU, CA, NZ, US, and GB) with a HIT approval rate of over 98%, with greater than 5000 HITs approved. Each annotator was assigned a single example, randomly chosen from four novels. It was mandatory for
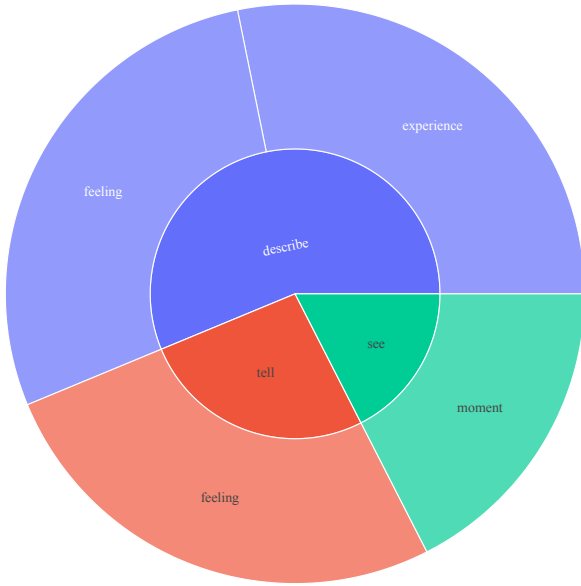
Figure 5: A nested pie chart of verb-noun structures in *structured* questions.

them to not only select the appropriate labels but also to write their rationales in complete sentences. Based on their choices and provided rationales, we identified and qualified annotators who were both accurate in their selection and provided sound rationales. This process resulted in 42 qualified workers out of an initial pool of 86.

Thirdly, annotators were tasked with completing the 2nd qualification HITs, designed to expose them to a variety of examples before attempting the main HITs. We presented 200 examples, each evaluated by two annotators. We also committed to offering our annotators a fair compensation of approximately $16/hour (with each HIT paying $1.0), which surpasses the minimum wage in the countries from which we recruited. This stage involved 38 annotators, yielding average scores for (1) summary quality: 0.95, (2) participants list accuracy: 0.96, (3) question quality: 0.94, (4) spatiotemporal consistency of the gold response: 0.75, and (5) personality consistency of the gold response: 4.82. We use Gwet's AC1 scores to measure inter-annotator agreement, with scores of (1) 0.9, (2) 0.91, (3) 0.87, (4) 0.5, and (5) 0.17. Note that we chose Gwet's AC1 score for its robust performance in situations of class imbalance (Gwet, 2008; Wongpakaran et al., 2013; Wong et al., 2021). Upon reviewing their selections and rationales, we observed that some annotators still struggled with the concept of spatiotemporal consistency. As a result, we excluded 9 workers, ultimately assembling a

team of 29 annotators fully prepared to undertake the main HITs.

Finally, we assigned workers to complete the main HITs. We presented 100 examples, each evaluated by two annotators. We compensated the workers at a rate of $1.00 per HIT, and a total of 19 workers completed their assignments. The final results of the main HITs were favorable, with average scores of (1) summary quality: **0.88**, (2) participants list accuracy: **0.92**, (3) question quality: **0.89**, (4) spatiotemporal consistency of the gold response: **0.95**, and (5) personality consistency of the gold response: 5.91, with substantial Gwet's AC1 scores of (1) **0.74**, (2) **0.85**, (3) **0.81**, (4) **0.90**, and (5) 0.19, respectively. We also highlight the agreement ratios between annotators choosing the same versus different labels: **0.78** vs. 0.22 for the quality of the summary, **0.86** vs. 0.14 for the accuracy of the participants' list, **0.84** vs. 0.16 for the quality of the questions, **0.90** vs. 0.10 for the spatiotemporal consistency of the responses, and 0.32 vs. 0.68 for the personality consistency of the responses. Unlike the evaluation of personality consistency, which is inherently subjective, assigning binary labels to the other criteria (e.g., spatiotemporal consistency) is more straightforward due to their objective nature and the presence of clear, definitive answers, as demonstrated by the results.

F   Character Time Points

We select 14 different characters across four different novel series, assigning a total of 219 unique {character, time point} pairs.

For the Harry Potter series, time points correspond to the Hogwarts school year for each character. We have identified four key periods within this timeline: the start of the school year on September 1st, Halloween, Christmas, and the end of the scene. Note that these time points (A total of 25 time points) are consistently applied across three main characters—**Harry Potter, Ronald Weasley, and Hermione Granger**. To facilitate a comparison of time points with those in the questions, we manually annotate each character's time point using the format {book_number - chapter_number}, categorized as follows:

- September 1st: [1-6, 2-5, 3-5, 4-11, 5-10, 6-7, 7-12]

- Halloween: [1-10, 2-8, 3-8, 4-16]

Criteria for Ranking the Interview Construction Pipeline:
We would like you to read the whole interview construction pipeline presented on the left first (which includes the **Extracted Scene, Event Summary, Participants List, Fake Method, Fake-based Interview Question, Temporal Label, Spatial Label, and Response from Character at their specific time points**), and then rate each component based on the following criterion. We advise that workers should ideally be fans of the respective novel series, or at the very least, have some familiarity with them.

1. Event Summary Quality:
The **Event Summary** should meet two key criteria. If these criteria are not satisfied, the summary should be deemed *bad*.
(1) First, the **Event Summary** should be *unique* throughout the series. For example, "Harry struggled with casting spells using a blackthorn wand." is not a unique **Event Summary**, since it occurs multiple times during the "Harry Potter" series.
(2) Second, the **Event Summary** should be *clear and not ambiguous*. An **Event Summary** with excessive use of pronouns or unclear nouns can be ambiguous when considered in isolation, without the context of the **Scene**. For instance, "Hermione Granger stormed out of the class after being insulted for her skeptical attitude towards the *subject*." is ambiguous.

2. Participants List Accuracy:
Based on the extracted **Scene**, check if all the characters, who are *physically present* in the **Scene**, are listed in the **Participants List**. If even a single character is missing or an extraneous individual is included, the list should be considered *inaccurate*.

3. Fake-based Interview Question Quality:
The quality of **Fake-based Interview Question** should be evaluated by two criteria. If these criteria are not satisfied, the question should be deemed *bad*.
(1) First, **Fake-based Interview Question** *shouldn't be directly answered* with the **True Answer** in **Temporal Label**. For example, "Fake Question: What was Harry instructed to prepare by Aunt Petunia for Dudley's birthday breakfast to ensure it didn't fall apart?" can be directly answered with the "True Answer: Harry was ordered to watch the bacon for Dudley's birthday breakfast and warned not to let it burn.".
(2) Second, check if the **Fake-based Interview Question** is made properly from the **Event Summary** using the *specified Fake Method*. For instance, "Fake Question: After a harrowing encounter with a three-headed dog, who scolded Harry and Ginny for not noticing it was standing on a trapdoor, suggesting it was guarding something?" made from "Event Summary: After a harrowing encounter with a three-headed dog, Hermione scolded Harry and Ron for not noticing it was standing on a trapdoor, suggesting it was guarding something." with "Fake Method: 4. Switch the action" is *bad*, since the **Question** is generated with 1. Change the character (=Ron to Ginny) method instead of 4. Switch the action method.

4. Response Quality (Spatiotemporal Consistency):
The **Response** *should adhere to* the provided **Temporal Label** and **Spatial Label**. In particular, for the **Response** to the **Fake Question**, it is necessary to check whether the **Response** matches the **True Answer** given at the last line of the provided **Temporal Label**. The **Response** should not skip over errors included in the **Fake Question**, nor should it state incorrect information. For example,
- **Fake Question**: "What was the reason behind Dobby enchanting the Sorting Hat to sing a warning to Harry at Hogwarts?"
- **Answer at the last line of the Temporal Label**: Dobby did not enchant the Sorting Hat to sing a warning; he sealed the gateway at King's Cross and tampered with the Bludger to protect Harry from danger at Hogwarts.
- **Wrong Response from 7th-year Harry Potter on Christmas**: "*It is because he was trying to protect me in his own, erm, unique way*. You know Dobby, always looking out for me, even if his methods are a bit unconventional. Maybe he heard something dangerous was going to happen at Hogwarts and figured a singing hat would be the best way to give me a heads up. Sounds like something straight out of one of his plans, doesn't it? Always with the best intentions, but you can never expect what he's going to do next."

5. Response Quality (Personality Consistency):
The **Response** should reflect the **Character**'s personalities in **Personality Label**, including their preferences, values, and convictions. Please note that the ratings can be *subjective*, varying from one evaluator to another. Here's a breakdown of the seven-point scale for clarity.
- **Not at All Reflective (1)**: The response shows no alignment with the character's known personality traits or behaviors.
- **Slightly Reflective (2)**: The response has minimal alignment, indicating very rare or barely perceptible characteristics of the character's personality.
- **Somewhat Reflective (3)**: There are noticeable elements that align with the character's personality, though inconsistencies are present.
- **Moderately Reflective (4)**: The response fairly represents the character's personality, with a balanced mix of consistent and inconsistent elements.
- **Quite Reflective (5)**: The response is largely in line with the character's personality, with only minor discrepancies.
- **Highly Reflective (6)**: The response closely aligns with the character's established personality traits, with very few exceptions.
- **Perfectly Reflective (7)**: The response is entirely in keeping with the character's known personality, without any deviations.

Figure 6: Instructions in the UI design of Amazon Mechanical Turk to collect human annotations for data quality.

- Christmas: [1-12, 2-12, 3-11, 4-23, 5-23, 6-16, 7-19]

- End of the scene: [1-17, 2-18, 3-22, 4-37, 5-38, 6-30, 7-36]

For the Lord of the Rings series, we select five main characters: **Frodo Baggins, Samwise Gamgee, Gandalf, Aragorn, and Legolas**. Differing from the approach taken with the Harry Potter series, we assign specific time points individually for each character due to the lack of recurring events throughout the books. In addition, we annotate the character's time point using the format {volume_number - book_number - chapter_number}, as follows:

- Frodo Baggins (Total 12 time points): at Bilbo Baggins's Farewell party (1-1-1), at the moment when Frodo was stabbed by one of the Ringwraiths (1-1-11), at the moment when The Fellowship loses Gandalf in Moria (1-2-5), end of Volume 1 (1-2-10), at the moment when encountering Gollum and decides to spare his life (2-4-1), at the moment of encountering the Black Gate of Mordor (2-4-3), at the moment when Frodo captured by Faramir (2-4-5), at Shelob's lair (2-4-9), end of Volume 2 (2-4-10), at the event when captured by Orcs at the Tower of Cirith Ungol (3-6-1), at the moment when rescued from Mount Doom by the eagles (3-6-4), end of Volume 3 (3-6-9).

Figure 7: An example of the UI design of Amazon Mechanical Turk to collect human annotations for data quality.

- Samwise Gamgee (Total 12 time points): at Bilbo Baggins's Farewell party (1-1-1), at the moment when Frodo was stabbed by one of the Ringwraiths (1-1-11), at the moment when The Fellowship loses Gandalf in Moria (1-2-5), end of Volume 1 (1-2-10), at the moment when encountering Gollum and decides to spare his life (2-4-1), at the moment of encountering the Black Gate of Mordor (2-4-3), at the moment when captured by Faramir (2-4-5), at Shelob's lair (2-4-9), end of Volume 2 (2-4-10), at the event when captured by Orcs at the Tower of Cirith Ungol (3-6-1), at the moment when rescued from Mount Doom by the eagles (3-6-4), end of Volume 3 (3-6-9).

- Gandalf (Total 12 time points): at Bilbo Baggins's Farewell party (1-1-1), at the moment when the Fellowship was formed at the council of Elrond (1-2-3), at the moment when The Fellowship loses Gandalf in Moria (1-2-5), end of Volume 1 (1-2-10), at the moment when Gandalf the white met Aragorn, Legolas, and Gimli at Fangorn Forest after the fall at Moria (2-3-5), at the moment when arriving at Isengard after the battle of Helm's Deep (2-3-8), at the moment when Gandalf met Saruman at Isengard (2-3-10), end of Volume 2 (2-4-10), at the moment when Gandalf arrived at Minas Tirith with Pippin (3-5-1), during the Battle of the Pelennor Fields (3-5-6), at the coronation of King Elessar(Aragorn) (3-6-5), end of Volume 3 (3-6-9).

- Aragorn (Total 12 time points): at the moment when Aragorn first met Frodo and his companions (1-1-9), at the moment when the Fellowship was formed at the council of Elrond (1-2-3), at the moment when The Fellowship loses Gandalf in Moria (1-2-5), end of Volume 1 (1-2-10), at the moment when Gandalf the white met Aragorn, Legolas, and Gimli at Fangorn Forest after the fall at Moria (2-3-5), at the moment when Aragorn arrived at Isengard after the battle of Helm's Deep (2-3-8), at the moment when Aragorn met Saruman at Isengard (2-3-10), end of Volume 2 (2-4-10), at the moment when Aragorn arrived at the paths of the dead (3-5-2), at the Battle of the Pelennor Fields (3-5-6), at the coronation of King Elessar(Aragorn) (3-6-5), end of Volume 3 (3-6-9).

- Legolas (Total 12 time points): at the moment when the Fellowship was formed at the council of Elrond (1-2-3), at the moment when The Fellowship loses Gandalf in Moria (1-2-5), at the moment of leaving Lothlórien (1-2-8), end of Volume 1 (1-2-10), at the moment when Leoglas met Gandalf the white at Fangorn Forest after Gandalf's fall at Moria (2-3-5), at the moment when Legolas arrived at Isengard after the battle of Helm's Deep (2-3-8), at the moment when Legolas met Saruman at Isengard (2-3-10), end of Volume 2 (2-4-10), at the moment when Legolas arrived at the paths of the dead with Aragorn (3-5-2), at the Battle of the Pelennor Fields (3-5-6), at the coronation of King Elessar(Aragorn) (3-6-5), end of Volume 3 (3-6-9).

For the Twilight series, we select three main characters: **Bella Swan, Edward Cullen, and Jacob Black**. We annotate each character's time point using the format {book_number - chapter_number}, as follows:

- Bella Swan (Total 16 time points): at the moment when Bella moved from Phoenix to Forks (1-1), at the moment when Bella first confirmed Edward's true nature as a vampire (1-9), at the moment when Bella first visited the Cullens (1-15), end of book 1 (1-25), on Bella's 18th birthday (2-1), at the moment when Bella jumps off the cliff into the ocean (2-15), at Volterra (2-20), end of book 2 (2-25), at the moment when Bella was grounded

by her father (3-1), at the moment when Bella learns about the history of the Quileute tribe and the Cullens (3-11), at the moment when Bella receives an engagement ring from Edward (3-20), end of book 3 (3-27), at Bella and Edward's Wedding (4-3), at the moment when Renesmee was born (4-18), at the moment when Bella forges passports and IDs for Renesmee and Jacob from J. Jenks (4-33), end of book 4 (4-39).

- Edward Cullen (Total 16 time points): at the moment when Edward saves Bella from a Van (1-3), at the moment when Bella first confirm Edward's true nature as a vampire (1-9), at the moment when Bella first visited the Cullens (1-15), end of book 1 (1-25), on Bella's 18th birthday (2-1), at the moment when Edward tells Bella that he and the Cullens are leaving Forks (2-3), at Volterra (2-20), end of book 2 (2-25), at the moment when Edward rewarded Alice for watching Bella by giving her the canary yellow Porsche from Italy (3-6), at Bella's graduation ceremony (3-16), at the moment when Jacob crawled into the sleeping bag beside Bella at the campsite, chosen for Bella's hiding place (3-22), end of book 3 (3-27), at Bella and Edward's Wedding (4-3), at the moment when Renesmee was born (4-18), at the moment when Esme has renovated a cottage on the property for Bella, Edward, and now Renesmee (4-24),end of book 4 (4-39).

- Jacob Black (Total 16 time points): at the beach at La Push, when Jacob met Bella and her friends (1-6), on March 10, 2005 (1-11), on March 13, 2005 when Bella found Jacob before watching a baseball game with Edward (1-17), end of book 1 (1-25), at the moment when Jacob and Bella worked together on repairing two old motorcycles (2-6), at the moment when Bella first discovers Jacob's werewolf identity (2-10), at the moment when Jacob pulled out Bella from drowning (2-16), end of book 2 (2-25), on May 31, 2006 when Bella found Jacob on his motorcycle at the school (3-16), at Bella's graduation ceremony (3-7), at the moment when Jacob crawled into the sleeping bag beside Bella at the campsite, chosen for Bella's hiding place (3-22), end of book 3 (3-27), at Bella and Edward's Wedding (4-3), at the moment when Renesmee

was born (4-18), at Christmas, 2006 (4-34), end of book 4 (4-39).

For the Hunger Games series, we select three main characters: **Katniss Everdeen, Peeta Mellark, and Gale Hawthorne**. We annotate each character's time point using the format {book_number - chapter_number}, as follows:

- Katniss Everdeen (Total 12 time points): at the moment when Katniss volunteered to take her sister's place as the female tribute (1-2), at the start of the 74th Hunger Games (1-11), at the moment when Katniss found wounded Peeta hidden under a layer of mud (1-19), end of book 1 (1-27), at the moment when they arrived at District 11 for the first stop of the Victory Tour (2-4), at the announcement of the Quarter Quell (2-12), at the moment when Katniss first witnessed a heavy fog during the Quarter Quell (2-20), end of book 2 (2-27), at the first conversation about the bombing of district 12 with Gale (3-1), at the moment when Peeta suddenly warned of an impending attack on District 13 (3-9), at the moment when the squad 451 was attacked by the mutts in the tunnels (3-22), end of book 3 (3-28).

- Peeta Mellark (Total 12 time points): at the moment when Katniss volunteered to take her sister's place as the female tribute (1-2), at the start of the 74th Hunger Games (1-11), at the moment when Katniss found wounded Peeta hidden under a layer of mud (1-19), end of book 1 (1-27), at the moment when they arrived at District 11 for the first stop of the Victory Tour (2-4), at the announcement of the Quarter Quell (2-12), at the moment when Katniss first witnessed a heavy fog during the Quarter Quell (2-20), end of book 2 (2-27), at the moment when Peeta suddenly warned of an impending attack on District 13 (3-9), at the moment when Peeta was sent as new member of squad 451 by president Coin (3-18), at the moment when the squad 451 was attacked by the mutts in the tunnels (3-22), end of book 3 (3-28).

- Gale Hawthorne (Total 12 time points): at the moment when Katniss volunteered to take her sister's place as the female tribute (1-2), at the start of the 74th Hunger Games (1-11), at the moment when Katniss found wounded Peeta

hidden under a layer of mud (1-19), end of book 1 (1-27), at the moment when Katniss delivered the animals she caught before the Victory Tour (2-1), at the announcement of the Quarter Quell (2-12), at the moment when Katniss first witnessed a heavy fog during the Quarter Quell (2-20), end of book 2 (2-27), at the first conversation about the bombing of district 12 with Katniss (3-1), at the moment when Peeta suddenly warned of an impending attack on District 13 (3-9), at the moment when the squad 451 was attacked by the mutts in the tunnels (3-22), end of book 3 (3-28).

G   Details of Experiments on TIMECHARA

G.1   Implementation Details

For generating responses, we apply nucleus sampling with $p = 1$ and temperature $\tau = 0.2$ across all role-playing LLMs. We cap the maximum token length at 2048 tokens. For GPT-4 as judges, we set $p = 0.95$, a temperature of $\tau = 0.0$, and a maximum token length of 1024 tokens. We use a single NVIDIA RTX A6000 GPU to generate responses with Mistral. For AlignScore evaluation, we use a single NVIDIA Quadro RTX 6000 GPU.

G.2   Few-Shot Method

The examples are carefully selected to represent a range of question types: *future*, *past-presence*, *past-absence*, and *past-only*, each paired with a corresponding response. The responses are generated by GPT-4 and are then manually checked for spatiotemporal and personality consistency. We then append these four examples, complete with their questions and correct responses, before the question.

G.3   Self-Refine Method

We incorporate a self-feedback prompt to facilitate this process, as outlined in Table 18. The model iteratively refines its response for a maximum of three iterations. It considers both spatiotemporal and personality scores, each with a maximum of 3 points. The response is finalized when a combined score of 5 or more is reached.

Prompt for Self-Feedback of Self-Refine Method
We want to iteratively improve the provided responses, mimicking the character {agent_name}. To help improve, scores for each response on desired traits are provided: 1) Spatiotemporal Consistency and 2) Personality Consistency.
*** [Interactions] Interviewer: {question} {agent_name}: {response}
[Evaluation Criterion] - Spatiotemporal Consistency (0 or 3): Is the response consistent with the character's spatiotemporal knowledge? If the response includes information that the character couldn't have known (either because it pertains to a future event or a past event they were not present for), assign a score of 0. If the response accurately reflects only the knowledge and events the character has experienced or been aware of, give a score of 3. - Personality Consistency (1 to 3): Is the response consistent with the character's personality? Use the given scale from 1-3 to rate how well the response reflects the personalities, including preferences, values, and convictions of the character. 1 being not at all reflective of the character's personalities, and 3 being perfectly reflective of the character's personalities. ***
1. Read through the [Interactions] and evaluate the spatiotemporal consistency: print the single-sentence rationale with the score on its own line corresponding to the correct answer. 2. Read through the [Interactions] and evaluate the personality consistency: print the single-sentence rationale with the score on its own line corresponding to the correct answer. 3. Print the total score.

Table 18: Prompt for generating self-feedback during response generation with self-refine method.

G.4 Retrieval-Augmented Generation (RAG) Method

We employ the LangChain framework[5] to implement the RAG method. We retrieve up to six paragraphs from a raw text source based on the given question. These retrieved paragraphs are then attached to the end of the question, serving as additional context for generating the response.

G.5 Decomposed Reasoning via NARRATIVE-EXPERTS

G.5.1 NARRATIVE-EXPERTS

We detail the prompts used for two narrative experts: a temporal expert (refer to Table 19) and a spatial expert (refer to Table 20). For the temporal expert, we compare the outputs numerically to the character's time point, presented in the 'book number - chapter number' format, to determine if the question pertains to the future relative to the character's timeline. In addition, we outline the complete algorithm that describes how the role-playing LLM generates a response based on a question and the hints provided by these two experts, as presented in Algorithm 1.

G.5.2 NARRATIVE-EXPERTS-RAG-CUTOFF

It can be effective to combine NARRATIVE-EXPERTS with RAG-cutoff methods, because the RAG-cutoff addresses *past-only* type questions, especially those that are fake-based, and NARRATIVE-EXPERTS is compatible with any baselines. For NARRATIVE-EXPERTS-RAG-CUTOFF, we incorporate a retrieval module to aug-

ment the capabilities of both the temporal and spatial experts. Specifically, we retrieve up to six paragraphs from a raw text source based on the input question. These paragraphs are subsequently incorporated into the prompts for the two narrative experts, as detailed in Table 21 and Table 22. The role-playing LLM then generates an answer using the question, hints from both experts and the paragraphs corresponding to events before the character's defined period. Note that we utilize all six paragraphs to assist the narrative experts (adopting a naive RAG approach rather than RAG-cutoff), but we limit the paragraphs when producing the final response. Furthermore, we exclude paragraphs in the final response if the temporal expert predicts 'future' events since it slightly enhances spatiotemporal consistency performance. The algorithm for NARRATIVE-EXPERTS-RAG-CUTOFF is detailed in Algorithm 2.

G.6 Experimental Results on the 11K Dataset

While Table 5 presents experimental results from a randomly sampled set of 600 data instances, we additionally conduct experiments on the entire dataset comprising 10,895 instances.

Baseline methods. We utilize two different open-source LLMs as the backbone models for our role-playing agents: Mistral 7B Instruct (i.e., `mistral-7b-instruct-v0.2`) (Jiang et al., 2023) and Llama-2 Chat 13B (Touvron et al., 2023). Note that we exclude GPT-4 Turbo and GPT-3.5 Turbo from our backbone models due to their high costs. Following § 5.2, we implement zero-shot prompting and RAG-cutoff as our baseline methods.

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

Prompt for Book and Chapter Identification of NARRATIVE-EXPERTS Method

You will be given a question from {series_name} series at a specific time. Your task is to identify the exact {book_chapter_name} of the scene of the question. Below is the data:

***
[Question]
{question}
***
[Identification Criterion]
What is the exact {book_chapter_name} of the scene of the question?

1. Read through the [Question], recall the scene from the question, and describe it using the six Ws (Who, What, When, Where, Why, and How).
2. Identify the exact {book_chapter_name} of the scene of the question, in '{book_chapter_format}' format.

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then, print the output on its own line corresponding to the correct answer. At the end, repeat just the selected output again by itself on a new line.

Table 19: Prompt for identifying the exact book and chapter of the scene in the question.

Prompt for Event Participants Identification of NARRATIVE-EXPERTS Method

You will be given a question and a character from {series_name} series. Your task is to classify whether the character is a participant (i.e., present or absent) in the scene of the question. Below is the data:

***
[Question]
{question}
[Character]
{character}
***
[Classification Criterion]
Is the character a participant in the scene of the question?

[Classification Steps]
1. Read through the [Question], recall the scene from the question, and describe it using the six Ws (Who, What, When, Where, Why, and How).
2. Identify the exact {book_chapter_name} of the scene of the question.
3. Write a list of every character involved in the scene described in the question, including those not explicitly mentioned in the question but who were present in the scene.
4. Compare the list of participants to the character. Check if the list of participants contains the character.
5. If the list contains the character, classify it as 'present'. Otherwise, classify it as 'absent'.
***

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then, print the output on its own line corresponding to the correct answer. At the end, repeat just the selected output again by itself on a new line.

Table 20: Prompt for identifying whether the given character is a participant in the scene of the question.

Prompt for Book and Chapter Identification of NARRATIVE-EXPERTS-RAG-CUTOFF Method

You will be given a question and contexts from {series_name} series at a specific time. Your task is to identify the exact {book_chapter_name} of the scene of the question. Below is the data:

***
[Question]
{question}
***
[Contexts]
{contexts}
***
[Identification Criterion]
What is the exact {book_chapter_name} of the scene of the question?

1. Read through the [Question] and [Contexts], recall the scene from the question, and describe it using the six Ws (Who, What, When, Where, Why, and How).
2. Identify the exact {book_chapter_name} of the scene of the question, in '{book_chapter_format}' format.

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then, print the output on its own line corresponding to the correct answer. At the end, repeat just the selected output again by itself on a new line.

Table 21: Prompt for identifying the exact book and chapter of the scene in the question while using a retrieval module.

Prompt for Event Participants Identification of NARRATIVE-EXPERTS-RAG-CUTOFF Method

You will be given a question, a character, and contexts from {series_name} series. Your task is to classify whether the character is a participant (i.e., present or absent) in the scene of the question. Below is the data:

***
[Question]
{question}
[Character]
{character}
***
[Contexts]
{context}
***
[Classification Criterion]
Is the character a participant in the scene of the question?

[Classification Steps]
1. Read through the [Question] and [Contexts], recall the scene from the question, and describe it using the six Ws (Who, What, When, Where, Why, and How).
2. Identify the exact {book_chapter_name} of the scene of the question.
3. Write a list of every character involved in the scene described in the question, including those not explicitly mentioned in the question but who were present in the scene.
4. Compare the list of participants to the character. Check if the list of participants contains the character.
5. If the list contains the character, classify it as 'present'. Otherwise, classify it as 'absent'.
***

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then, print the output on its own line corresponding to the correct answer. At the end, repeat just the selected output again by itself on a new line.

Table 22: Prompt for identifying whether the given character is a participant in the scene of the question while using a retrieval module.

Evaluation metrics. Evaluating spatiotemporal or personality consistency with GPT-4 judges incurs an extremely high cost. As a result, we utilize AlignScore (Zha et al., 2023) for assessing factual inconsistencies across scenarios such as natural language inference and fact verification tasks. Specifically, we employ the RoBERTa-large model equipped with a 3-way classification head. Given the gold response and the predicted response from the role-playing agent, the RoBERTa model produces an AlignScore ranging from 0 to 1, which we then average.

Experimental results. Experimental results are presented in Table 23. As expected, methods based on Mistral achieve higher Align-Scores than those based on Llama 2, despite the smaller model size. NARRATIVE-EXPERTS and NARRATIVE-EXPERTS-RAG-CUTOFF outperform baseline methods across both Llama 2 and Mistral models. Note that the performance trend of the Mistral-based method mirrors that of the main experiment in § 5 (Table 5), with scores of 19.68, 17.94, 21.85, and 22.34 for zero-shot, rag-cutoff, NARRATIVE-EXPERTS, and NARRATIVE-EXPERTS-RAG-CUTOFF, respectively, in Table 23, compared to 18.50, 17.82, 20.57, 22.20 in Table 5. Although the AlignScore may not be entirely interpretable or reliable, our results indicate that employing our methods enhances spatiotemporal consistency across the entire dataset.

H   Further Analyses

H.1   Human Evaluation on Assessing Spatiotemporal Consistency

We supplement our findings with results from manual evaluations by human judges via Amazon Mechanical Turk on a subset of the dataset used in Table 5. To specify, we randomly selected 50 instances out of 600, ensuring an even distribution across the four data types (i.e., *future*, *past-absence*, *past-presence*, and *past-only* types). We then applied four methods (i.e., zero-shot, RAG-cutoff, narrative-experts, narrative-experts-RAG-cutoff) based on GPT-4 to generate four different responses for each instance. Subsequently, we tasked annotators with evaluating the spatiotemporal consistency of each response using a binary scale: 0 for inconsistency and 1 for consistency. Two annotators evaluated each response, and we calculated the average score by dividing the sum of the scores by 2. Specifically, we opted to assemble a new pool of annotators rather than rely on those who had previously worked on Appendix E.2. This is because we observed that evaluating 'predicted' responses posed more challenges than assessing gold responses. To be similar to Appendix E.2, we collected qualified human annotators via two-step qualification steps, ultimately forming a team of 27 annotators ready to proceed with the main HITs.

The results in Table 24 reveal a closely aligned trend between the two evaluation approaches (i.e.,

Algorithm 1: Decomposed Reasoning Method for Role-Playing LLM

Input: A question regarding a scene, a character with their time point
Output: A response
Function `TemporalExpert`(*question*)
  Identify the book and chapter of the scene of the question;
  if *scene is in the future* **then**
    return *"future", "Note that the period of the question is in the future relative to {character}'s time point. Therefore, you should not answer the question or mention any facts that occurred after {character}'s time point.";*
  end
  return *"past", "";*
end
Function `SpatialExpert`(*question, character*)
  Determine if character participates in the scene of the question;
  if *character is past-absent* **then**
    return *"Note that {character} had not participated in the scene described in the question. Therefore, you should not imply that {character} was present in the scene.";*
  end
  return *"";*
end
Function `RolePlayingLLM`(*question, hints*)
  Append hints to the prompt;
  Generate response based on enhanced prompt;
  return *response*;
end

Procedure `MainProcedure`(*question, character*)
  `// Invoke Temporal Expert`
  temporalStatus, temporalHint := TemporalExpert(*question*);
  spatialHint := "";
  if *temporalStatus == "past"* **then**
    `// Invoke Spatial Expert only if past`
    spatialHint := SpatialExpert(*question, character*);
  end
  `// Prepare hints`
  hints := temporalHint + spatialHint;
  `// Invoke Role-Playing LLM with hints`
  response := RolePlayingLLM(*question, hints*);
  return *response*;
end

GPT-4 evaluation and human evaluation) across methods and data types, with a marginal gap. In addition, we calculated Gwet's AC1 scores between two annotators, obtaining a value of **0.91**. We also highlight the agreement ratios between annotators who chose the same versus different labels: **0.94** vs. 0.06 for GPT-4 zero-shot, **0.94** vs. 0.06 for GPT-4 RAG-cutoff, **0.96** vs. 0.04 for GPT-4 narrative-experts, **0.94** vs. 0.06 for GPT-4, and narrative-experts-RAG-cutoff. This further step emphasizes our commitment to ensuring the reliability and validity of both the automatic evaluation process and our experimental findings.

H.2 Fine-grained Hallucination Detection

While the spatiotemporal consistency in Table 5 presents only accuracy, GPT-4 judges also provide rationales for their decision. To thoroughly analyze point-in-time hallucinations in the role-playing LLM, we first use four types of hallucination type from Mishra et al. (2024) and additionally define four types.

1. Entity: Incorrect facts about a person, place, or thing.

2. Relation: Wrong connections between characters or objects.

3. Contradictory: A statement that goes against known facts.

4. Invented: Making up facts or details that don't exist.

5. Temporal error (new): Getting the timing of events wrong, applicable to *future* type instances.

6. Spatial error (new): Incorrect event participants reasoning, relevant for *past-absence* and *past-presence* type instances.

7. Unclarified (new): Not correcting a wrong question, specifically not addressing false information in fake-based questions.

8. Incomplete (new): Omission of parts of an answer, especially when responding to *free-form* questions.

Then, we collect and manually annotate 199 instances where GPT-4 Turbo's responses are incorrect.

Algorithm 2: Decomposed Reasoning with RAG-cutoff

Input: A question regarding a scene, a character
Output: A response
Function RetrieveParagraphs(*question*)
   Retrieve up to six paragraphs based on the question from raw text source;
   return *paragraphs*;
end
Function LimitParagraphsBeforeCharacterPeriod(*paragraphs, character*)
   Filter paragraphs to only include those corresponding to events before the character's period;
   return *filteredParagraphs*;
end
Function TemporalExpertWithRAG(*question, character, paragraphs*)
   Incorporate paragraphs into the prompt to identify the book and chapter of the scene of the question;
   if *scene is in the future* **then**
      return *"future", "Note that the period of the question is in the future relative to {character}'s time point. Therefore, you should not answer the question or mention any facts that occurred after {character}'s time point.";*
   end
   return *"past", "";*
end
Function SpatialExpertWithRAG(*question, character, paragraphs*)
   Incorporate paragraphs into the prompt to determine if character participates in the scene of the question;
   if *character is past-absent* **then**
      return *"Note that {character} had not participated in the scene described in the question. Therefore, you should not imply that {character} was present in the scene.";*
   end
   return *"";*
end
Function RolePlayingLLMWithRAGCutoff(*question, hints, paragraphs, character, temporalStatus*)
   filteredParagraphs := ;
   if *temporalStatus == "future"* **then**
      filteredParagraphs := "";
   else
      filteredParagraphs := LimitParagraphsBeforeCharacterPeriod(*paragraphs, character*);
   end
   Append hints and filtered paragraphs to the prompt;
   Generate response based on the enhanced prompt;
   return *response*;
end

Procedure MainProcedureWithRAGCutoff(*question, character*)
   // Retrieve relevant paragraphs
   paragraphs := RetrieveParagraphs(*question*);
   // Invoke Temporal Expert with RAG
   temporalStatus, temporalHint := TemporalExpertWithRAG(*question, character, paragraphs*);
   spatialHint := "";
   if *temporalStatus == "past"* **then**
      // Invoke Spatial Expert with RAG only if past
      spatialHint := SpatialExpertWithRAG(*question, character, paragraphs*);
   end
   // Prepare hints and paragraphs
   hints := temporalHint + spatialHint;
   // Invoke Role-Playing LLM with RAG-cutoff
   response := RolePlayingLLMWithRAGCutoff(*question, hints, paragraphs, character, temporalStatus*);
   return *response*;
end

Method	AlignScore ↑
Llama-2 Chat 13B	
zero-shot	12.81±0.11
RAG-cutoff	12.22±0.11
narrative-experts (**Ours**)	**13.45±0.11**
narrative-experts-RAG-cutoff (**Ours**)	13.16±0.11
Mistral Instruct 7B (`mistral-7b-instruct-v0.2`)	
zero-shot	19.68±0.15
RAG-cutoff	17.94±0.16
narrative-experts (**Ours**)	21.85±0.17
narrative-experts-RAG-cutoff (**Ours**)	**22.34±0.18**

Table 23: Sub-experiment on all 11K data instances. We report the average scores with their standard error of the mean (SEM). A **bold** number indicates the highest average score, while an underline number denotes the second-best average score.

Table 25 shows a fine-grained hallucination-type distribution of 199 responses. As anticipated, 'Temporal Error' was the most common type, primarily because all the incorrect examples from the *future* type data instances fell under Temporal Error. On the other hand, most errors in the *past-absence* and *past-presence* were classified as 'Spatial Errors'. The distribution for *past-only* was the most balanced, with a notable number of cases falling into the 'Entity' category. This trend is largely due to instances where role-playing agents failed to detect swapped entities in fake-based questions.

H.3 Experimental Results per Novel Series

Table 26 shows the spatiotemporal consistency of the GPT-4 Turbo zero-shot method across different novel series, highlighting distinct performance trends. Specifically, the model tends to produce incorrect responses more often for the *future* and *past-absence* types within the Harry Potter series. In contrast, it exhibits lower performance for the *past-presence* type but performs better in other types within the Lord of the Rings series. Across all series, a common challenge is the model's struggle to accurately respond to instances of both *future* and *past-only (fake)* types.

H.4 Performance for Fact-Based vs. Fake-Based Questions

As shown in Table 26 and further supported by additional experimental results in Table 27, LLMs indeed do better at fact-based questions than fake-based ones, as expected. Although fake-based questions might not seem natural within a role-playing scenario, they are crucial from a fact-checking perspective to evaluate the agent's capability of avoiding point-in-time character hallucinations. Therefore, the inability of the agent to appropriately respond to fake-based questions remains a significant concern.

H.5 Accuracy of Narrative Experts

While we show the effectiveness of NARRATIVE-EXPERTS and NARRATIVE-EXPERTS-RAG-CUTOFF, we further analyze the performance of each expert. To evaluate the temporal expert, we use all 600 data instances used in Table 5, since all data instances are classified as either future or past. On the other hand, we only use 200 *fact-based structured* data instances from *past-absence* and *past-presence* types to evaluate the spatial expert.

Table 28 demonstrates the accuracy of narrative experts using three backbone LLMs.

NARRATIVE-EXPERTS *vs.* **NARRATIVE-EXPERTS-RAG-CUTOFF.** Despite lacking access to external knowledge, narrative experts significantly outperform random selection in classifying past/future and presence/absence scenarios. This indicates that LLMs rely on their parametric memories for spatiotemporal reasoning. Moreover, NARRATIVE-EXPERTS-RAG-CUTOFF outperforms NARRATIVE-EXPERTS on average, highlighting the benefits of incorporating retrieval modules.

Comparison of backbone models. As expected, narrative experts using GPT-4 Turbo achieve the highest accuracy on average. While the accuracy of temporal experts with both GPT-4 Turbo and GPT-3.5 Turbo in predicting future events exceeds 75%, the temporal expert with Mistral faces challenges in future predictions. Conversely, Mistral surpasses GPT-3.5 Turbo in predicting past events and absences.

Instruction following capability. The temporal expert of GPT-3.5 Turbo NARRATIVE-EXPERTS-RAG-CUTOFF shows an impressive 89.5% accuracy in detecting future events. However, its performance on the *future* type in Table 5 reveals that the GPT-3.5 Turbo model, when acting as a role-playing agent, often fails to follow provided instructions. Furthermore, we demonstrate that the underperformance of NARRATIVE-EXPERTS and NARRATIVE-EXPERTS-RAG-CUTOFF in the *past-presence* type, as shown in Table 5, is often due to the temporal and spatial experts' occasional failures to predict past events and presences accurately.

Method	Spatiotemporal Consistency (**GPT-4 Evaluation**, %) ↑					Spatiotemporal Consistency (**Human Evaluation**, %) ↑				
	Future	Past-absence	Past-presence	Past-only	Avg.	Future	Past-absence	Past-presence	Past-only	Avg.
GPT-4 Turbo (gpt-4-1106-preview)										
zero-shot	4.0 / 12	11.0 / 12	13.0 / 13	5.0 / 13	33.0 / 50	4.5 / 12	9.0 / 12	13.0 / 13	5.0 / 13	31.5 / 50
RAG-cutoff	6.0 / 12	11.0 / 12	13.0 / 13	5.0 / 13	35.0 / 50	6.0 / 12	10.5 / 12	13.0 / 13	5.0 / 13	34.5 / 50
narrative-experts	11.0 / 12	12.0 / 12	13.0 / 13	6.0 / 13	42.0 / 50	10.0 / 12	11.0 / 12	13.0 / 13	6.0 / 13	40.0 / 50
narrative-experts-RAG-cutoff	12.0 / 12	12.0 / 12	12.0 / 13	6.0 / 13	42.0 / 50	10.5 / 12	12.0 / 12	12.0 / 13	6.0 / 13	40.5 / 50

Table 24: Spatiotemporal consistency for 50 sampled data instances, evaluated by both GPT-4 and human judges.

Hallucination Type	Future	Past-absence	Past-presence	Past-only	Overall
Entity	0/103	0/22	0/8	**24**/66	24/199
Relation	0/103	0/22	0/8	2/66	2/199
Contradictory	0/103	0/22	1/8	15/66	16/199
Invented	0/103	0/22	0/8	14/66	14/199
Temporal Error	**103**/103	0/22	0/8	0/66	**103**/199
Spatial Error	0/103	**22**/22	**7**/8	0/66	29/199
Unclarified	0/103	0/22	0/8	9/66	9/199
Incomplete	0/103	0/22	0/8	2/66	2/199

Table 25: Fine-grained hallucination-type distribution from GPT-4 Turbo's zero-shot responses.

Novel Series	Future	Past-absence	Past-presence	Past-only (Fact)	Past-only (Fake)
Harry Potter	19/50	16/25	24/25	15/25	6/25
The Lord of the Rings	27/50	21/25	21/25	23/25	17/25
Twilight	23/50	19/25	23/25	18/25	5/25
The Hunger Games	24/50	19/25	22/25	18/25	16/25

Table 26: Spatiotemporal consistency of GPT-4 Turbo zero-shot per novel series.

Method	Past-only (Fact)	Past-only (Fake)
GPT-4 Turbo (gpt-4-1106-preview)		
zero-shot	74.0	44.0
zero-shot-cot	78.0	44.0
few-shot	77.0	57.0
self-refine	82.0	48.0
RAG	**90.0**	54.0
RAG-cutoff	84.0	60.0
NARRATIVE-EXPERTS	79.0	56.0
NARRATIVE-EXPERTS-RAG-CUTOFF	85.0	**64.0**

Table 27: Spatiotemporal consistency of GPT-4 Turbo zero-shot per fact-based vs. fake-based questions.

Method	Temporal Expert		Spatial Expert	
	Future	Past	Absence	Presence
Random	50.0	50.0	50.0	50.0
Mistral Instruct 7B (mistral-7b-instruct-v0.2)				
NARRATIVE-EXPERTS	30.5	81.0	70.0	87.0
NARRATIVE-EXPERTS-RAG-CUTOFF	63.5	89.3	76.0	88.0
GPT-3.5 Turbo (gpt-3.5-turbo-1106)				
NARRATIVE-EXPERTS	75.5	79.8	68.0	90.0
NARRATIVE-EXPERTS-RAG-CUTOFF	**89.5**	76.8	68.0	91.0
GPT-4 Turbo (gpt-4-1106-preview)				
NARRATIVE-EXPERTS	83.5	**94.5**	76.0	**98.0**
NARRATIVE-EXPERTS-RAG-CUTOFF	85.0	94.3	**84.0**	96.0

Table 28: Accuracy of temporal and spatial experts.

Character	6 years	5 years	4 years	3 years	2 years	1 years	0 year
GPT-3.5 Turbo (gpt-3.5-turbo-1106) zero-shot							
Harry Potter	25.0	25.0	0.0	12.5	12.5	12.5	0.0
Hermione Granger	71.4	57.1	42.9	42.9	42.9	42.9	14.3
Total	46.7	40.0	20.0	26.7	26.7	26.7	6.7
GPT-4 Turbo (gpt-4-1106-preview) zero-shot							
Harry Potter	100.0	100.0	100.0	100.0	100.0	100.0	25.0
Hermione Granger	100.0	100.0	100.0	100.0	85.7	85.7	71.4
Total	100.0	100.0	100.0	93.3	93.3	93.3	46.7

Table 29: Spatiotemporal consistency across *future* type instances, divided by the temporal gap between the role-playing character's time point and the question's time point.

Consequently, this leads to the role-playing LLMs following 'incorrect' instructions.

H.6 Evaluating Role-Playing LLM Agents on Time Points More Distant in the Future

While we presented Figure 1 as a motivational example of TIMECHARA, we observed that the challenge of detecting character hallucination, particularly with *future* type questions, diminishes as the temporal gap (i.e., the gap between the character's time point and the question's time point) widens. Conversely, it remains challenging for LLMs to identify hallucinations when the temporal difference narrows, as illustrated in Table 5. To empirically validate this observation, we designed an experiment as follows: We selected 15 questions from the latter part of "Harry Potter and the Deathly Hallows" (book 7), allocating them between Harry Potter and Hermione Granger - 8 for Harry and 7 for Hermione. We then positioned the characters at seven distinct time points (i.e., the start of each school year from year 1 to year 7), resulting in a total of $15 * 7 = 105$ unique *future* type instances. Note that this scalable approach enabled us to generate examples that assess models at time points distant in the future.

The Table 29 presents the spatiotemporal consistency across 105 instances, illustrating the effect of the temporal gap between the character's time point and the question's time point on model performance. According to the results, LLMs (with zero-

shot prompt) exhibit an increase in spatiotemporal consistency as the temporal distance between the character's time point and the question's time point widens. Specifically, 'GPT-3.5 zero-shot' showed a maximum of 46.7% spatiotemporal consistency, indicating persistent hallucinations, while 'GPT-4 zero-shot' achieved 100%. This suggests that role-playing LLMs, especially GPT-4, with a higher memorization capability, are able to recognize this extent of temporal difference as the temporal gap between the character and the question expands in *future* type instances. However, it remains challenging for LLMs to maintain spatiotemporal consistency when the temporal distance between the character's time point and the question's time point decreases (e.g., GPT-3.5: 6.7%, GPT-4: 46.7%), as also depicted in Table 5.