Character is Destiny: Can Role-Playing Language Agents Make Persona-Driven Decisions?

Rui Xu[1], **Xintao Wang**[1], **Jiangjie Chen**[1], **Siyu Yuan**[1], **Xinfeng Yuan**[1],
Jiaqing Liang[1], **Zulong Chen**[2], **Xiaoqing Dong**[2], **Yanghua Xiao**[1]
[1]Fudan University [2]Alibaba Group
{ruixu21, xtwang21, syyuan21, xfyuan23}@m.fudan.edu.cn
{jjchen19, liangjiaqing, shawyh}@fudan.edu.cn
{zulong.czl, xiaoqing.dongxq}@alibaba-inc.com

Abstract

Can Large Language Models (LLMs) simulate humans in making important decisions? Recent research has unveiled the potential of using LLMs to develop role-playing language agents (RPLAs), mimicking mainly the knowledge and tones of various characters. However, imitative decision-making necessitates a more nuanced understanding of personas. In this paper, we benchmark the ability of LLMs in persona-driven decision-making. Specifically, we investigate whether LLMs can predict characters' decisions provided by the preceding stories in high-quality novels. Leveraging character analyses written by literary experts, we construct a dataset LIFECHOICE comprising 2,512 characters' decision points from 470 books. Then, we conduct comprehensive experiments on LIFECHOICE, with various LLMs and RPLA methodologies. The results demonstrate that state-of-the-art LLMs exhibit promising capabilities in this task, yet substantial room for improvement remains. Hence, we further propose the CHARMAP method, which adopts persona-based memory retrieval and significantly advances RPLAs on this task. Resources are available at `https://github.com/airaer1998/LifeChoice`.

1 Introduction

Every man has but one destiny.
– *The Godfather*. Mario Puzo.

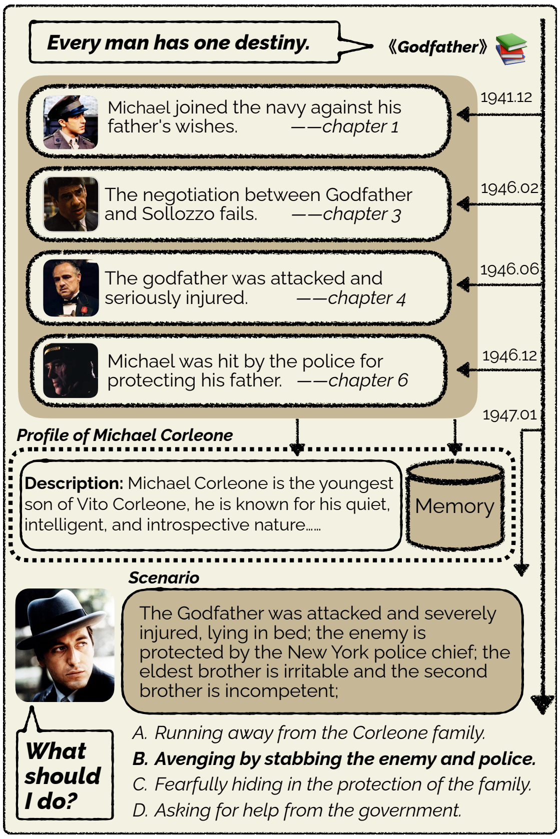

Every man has one destiny.
《Godfather》
Michael joined the navy against his father's wishes. ——chapter 1
1941.12
The negotiation between Godfather and Sollozzo fails. ——chapter 3
1946.02
The godfather was attacked and seriously injured. ——chapter 4
1946.06
Michael was hit by the police for protecting his father. ——chapter 6
1946.12
1947.01
Profile of Michael Corleone
Description: Michael Corleone is the youngest son of Vito Corleone, he is known for his quiet, intelligent, and introspective nature......
Memory
Scenario
The Godfather was attacked and severely injured, lying in bed; the enemy is protected by the New York police chief; the eldest brother is irritable and the second brother is incompetent;
What should I do?
A. Running away from the Corleone family.
B. Avenging by stabbing the enemy and police.
C. Fearfully hiding in the protection of the family.
D. Asking for help from the government.

Figure 1: An example of LIFECHOICE. Given a character, a decision point and the preceding context, RPLAs are expected to reproduce the original decision. Typically, RPLAs are constructed by parsing the context into the character's description and memory.

With the recent advancements in large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023), Role-Playing Language Agents (RPLAs) have emerged as a flourishing field of AI applications and research (Chen et al., 2024). RPLAs are LLM-based AI systems that simulate assigned personas, reproducing their tones, knowledge, personalities, and even decisions (Park et al., 2023; Gao et al., 2024; Wang et al., 2024; Xie et al., 2024). They emulate various characters across extensive applications, including fictional characters in chatbots and video games (Wang et al., 2023, 2024), as well as digital clones (Gao et al., 2023) or personalized assistants (Xu et al., 2022; Salemi et al., 2024) for real-world individuals.

Can RPLAs reliably make decisions that align with their personas, as humans do? This question is vital for the practical usage of RPLAs, yet remains underexplored. Previous studies primarily investigate RPLAs' character fidelity in terms of their

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15038–15059
November 4-9, 2025 ©2025 Association for Computational Linguistics

tones (Wang et al., 2023) and knowledge (Shao et al., 2023), which could be readily replicated by existing RPLAs via style imitation and knowledge retrieval. However, these features are relatively superficial compared with the underlying thinking and mindset of characters. Recent efforts (Wang et al., 2024) study the personality fidelity of RPLAs, but they fail to capture the nuances and dynamics of characters' mindsets. Hence, it remains an understudied question whether RPLAs could simulate persona-driven decisions, which challenges their comprehensive understanding of the personas and reasoning about unobserved behaviors.

In this paper, we systematically study the capability of RPLAs to simulate persona-driven decisions, based on characters from high-quality novels. In high-quality novels, characters' life choices are carefully plotted and aligned with their personas. Hence, we introduce the LIFECHOICE dataset, which evaluate whether RPLAs can faithfully reproduce the characters' life choices in the narratives. Specifically, LIFECHOICE comprises 2,512 character decisions from 470 novels, leveraging expert-written character analyses. Each sample is presented as a multiple-choice question with the preceding context before the decision point. As depicted in Figure 1, RPLAs are expected to identify and reason over relevant knowledge about the characters to simulate their decisions. The construction of LIFECHOICE primarily involves three steps: decision point selection, multiple-choice question construction, and manual examination.

Compared with previous methods for RPLA evaluation, our task and dataset benefit from higher-quality data and are more challenging. First, compared to most role-playing benchmarks that use LLMs as judges (Wang et al., 2023; Zhou et al., 2023), our questions and decisions are derived from the original books and can be seen as ground truth. Second, our task is more challenging as it requires RPLAs to comprehensively understand and reason based on the personas, including their knowledge, experiences, and personalities. Specifically, LIFECHOICE poses the following challenges: 1) *Long-context understanding*, where RPLAs' decisions are related to their entire long context, requires identifying sparse relevant motivations from massive character contexts. 2) *Temporal intelligence*, where RPLAs should intelligently adapt to the dynamic evolution of characters and environments. 3) *Intricate motives*, where RPLAs are required to reason through complex and entangled backgrounds and motives to arrive at the decisions.

To evaluate RPLAs on the LIFECHOICE, we conduct extensive experiments. These experiments encompass a variety of LLMs and RPLA frameworks, including RPLAs based on description, memory, a combination of both, and our proposed method CHARMAP, which enhances memory retrieval through character storylines. Our findings indicate that current RPLAs achieve a maximum accuracy of up to 61.12% on LIFECHOICE, which underscores the task's challenge. Notably, CHARMAP framework significantly advances RPLA performance on this task, reaching an accuracy of 65.28% and outperforming previous baselines by 4.16%. However, a considerable gap remains when compared to human performance (92.01%), highlighting substantial room for improvement. Furthermore, our observations underscore the critical importance of well-summarized character descriptions and accurate memory retrieval for effective RPLAs.

In summary, our contributions include:

- We propose to explore RPLAs' ability in simulating persona-driven decisions, which is crucial for future RPLA applications and challenges existing RPLAs.
- We delicately craft LIFECHOICE, the first benchmark for persona-driven decisions of RPLAs, based on characters' life choices from high-quality novels. Besides, we propose CHARMAP, which adopts persona-based memory retrieval for better decision-making of RPLAs.
- Based on LIFECHOICE, we conduct extensive experiments. The results demonstrate the promising performance of RPLAs in decision simulation. Then, we analyze and compare methodologies for RPLA development, and show the effectiveness of CHARMAP.

2 Related Work

Character Role-Playing Research on character-related studies evolves through multiple stages. Initial studies focus on character understanding, where models demonstrate comprehension through character recognition and personality prediction in context (Chen and Choi, 2016; Sang et al., 2022; Xu et al., 2025). While Yuan et al. (2024) analyzes character motivations using LLMs, their approach treats motivation analysis as an information extraction task from character profiles, rather than reasoning about events not explicitly stated in the profiles. Current research advances to character role-playing (Chen et al., 2024), where LLMs simulate character traits through dialogue and behavior. This simulation includes replicating personalities, speaking styles, and character knowledge (Shao et al., 2023; Xu et al., 2024). Although instruction fine-tuning and memory retrieval techniques show success in character recreation (Zhou et al., 2023), our research extends beyond these methods. We evaluate RPLAs through the lens of behavior and decision-making, which requires models to perform complex reasoning based on memory. This approach presents new challenges in character understanding and portrayal.

Personal LLM assistants With the rapid development of artificial intelligence technology, there are now many personal intelligent agents embedded in mobile devices, providing personalized services through analyzing user data and equipment (Kaplan and Haenlein, 2019; Hoy, 2018).These agents model user profiles and preferences through historical data (Gurrin et al., 2014; Dodge and Kitchin, 2007), such as extracting personality from text (Majumder et al., 2017; Štajner and Yenikent, 2020), detecting emotions from images (Jaiswal et al., 2020; Zad et al., 2021), and analyzing interaction patterns (Tang et al., 2019; Li et al., 2018). While these capabilities enhance decision-making and user experience, collecting real user data raises privacy concerns. We address this challenge by modeling characters from novels, creating a benchmark where models predict decisions based on story context.

3 Dataset and Task Setups

3.1 Dataset Construction

We construct a comprehensive dataset called LIFECHOICE. As shown in Table 1, the sample for each decision point includes the preceding context p from the original book, the current scenario s, a question q outlining a decision faced by that character c, a list of options $a = \{a_i\}_{i=1}^4$, the correct answer y, and the motivation m explaining the character's choice. Our data is sourced from the website *Supersummary*[1], which provides three pieces of content written by literary experts: *key character descriptions*, *chapter summaries*, and *book analyses*. We contact the website and obtain authorization to use the data for academic research. The dataset construction comprises the following three main steps:

Book: Les Misérables
Character: Jean Valjean
Context:
In 1815 Monsieur Charles-François-Bienvenu Myriel was Bishop of Digne. He was then......Jean Valjean reflections gave him a sort of frightening aspect. He was subject to one of those violent inner tearings, which was not unknown to him.
Scenario:
In the courtroom, an innocent man was wrongfully accused, because he bore a resemblance to Jean Valjean. If Jean Valjean did not come forward, this innocent man would be sent to the gallows in his place. At this time, Jean Valjean had transformed his identity and become a respected town mayor, and he had also adopted a young girl named Cosette, with whom he had a new life.
Question:
You will play the role of Jean Valjean. What will you choose to do when you discover that man is about to be convicted due to being mistaken for you?
Options:
A. Keep silent, letting an innocent person take the punishment in one's place.
B. Persuade the person to run away, in order to protect both from the disaster of jail.
C. Go to court and reveal the truth, sacrificing oneself to save the innocent person.
D. Look for legal loopholes, trying to save both the person and oneself.

Correct Answer: C
Motivation:
[Values and Beliefs] Jean Valjean is a person who values honesty and justice, possessing a strong sense of morality and righteousness. He decides to turn himself in to save another innocent person, fulfilling his inner need for morality and justice.

Table 1: Case study of LIFECHOICE. A complete set of data includes book, character, scenario, question, options, correct answer, motivation, and input.

Selecting Decision Points To prevent data leakage, we first filter novels on the site using the following criteria: (1) The narrative must exclude non-fiction genres like biographies or documentary literature. (2) The narrative perspective must be in the first or third person. (3) The progression of narrative time should be linear, avoiding stories with complex timelines or flashbacks. (4) Exclude overly popular books, as measured by a high number of reviews on literary review websites. For each book that passes these filters, we provide GPT-4 with content written by literary experts, requesting it to output each key character's life choice decision points and their corresponding gold motivations. Additionally, we also ask GPT-4 to output the chap-

[1] https://www.supersummary.com/

Dataset	Source	Context Length	Task Format	Has Explanation
TVSHOWGUESS	TV show transcripts	∼50k	Character Identification	✗
ROCStories	Commonsense short stories	∼100	Character Behavior Prediction	✗
LiSCU	Literature	∼1000	Character Identification	✗
LifeChoice	Literature	∼150k	Character Behavior Prediction	✓

Table 2: Comparison between LifeChoice and previous character understanding benchmarks: data source, context length, task format, and whether the benchmark has explanations.

ter numbers related to the decision based on the extracted motivations. As shown in the example in Figure 1, the literary expert's analysis of the book suggests that *Michael Corleone*'s motivation for choosing to assassinate the enemy includes both avenging his father and witnessing the collusion between the police and the enemy, which exposes him to the darker side of the government. We then identify two corresponding chapters in the original book based on these motivations, providing more refined data for constructing multiple-choice questions.

Constructing Multiple-Choice Questions We input the content written by literary experts and the corresponding chapters identified based on motivation into GPT-4. Our goal is to generate multiple-choice questions that capture the complexity of the characters' decision-making processes. The correct option reflects the decision made by the characters in the original books, whereas the distractors are designed to be plausible for an arbitrary person. As shown in the example in Figure 1, *Michael Corleone* can ask for help from the government because he was once a Navy officer who trusted the government. However, in the preceding text, *Michael* witnesses the dark side of the government, so he ultimately chooses to stab the police.

Manual Examination We invite ten native English-speaking university students to filter the data and pay them according to local minimum wage standards. We supply the annotators with content written by literary experts and the multiple-choice questions, asking them to assess whether the model-created questions are challenging and reasonable. They are also tasked with filtering out data they deem low quality.

Ultimately, we collect 2,512 characters from 470 books and their corresponding life choices. Table 1 shows a complete data example. The specific prompts and more detailed data construction process can be found in Appendix B.

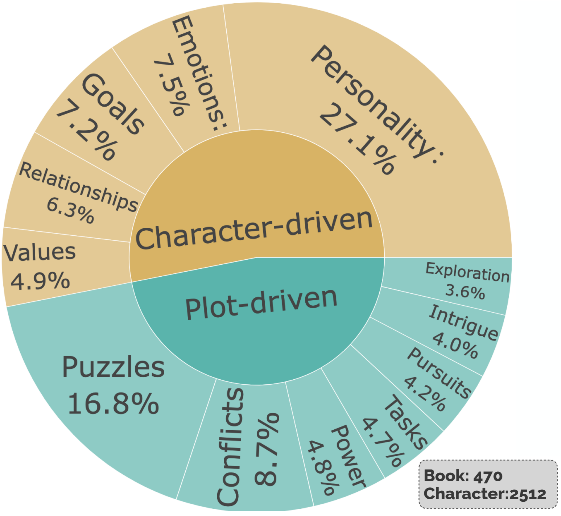


Figure 2: Statistics of motivation types in LifeChoice, with the first words for each motivation type.

3.2 Dataset Analysis

We refer to the drama theory of Aristophanes (Sommerstein, 2013; Silk, 2002) as the system prompt and use GPT-4o to classify the motivations for character decisions into two meta-motivations and several accompanying sub-motivations:

Character-driven Motivation Character-driven behavior revolves around the character's inner world, personality, and transformation. Sub-motivations of character-driven behavior include *Personality and Traits*, *Emotions and Psychological State*, *Social Relationships*, *Values and Beliefs*, and *Desires and Goals.*

Plot-driven Motivation Plot-driven behavior stems from a series of external events and conflicts unfolding. Characters often react passively within a larger narrative structure, with their actions led by external events. Sub-motivations of plot-driven behavior include *External Conflicts*, *Tasks and Goals*, *Puzzles and Secrets*, *Pursuits and Escapes*, *Exploration and Discovery*, *Power and Control*, and *Intrigue and Betrayal.*

Note that each topic is assigned one category of motivation. Figure 2 shows the proportion of different motivations. Detailed introductions for each sub-motivation are in Appendix C.2.

3.3 Task Setups

This task can be formulated as a prediction problem $P(y|x)$. The input x combines five components: the preceding context p, the current scenario s, the character name c, the question q, and the answer options a. Based on these inputs, the RPLA predicts the output y, which represents the character's actual decision in the narrative. For evaluation, we directly use the accuracy of multiple-choice question answering. As shown in Table 2, compared to other character understanding tasks, LifeChoice requires understanding the character through a more extended context to make decisions. RPLAs must locate relevant information related to the current scene in vast personal data. This behavior demands a more profound understanding of the characters.

4 Experiments

Because our inputs generally exceed 100k, it is difficult for LLMs to handle them directly. Therefore, our approach is divided into two steps: 1) **Character Profile Construction**, which includes the character's description and memories; 2) **Reasoning for Decisions**, where different LLMs use the constructed profile to answer the questions.

4.1 Character Profile Construction

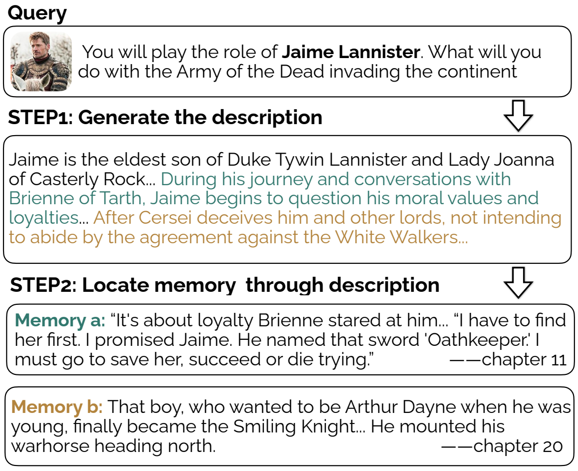
As shown in Figure 1, the character profile consists of two parts. The first part is the character's **description**, including their personality, experiences, hobbies, etc. The second part is the character's **memories**, specific segments from the preceding text. The methods for constructing these two parts are as follows:

Description Construction We adopt two automatic methods to construct character descriptions: (1) Recursive summary merging (Wu et al., 2021): Books are divided into chunks that fit within the LLM context window. The LLM summarizes each chunk, then merges and summarizes adjacent summarized chunks iteratively to produce the final description. (2) Progressive summary building (Chang et al., 2023): Books are divided into chunks and summarized sequentially, and the description is updated and refined incrementally by concatenating summarized chunks. The summarization model for both automated methods is GPT-3.5. Additionally, using the (3) expert-written descriptions from *Supersummary*, we employ GPT-4 to identify the positions of the decision points and truncate the text, providing only the data before these points. All descriptions are kept within 5k tokens, the maximum for expert-written descriptions.

Profile Construction	**Reasoning Model**	**Accuracy** *(%)*
	Description Construction	
Recursive Summary Merging	LLaMA-4	38.23
	Qwen-3	41.22
	Claude-3.5-sonnet	43.95
	GPT-4o	45.08
	Gemini-2.5-pro	46.11
Progressive Summary Building	LLaMA-4	39.07
	Qwen-3	40.88
	Claude-3.5-sonnet	44.56
	GPT-4o	48.18
	Gemini-2.5-pro	48.92
Expert-written Descriptions	LLaMA-4	53.04
	Qwen-3	52.77
	Claude-3.5-sonnet	54.08
	GPT-4o	55.18
	Gemini-2.5-pro	56.90
	Memory Retrieval	
BM25	GPT-4o	27.98
	Gemini-2.5-pro	28.51
Embedding	GPT-4o	35.19
	Gemini-2.5-pro	35.80
	Description & Memory	
Direct concatenation	LLaMA-4	58.01
	Qwen-3	57.84
	Claude-3.5-sonnet	59.21
	Deepseek-R1	58.36
	GPT-4o	60.07
	o1	61.12
	Gemini-2.5-pro	61.09
CharMap	LLaMA-4	61.43
	Qwen-3	62.87
	Claude-3.5-sonnet	62.18
	Deepseek-R1	62.90
	GPT-4o	64.22
	o1	**65.28**
	Gemini-2.5-pro	64.89

Table 3: Results of different LLMs and RPLA methods on LifeChoice.

Memory Retrieval We use two memory retrieval methods: (1) BM25 (Robertson et al., 2009): Scores documents based on term relevance and length, optimizing retrieval using term frequency and distribution. (2) Embedding-based retrieval: Uses dense vectors representing documents and queries to assess semantic similarity through vector distance. For the embedding model, we use OpenAI's text-embedding-ada-002(Neelakantan et al., 2022) model.



Query

You will play the role of **Jaime Lannister**. What will you do with the Army of the Dead invading the continent

STEP1: Generate the description

Jaime is the eldest son of Duke Tywin Lannister and Lady Joanna of Casterly Rock... During his journey and conversations with Brienne of Tarth, Jaime begins to question his moral values and loyalties... After Cersei deceives him and other lords, not intending to abide by the agreement against the White Walkers...

STEP2: Locate memory through description

Memory a: "It's about loyalty Brienne stared at him... "I have to find her first. I promised Jaime. He named that sword 'Oathkeeper.' I must go to save her, succeed or die trying." ——chapter 11

Memory b: That boy, who wanted to be Arthur Dayne when he was young, finally became the Smiling Knight... He mounted his warhorse heading north. ——chapter 20

Figure 3: An overview of CHARMAP, a two-step scenario-specific character profile building approach.

Description & Memory Using only Description or Memory alone may lead to information loss (Wang et al., 2024). Therefore, we also experiment by combining the results of both methods to form the character's profile. We adopt two methods: (1) Direct concatenation: This method concatenates the results from both approaches by prompting the model to role-play the corresponding character. By default, it uses the results from Human Description and Embedding retrieval. (2) CHARMAP: To better utilize the information in the Description, we propose CHARacter MAPping Profile Synthesis (CHARMAP), constructing a more scenario-specific profile in two steps. As shown in Figure 3, first, after obtaining the description, we input it along with the question into the model, asking it to locate the plot in the Description relevant to the current scene based on the question. Second, we use these episodes as queries to retrieve related memories and then input them into the inference model and the description. This leverages the overall character storyline in the description, thereby better retrieving related memories.

4.2 Reasoning for Decisions

After compressing the original input x into a character profile, we feed it into the LLMs. For methods using only description or memory, we use Llama-4 (Meta, 2025), Qwen-3 (Yang et al., 2025), Claude-3.5 (Anthropic, 2024) and Gemini-2.5-pro (Team, 2024). For methods using both, we also include DeepSeek-R1 (DeepSeek-AI, 2025),GPT-4 and o1 (OpenAI et al., 2024)[2].

	Raw text	Concat.	CHARMAP
GPT-4o	-	60.07	64.22
human	92.01	66.83	72.17

Table 4: Results of the human evaluation. Concat. refers to the direct concatenation of Description and Memory.

5 Analysis

In the experiments, we wish to answer two research questions: *RQ1)* Can RPLAs make decisions based on historical data? *RQ2)* What influences the decision-making of RPLAs?

5.1 *Can RPLAs Make Decisions Based on Historical Data?*

Analysis of Model Results Table 3 presents the accuracy results of different RPLA methods on the LIFECHOICE dataset, leading to the following observations: First, even the best model (Gemini-2.5-pro, o1), under optimal settings, only achieves an accuracy of approximately 61.12%, indicating that LIFECHOICE is a challenging task. Second, the method that uses both Description and Memory surpasses the one that uses only one, suggesting that both holistic and detailed data of key characters are essential in final decision-making. Thirdly, although Gemini-2.5-pro performs best among all models, the performance gap among different LLMs is insignificant while reasoning the answer. This indicates that the main factor for the result is the generated profile rather than reasoning ability. Last, CHARMAP outperforms the method that directly concatenates Description and Memory by 4.16%, proving its effectiveness. This scenario-specific profile better assists RPLA in decision-making.

Humans are Good Decision-makers We invite three native English-speaking university students to take a test in which we select six novels they have never heard of before. Each novel has between 3 to 5 characters and their corresponding multiple-choice questions. We provide each person with three data sets for each key character in two books: the full original text before the decision point, direct concatenation Description and Memory result, and the result from CHARMAP. As

[2] The versions in this paper are `Llama-4-Maverick`, `qwen3-235b-a22b`, `Claude-3.5-Sonnet`, `Deepseek-R1`, `gpt-4o`, `o1` and `gemini-2.5-pro` respectively

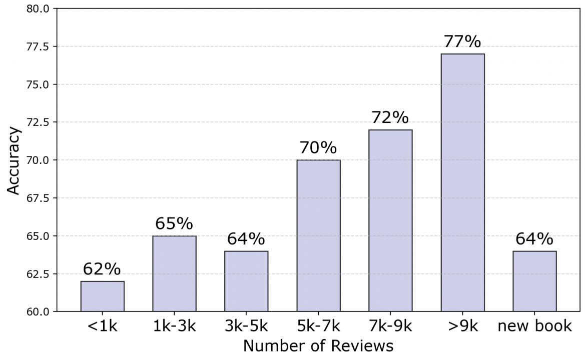


Figure 4: The impact of the number of book reviews on accuracy in LIFECHOICE, with *new books* being those not present in the training corpus of LLMs.

shown in Table 4, compared to direct concatenation, the CHARMAP results are easier for humans to understand. Additionally, humans slightly outperform GPT-4 in reasoning answers based on the profiles, indicating that humans can understand subtle character decisions better than models. When given the raw text, humans can achieve an accuracy rate of 92.01%, suggesting there is still significant room for improvement in RPLA methods.

Analysis and Mitigation of Data Leakage Data leakage presents a potential challenge as our experimental data may exist in the model's pre-training corpus. We implement preventive measures during data collection (section 3.1), including an entity replacement strategy that substitutes character names, locations, and other entities with placeholders during evaluation. To quantify the impact of data leakage, we examine the relationship between a book's popularity and model performance. We use review counts from a book review platform[3] as a proxy for popularity, which correlates with the likelihood of inclusion in LLM training data. Our analysis uses CHARMAP for profile construction and GPT-4 for role-playing, testing on thirty books across different popularity levels and thirty control books published after the model's training cutoff date (August 6th, 2024 for *gpt-4o*). Figure 4 reveals a clear pattern: model accuracy increases significantly for books with over 5,000 reviews, while books below this threshold show performance comparable to the control group. This finding suggests that data leakage minimally affects CHARMAP's performance on less popular books. Based on these results, we establish 5,000 reviews as the popularity threshold for book selection in section 3.1.

[3] https://www.douban.com/

LLMs	Method	Accuracy
Claude-3.5-sonnet	long-context	59.75
Claude-3.5-sonnet	CHARMAP	62.18
Gemini-2.5-pro	long-context	60.32
Gemini-2.5-pro	CHARMAP	64.89

Table 5: The results of using long-context models for LIFECHOICE.

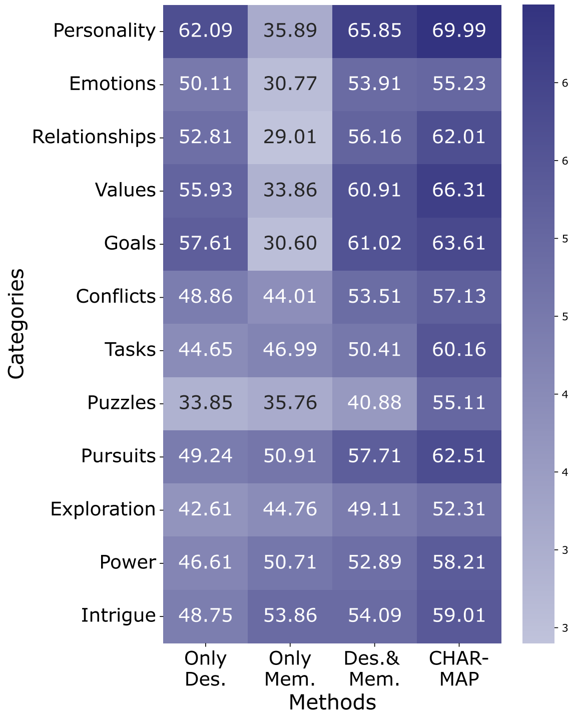


Figure 5: Heatmap of the impact of motivation types on the results. The results are predicted from the Progressive Summary Building, the embedding-retrieved memory, the direct concatenation of both, and CHARMAP. The reasoning model uses GPT-4.

Analysis of Long-Context LLMs LIFECHOICE requires processing long context, making it suitable for evaluating long-context LLMs. This evaluation focuses on models' ability to understand global information and make character-based decisions. We test two recent long-context models: Claude-3.5 and Gemini-2.5-pro. Results in Table 5 show that while these models perform below CHARMAP, they demonstrate competence in role-playing tasks. The multi-step reasoning and comprehensive context understanding required by LIFECHOICE position it as an effective benchmark for assessing long-context model capabilities.

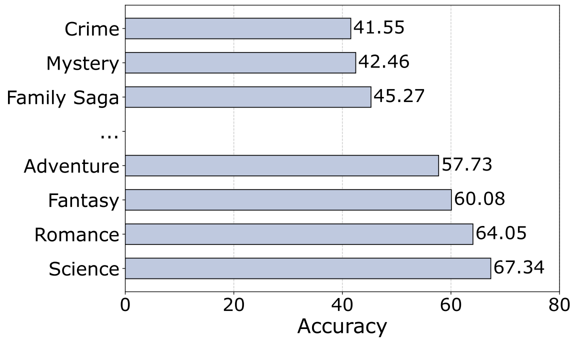


Figure 6: The result of the impact of different novel genres on accuracy.

5.2 *What Influences the Decision-making of RPLAs?*

The Impact of Motivation Types In line with the motivation types presented in Section 3.2, we examine how different types of motivation influence characters' decision-making. For profiles, we evaluate four methods: the Incremental updating, the embedding-retrieved memory, the direct concatenation of both, and CHARMAP. For reasoning, we use GPT-4 uniformly. The results are shown in Figure 5. We find that tasks requiring coherent reasoning, such as puzzles and mysteries, are not well answered for all methods. This might be because these questions need multi-step reasoning and details from various memories. Moreover, plot-driven questions have lower accuracy when descriptions are used only for the profile. Conversely, character-driven questions are challenging to answer when relying only on memories. We believe this is because character summaries in descriptions better capture the overall essence of the characters, while memories provide direct access to relevant events.

The Impact of Novel Genres We analyze character selection accuracy across genres using novel tags from the website. Experiments compare direct concatenation of descriptions and memories with GPT-4 as the reasoning model. Figure 6 shows high accuracy for science fiction, fantasy, and romance novels, likely due to their stylized characters and established archetypes. Conversely, crime and mystery novels show lower accuracy, possibly due to their complex logical narratives and characters' unpredictable behaviors. Further details can be found in Appendix C.1.

The Impact of Temporal Data A character's personality and circumstances may change over time. We study the decisions of the same character with memories from different time periods. Specifically, we randomly sample 40 characters, half character-driven, and half plot-driven. We split the content preceding the decision points into five equal sections and used these various content lengths as input. We conduct experiments on the combination of human description and embedding-retrieved memories, and the reasoning model is GPT-4. As shown in Figure 7, in the early stages, the accuracy of most characters' decisions is close to random (25%), potentially due to insufficient information. As more information becomes available, the characters' decisions tend to be closer to the correct choice. For character-driven decisions, accuracy tends to be stable. For plot-driven, the accuracy rate may change abruptly. This could be due to the relatively stable characteristics of a character, while some sudden events may greatly influence the final choices of the character.

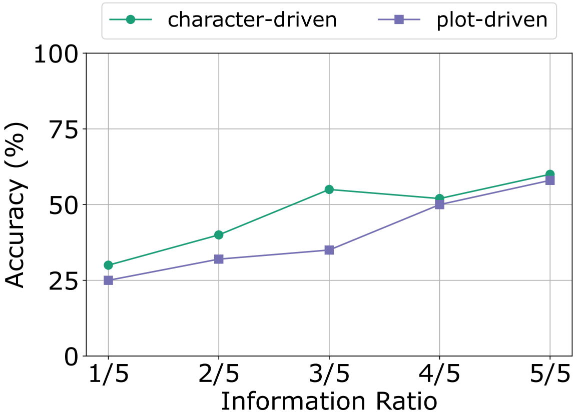


Figure 7: Analysis of whether character selection will change. The x-axis represents the input length relative to the point truncation.

6 Conclusion

In this work, we introduce a novel evaluation framework for RPLAs that focuses on decision-making capabilities. This framework tests whether LLMs can accurately reconstruct character decisions using historical data. To support this evaluation, we develop LIFECHOICE, a comprehensive dataset containing 2,512 characters from 470 books and their documented life choices. Our experiments on LIFECHOICE show that for RPLAs engaged in role-play based on extensive texts, reconstructing decisions from the original books is a challenging task, requiring accurate memory recall and a certain degree of reasoning.

Limitations

In this paper, we primarily investigate whether fictional characters can recreate their choices within a book. Although we have controlled the quality of the novels, there may still be issues with the plot and characters since the author designed the storyline, which can result in illogical choices within the book. Furthermore, as our research focuses mainly on fictional characters, there is a certain gap compared to real-world humans. For example, the author fictionalizes some story backgrounds, which may impact the model's generation results. Additionally, to aid in the construction and design of our dataset, we utilized large language models such as GPT-4. While all data subsequently underwent thorough manual verification, the inherent complexities of AI-assisted data generation mean that subtle inaccuracies or biases might persist despite our diligent review processes. Beyond these limitations, it is crucial to consider the potential social impacts of role-playing models. Despite all our data undergoing manual review, there remains a possibility that our methods could be referenced to construct models and data with the potential for social harm, such as role models of figures like Hitler or characters designed for psychological manipulation. Therefore, more robust safety and regulatory measures are essential for large language model-based role-playing.

Ethics Statement

Use of Human Annotations Our institution recruits annotators to implement the annotations of motivation recognition dataset construction. We ensure the privacy rights of the annotators are respected during the annotation process. The annotators receive compensation exceeding the local minimum wage and have consented to use motivation recognition data they process for research purposes. Appendix D provides further details on the annotations.

Risks Although the LIFECHOICE dataset is carefully compiled from literary novels and expert analyses, the novels themselves as data sources—originating from different authors and historical contexts—may contain language, themes, or character portrayals that appear sensitive in contemporary contexts, or reflect inherent latent biases in the original texts.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

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

Yu-Hsin Chen and Jinho D Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 90–100.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Martin Dodge and Rob Kitchin. 2007. 'outlines of a world coming into existence': pervasive computing and the ethics of forgetting. *Environment and planning B: planning and design*, 34(3):431–445.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. *arXiv preprint arXiv:2306.08401*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Cathal Gurrin, Alan F Smeaton, Aiden R Doherty, et al. 2014. Lifelogging: Personal big data. *Foundations and Trends® in information retrieval*, 8(1):1–125.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Matthew B Hoy. 2018. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88.

Akriti Jaiswal, A Krishnama Raju, and Suman Deb. 2020. Facial emotion detection using deep learning. In *2020 international conference for emerging technology (INCET)*, pages 1–5. IEEE.

Andreas Kaplan and Michael Haenlein. 2019. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business horizons*, 62(1):15–25.

Yuanchun Li, Ziyue Yang, Yao Guo, Xiangqun Chen, Yuvraj Agarwal, and Jason I Hong. 2018. Automated extraction of personal knowledge from smartphone push notifications. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 733–742. IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, and Ahmed El-Kishky. 2024. Openai o1 system card.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. Tvshowguess: Character comprehension in stories as speaker guessing. *arXiv preprint arXiv:2204.07721*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Michael Stephen Silk. 2002. *Aristophanes and the Definition of Comedy*. Oxford University Press, USA.

Alan Sommerstein. 2013. Aristophanes. *The Encyclopedia of Ancient History*.

Sanja Štajner and Seren Yenikent. 2020. A survey of automatic personality detection from texts. In *Proceedings of the 28th international conference on computational linguistics*, pages 6284–6295.

Xiaoli Tang, Tengyun Wang, Haizhi Yang, and Hengjie Song. 2019. Akupm: Attention-enhanced knowledge-aware user preference model for recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1891–1899.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, and Amjad Almahairi. 2023. Llama 2: Open foundation and fine-tuned chat models.

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. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*.

Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022. Cosplay: Concept set guided personalized dialogue generation across both party personas. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22. ACM.

Rui Xu, Dakuan Lu, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, Wei Chu, and Yinghui Xu. 2024. Mindecho: Role-playing language agents for key opinion leaders. *arXiv preprint arXiv:2407.05305*.

Rui Xu, MingYu Wang, XinTao Wang, Dakuan Lu, Xiaoyu Tan, Wei Chu, and Yinghui Xu. 2025. Guess what i am thinking: A benchmark for inner thought reasoning of role-playing language agents. *arXiv preprint arXiv:2503.08193*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, and Binyuan Hui. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. *arXiv preprint arXiv:2404.12726*.

Samira Zad, Maryam Heidari, H James Jr, and Ozlem Uzuner. 2021. Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0255–0261. IEEE.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

A Analysis

A.1 Generative Evaluation

LIFECHOICE utilizes a multiple-choice question format primarily for two reasons: 1) to enable better quantitative evaluation, as we recognize that the space of personal "life choices" is inherently vast, and 2) the multiple-choice format effectively constrains this extensive decision space. However, in practical applications, choices are often generated rather than selected from a predefined set, which can introduce a specific gap. To address this, we conducted additional experiments. We allowed the model to freely generate the following possible behavior based on the scenario and then compared it with the actual behavior for evaluation. We employed two types of evaluation methods: 1) automatic text evaluation metrics, using BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) for assessment, and 2) model-based automatic evaluation methods. For the latter, we used the model mDeBERTa-v3-base-xnli[4], an NLI model trained on DeBERTa (He et al., 2021). This model classifies text pairs into entailment, contradiction, or neutral relationships, and we used the entailment probability as the evaluation score. Table 6 presents the results of these experiments. As shown, the evaluation outcomes for free generation are largely consistent with the trends observed with our multiple-choice format. Nevertheless, these metrics are noticeably lower, which is attributable to the considerably larger generation space.

B Prompts

In this section, we provide the key prompts we used, including prompts for selecting decision points, locating the node's position, constructing multiple-choice questions, system prompts for role-playing characters, and the prompt of CHARMAP.

B.1 Selecting Decision Points

As mentioned in section 3.1, the first step in constructing our data is selecting the character's Decision Points. In this step, our input data consists of all raw data from *Supersummary*, including the current character's description, all chapter summaries, and book analysis. We provide this data to GPT-4, requesting it to output three types of data: the character's decision point, the corresponding gold motivation, and the chapters related to this choice. Table 7 presents the character's description, Table 8 presents the chapter summaries, and Table 9 presents the book analysis, these examples are all from the 2024 novel *"A Calamity of Souls."*. Table 10 illustrates the prompt for selecting decision points.

B.2 Locating the Node

Furthermore, for the character's decision points, we provide the previously identified decision points and their corresponding chapters from the original book to GPT-4. This allows it to precisely determine the position in the original book that should be segmented, helping to avoid data leakage. Table 11 shows the prompt for locating.

B.3 Constructing Multiple-Choice Questions

After selecting the Decision Points, our next step is constructing Multiple-Choice Questions. In this step, our input data consists of all the input and output data from the previous step(Appendix B.1). We ask GPT-4 to construct a multiple-choice question regarding the character's decision based on this data, outputting the scenario in which the character is situated, the question, and four options. Table 12 shows the prompt for constructing multiple-choice questions.

B.4 System Prompts for Role-Playing

The prompt for role-playing as the character can be found in Table 13.

B.5 Prompts for CHARMAP

The prompt for CHARMAP can be found in Table 14.

C Dateset Details

C.1 Categories of novel

Below is a complete classification of novel genres, from the literary experts at the Supersummary website:

Mystery Novels: The mystery genre includes general mystery, noir mystery, historical mystery, police procedural mystery, and supernatural mystery.

Thriller Novels: The thriller genre includes supernatural thrillers, historical thrillers, environmental thrillers, medical thrillers, legal thrillers, political thrillers, military thrillers, and espionage stories.

[4] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

Profile Construction	**Reasoning Model**	**Accuracy** *(%)*	**BLEU**	**ROUGE**	**NLI Score** *(%)*
Description Construction					
Recursive Summary Merging	LLaMA-4	38.23	2.51	15.32	31.56
	Qwen-3	41.22	3.15	17.08	33.19
	Claude-3.5-sonnet	43.95	3.92	18.55	34.80
	GPT-4o	45.08	4.20	19.17	35.72
	Gemini-2.5-pro	46.11	4.83	20.05	36.15
Progressive Summary Building	LLaMA-4	39.07	2.70	15.93	32.11
	Qwen-3	40.88	3.01	16.82	32.88
	Claude-3.5-sonnet	44.56	3.87	18.90	35.13
	GPT-4o	48.18	4.55	19.83	37.04
	Gemini-2.5-pro	48.92	5.02	20.56	37.85
Expert-written Descriptions	LLaMA-4	53.04	15.28	35.70	40.27
	Qwen-3	52.77	15.03	35.11	39.85
	Claude-3.5-sonnet	54.08	15.88	36.23	41.02
	GPT-4o	55.18	16.15	36.98	41.95
	Gemini-2.5-pro	56.90	16.97	37.85	42.88
Memory Retrieval (NLI Model: ST5-L for these entries)					
BM25	GPT-4o	27.98	0.85	7.82	27.98
	Gemini-2.5-pro	28.51	0.92	7.99	28.51
Embedding	GPT-4o	35.19	1.02	8.11	35.19
	Gemini-2.5-pro	35.80	1.15	8.35	35.80
Description & Memory					
Direct concatenation	LLaMA-4	58.01	16.95	38.02	43.15
	Qwen-3	57.84	16.82	37.88	42.97
	Claude-3.5-sonnet	59.21	17.15	38.93	44.02
	Deepseek-R1	58.36	17.03	38.15	43.55
	GPT-4o	60.07	17.98	39.81	45.10
	o1	61.12	18.53	40.15	45.92
	Gemini-2.5-pro	61.09	18.22	40.03	45.85
CharMap	LLaMA-4	61.43	18.93	40.88	46.13
	Qwen-3	62.87	19.55	41.93	47.05
	Claude-3.5-sonnet	62.18	19.02	41.10	46.89
	Deepseek-R1	62.90	19.89	42.03	47.11
	GPT-4o	64.22	20.15	43.05	48.23
	o1	**65.28**	**21.08**	**44.19**	**49.15**
	Gemini-2.5-pro	64.89	20.86	43.87	48.90

Table 6: Results of generative evaluation for different LLMs and methods on LifeChoice.

Science Fiction Novels: Science fiction stories take place in the future or the past but are almost always set in a dimension different from our present. They are characterized by entirely new, imagined realities and universes, where the setting is indispensable. High technology also plays an important role in these stories. Space opera, romantic science fiction, military science fiction, alternate history, dystopian and utopian tales, as well as steampunk, are considered sub-genres of science fiction.

Romance Novels: Romance novels feature romantic relationships between at least two people, characterized by tension and desire. Romance novel themes include supernatural romance, contemporary romance, historical romance, western romance, gothic romance, regency romance, and romantic suspense.

Fantasy Novels: Fantasy stories are centered around mythical kingdoms and magic. Fantasy novel genres include contemporary fantasy, traditional fantasy, horror fantasy, weird fantasy, epic fantasy, historical fantasy, dark fantasy, urban fantasy, and anime fantasy.

Action Adventure Novels: Action-adventure novels place the protagonist in various realistic dangers. This is a fast-paced genre where the climax should provide some form of thrill for the audience or reader.

Speculative Novels: Speculative fiction is characterized by overlapping with our world but differing in key aspects, introducing "what if" scenarios.

Mystery Thriller Novels: Mystery thriller stories are usually filled with suspense, with one or more characters' lives in danger. In gripping scenes, these characters are often chased and manage to escape narrowly.

Young Adult Novels: Young Adult fiction, commonly abbreviated as YA, is intended for teenagers aged 12-18. Most YA novels feature coming-of-age stories, often with elements of science fiction or fantasy.

New Adult Novels: New Adult novels target college-aged adults and usually explore stories of first adventures on one's own.

Horror and Supernatural Novels: Horror, supernatural, and ghost story genres aim to scare the reader and audience by playing on common fears. The protagonist usually has to overcome supernatural threats, and the stories often include supernatural elements.

Crime Mystery Novels: Crime mystery stories focus on a central problem or crime to be solved, or a mysterious event that must be answered. Throughout the story, the reader or audience and characters are given clues that help the protagonist eventually find the solution.

Detective Novels: In detective fiction, a common element is a police officer or detective embarking on solving a crime. The plot is filled with evidence gathering, forensic studies, and legal drama.

Historical Novels: Historical novels are fictional stories set against the backdrop of real historical events or historical settings. Historical fiction may also portray real historical figures.

Western Novels: Stories with a western theme take place in the old times of the American West, filled with adventure, cowboys, and pioneers. There are also Italian western novels, Asian western novels, space westerns, and other stories about the American West.

Family Saga Novels: Family saga novels typically tell the stories of several generations of family members dealing with family affairs, family curses, and family adventures. These stories usually follow a timeline and deal with conflicts in the present.

Women's Novels: Women's fiction plotlines revolve around the challenges and crises that women face in real life, including interpersonal relationships, work, family, politics, and religion.

Magical Realism Novels: Magical realism stories take place in the real world but have characters who take magical elements for granted. These mystical elements do not exist in real life, but they are perfectly normal in magical realism.

C.2 Categories of motivations

Below are the motivations for each topic and their corresponding proportions:

Character-driven motivation Character-driven narrative is centered on the inner world, growth, and transformation of characters. In character-driven stories, the progression of the plot and the resolution of conflicts are often propelled by the characters' personalities, desires, fears, and psychological development. Such stories typically delve deeply into the characters' mental states and development, focusing on how characters influence each other and how their actions reflect their inner emotions and thoughts. The choices and changes of the characters serve as the main engine for the story's development, influencing the direction of the plot. Sub-motivations of character-driven behavior include:

Personality and Traits: (27.12%) These refer to a character's characteristics such as being introverted, extroverted, brave, or guilt-ridden, which influence their choices and lifestyle.

Emotions and Psychological State: (7.53%) A character's emotional responses, psychological traumas, or sense of personal well-being are key elements that drive the story forward.

Social Relationships: (6.31%) The character's status and changes in family, love, friendship, or other social connections can propel the story's development.

Values and Beliefs: (27.12%) The character's moral convictions, religious beliefs, or life philosophy can serve as motivation for action.

Desires and Goals: (7.22%) Personal desires, career aspirations, or specific life goals of a character are pivotal in advancing the plot.

Plot-driven motivation Plot-driven narrative emphasizes the creation and resolution of external conflicts in the story. In such stories, the driving force of the plot comes from a series of events and conflicts themselves, while characters are often the responders to these events. Plot-driven stories typically highlight tense drama, complex plot structure, and frequent changes in external actions, rather than changes in the character's internal world. In this type of narrative, characters may act in response to the demands of the plot, rather than the plot following the development of the characters' inner world. Sub-motivations of plot-driven behavior include:

External Conflicts: (8.76%) Conflicts from the outside world, such as war, natural disasters, or social upheaval, can propel the plot.

Tasks and Goals: (4.7%) Tasks or specific goals that characters must accomplish often become the driving force behind the story's progression.

Puzzles and Secrets: (7.22%) Secrets that need revealing or mysteries that need solving can form the core of a story.

Pursuits and Escapes: (4.25%) Characters might chase something (e.g., power, wealth, knowledge) while avoiding or fleeing from certain situations (e.g., pursuit, personal past).

Exploration and Discovery: (3.66%) Characters' adventures or discoveries in new realms (physical, scientific, or spiritual) can move the plot forward.

Power and Control: (4.81%) The pursuit or struggle for power and control often serves as motivation for characters.

Intrigue and Betrayal: (4.09%) Complex plots and betrayals can catalyze the progression of the story.

D Manual Annotation

This is a supplement to Section 3.1. After constructing the multiple-choice question data using GPT-4, we perform a manual examination. For each annotator, we provide key character descriptions, chapter summaries, and book analyses written by human literature experts on the *Supersummary*. Each annotator is asked to score the questions constructed by GPT-4 based on the rules shown in Table 15. We evaluated the scores of each annotator and only retained the data with an average score of more than 6 points.

We provide compensation based on the local minimum hourly wage for all individuals involved in the annotation.

E Future direction

Building personal agents for everyone is an exciting topic. We have explored how fictional characters can determine their subsequent actions based on historical data, proposing possibilities for combining role-playing with personalized models. We believe there are additional directions to explore:

- **Real-life version of LIFECHOICE** The behavior of fictional characters often stems from the author's design, which can lead to logical inconsistencies. In contrast, real-world human behavior data do not have this issue. How to construct real human historical data and related behavior data is a question worth exploring.

- **Improving RPLA performance in life choices** Although CHARMAP has achieved decent results, better methods are needed to balance reasoning efficiency (which depends on the length of input tokens) while achieving superior outcomes.

- **More complex downstream decision tasks** The decisions we select are often significant choices for fictional characters, resulting in a large decision space without a fixed task framework. Identifying more systematic tasks by integrating social sciences is a challenge that needs to be addressed in the future.

Key Character Description
John “Jack” Robert Lee
Jack Lee is the protagonist of the novel. He is a white man who sees himself as intelligent, having done well in school and having a law degree, yet he acknowledges that he did not go to one of the best law schools. At the beginning of the novel, he turns 33. He is disappointed with much of his professional life, having been out of law school for eight years and still “just getting by.” He believes that he has largely failed to “change” the world—which was one of his goals in becoming a lawyer. Jack changes throughout the course of the novel. At the novel’s start, he recognizes the racist actions of his mother and is glad that segregation is ending. Because of his love of books, he has a vast knowledge of Black history and the hardships that Black people have faced, yet he largely ignores these hardships, as they do not affect him. He largely chooses to go along with the way things are and scolds himself for not being a “risk-taker.” However, he realizes how very real injustice is for people like Jerome. He sets aside his own fear and faces the danger of representing him. Initially, he dislikes DuBose’s interest in speaking with the press and trying to use Jerome’s trial as anything other than a chance to save Jerome’s life. By contrast, he speaks to the press for the first time at the novel’s conclusion, making an impassioned plea about the importance of coming together as a community.
Desiree DuBose
DuBose is a Black lawyer from Chicago. She is extremely intelligent, having gone to college at age 16 and graduated from Yale Law School after six years. She has a vast array of experience working with the NAACP and the Legal Defense Fund to fight against racist legislation, including, mostly recently, the Loving case, which allowed Black and white people to marry. She initially comes to Virginia with the intent of taking over Jerome’s defense from Jack, but after seeing how committed he is to the case, agrees to be his co-counselor. DuBose contrasts with Jack. She is very different from him on the surface level: She is a woman, Black, has served as a lawyer in dozens of murder trials, and recognizes the importance of Jerome’s trial to the larger picture of civil rights. However, like Jack, she is a dynamic character in that she changes throughout the text. After Jack asks her to stop focusing on mistrials or appeals, she brings her full effort to the courtroom, fighting back against the admission of the murder weapon instead of utilizing it as a chance for appeal. At the novel’s end, she succeeds in Overcoming Personal Bias by putting aside her fear and hesitancy to be involved with Jack.
Jerome Washington
Jerome is a Black man on trial for the murder of a wealthy white couple. He is a veteran of the Vietnam War and is described as “large” and “strong,” standing at 6’5”. He is dedicated to his job working for the Randolphs before their death, riding his bike five miles each way, never missing a day of work. He is also dedicated to his wife, Pearl and three children; he is willing to go to prison for life if it means that Pearl is able to be acquitted, then is willing to accept a plea of five years, despite the overwhelming evidence in his favor. Jerome is a flat character in the novel, meaning that he doesn’t change. He serves primarily as a plot device and a way to illuminate the systemic racism and injustice of the time. The narrative offers little information about him. Jack and DuBose repeatedly choose not to put him on the stand to speak even in his own defense. This reflects his status in 1968 in the American South: He has little control of his own life, and is at the mercy of the white people around him and the racist system that he lives in.
Hilda “Hilly” Lee
Hilly is Jack’s mother. A homemaker, she cares for her developmentally disabled daughter, Lucy, well into adulthood while blaming herself for Lucy’s disability. Jack sees his mother as “complicated.” Hilly is upset by Martin Luther King Jr.’s death and helps the Black men who work with her husband, yet is vocal about her belief in segregation. DuBose initially thinks of her as a “typical racist” who perpetuates the idea that she is somehow superior to the Black people around her. Hilly is a dynamic character who changes throughout the text. As the text progresses, she reverts back to who she originally was prior to the events of the novel—a kind, nonracist woman. After Lucy’s death, she gets to know DuBose and invites her into her home, even lending her clothing. She then reveals that she used to love a Black man, but was forced apart from him. Additionally, she was told by a preacher that Lucy’s disability was a punishment from God in return for loving a Black man. These experiences, and the society in which she lived, led to years of acceptance and eventual perpetuation of racism.
Howard Pickett
Pickett is an antagonist, or villain, of the text. He is a wealthy man who owns coal mines. He is interested in the trial because he wants to use it as a talking point for campaigning for George Wallace to win the 1968 presidential election. Throughout the novel, he speaks with the presses and stresses the importance of Jerome being convicted to emphasize that segregation should be legalized again. Whether Pickett believes these ideas or is simply using them to drum up support for his political campaign is unclear; however, he is unapologetic in his racist language. Pickett represents the true problems that need to be addressed in America. He distracts the working class by perpetuating racism and stressing that Black people are the problem. In this way, he masks that the true issue lies with the greed and theft of the wealthy from the working class.

Table 7: Data examples of key character descriptions written by literary experts, sourced from the 2024 novel *"A Calamity of Souls."*

Chapter Summaries
Chapter 1
In Freeman County, Virginia, in 1968, an elderly white couple is dead in their home. The husband is sprawled across the floor, while his wife's body lays across a chair. Two white officers—Raymond Leroy and Gene Taliaferro —have a Black man in handcuffs on the floor, referred to as "the only suspect in the room". Raymond struggles to read him his Miranda Rights off an index card, a new policy recently enacted in the police force. The idea of reading them annoys both Raymond and Gene, who are bothered by the idea of criminals getting representation, especially "those people, who had committed crimes, usually against white folks". Gene interrupts Raymond to hit the suspect with his club. He forces the suspect to lie down, then hits him again, then forces him to kneel again. Gene goads the suspect into getting angry by asking him about his wife and family. When the suspect reacts with rage, struggling against his handcuffs, Gene is excited that he can now claim the suspect was "resistin' arrest" and raises his club to beat him.
Chapter 2
John "Jack" Robert Lee is a lawyer from Freeman County, Virginia. He is white, single, and grew up in a working-class home, with a love of books and debate. He arrives at his parents' home to celebrate his 33rd birthday. He is greeted by his older sister, Lucy. She is 37. Due to their mother's exposure to nitrous oxide at the dentist while pregnant, she is developmentally disabled.
Chapter 3
Jack's mother, Hilda "Hilly" Lee," has always cared for her home and children while her husband works. She harbors some guilt over what happened to Lucy and chose to have Jack and his younger brother, Jefferson, without any pain killers—even aspirin. She refers to Jack as "Robert," insisting that she would have named him Robert E. Lee if her husband did not have a say. Despite naming her son after a Confederate general, she still tells Jack how upset she is about the deaths of Martin Luther King Jr. and Robert Kennedy. Jack finds it "bewildering" that she respects Lee while mourning King and Kennedy, who "held views diametrically opposed to all the Confederacy had stood for." Hilly tells Jack that Miss Jessup was by earlier looking for him. Miss Jessup is one of the only Black women in the area. She is a housemaid to Ashby, a wealthy, retired lawyer who lives down the street.
...
Chapter 93
Jack flies to Chicago. On the flight there he thinks of his injury, and how the bullet barely missed doing any major damage. He was also saved by the fact that the bullet went through Jerome first; he is saddened that he can't thank him for saving his life. He goes to DuBose's apartment and surprises her. He tells her that he came to Chicago to work with her. She makes it clear that they can only work together, and Jack admits that he cares for her. She tells him that she once lost a man she loved because of the work that they do, and she can't go through that pain again. She compares it to Jack's losing Lucy. Jack insists that meeting her was one of the best things that ever happened to him, despite the damage it caused. When DuBose doesn't respond, Jack turns to leave. She stops him, informing him that it will "be far tougher" than he thinks. Jack insists that he is now "far tougher" than he thought he would ever be.

Table 8: Data examples of chapter summaries written by literary experts, sourced from the 2024 novel *"A Calamity of Souls."*

Book Analysis
Chapters 1-10
The setting of the novel plays a pivotal role. Set in the South in 1968, the country is on the verge of moving forward with the end of Jim Crow and segregation, while people throughout the South fight against this. The novel discusses the death of Martin Luther King Jr. as well as the impending presidential election as important moments in history that will decide the future. Jack battles with Overcoming Personal Bias. He has grown up in the South with a mother who encourages segregation, and up until the novel's events, has not fought against racism or done something meaningful with his law degree. He struggles with his lack of action, and contemplates whether the danger that will surround the trial is worth it. He recognizes his own bias, and doesn't yet see racism as an important enough cause to risk his safety. However, when he is confronted with violence on taking Jerome's case, it has the opposite of its intended effect: Instead of discouraging him from defending Jerome, the violence shows him that he has ignored the problem of racism for too long. Jack's mother, Hilly, represents a vast number of white people throughout the South. She is complex: She helps Black people in need but also adamantly believes in segregation. This reflects the beliefs that many people hold throughout the novel. Although she does not believe herself to be racist and feels sympathy for Black people, she still exhibits racist biases and does not want to go against the status quo. Like Jack, she battles with her own personal bias and reflects on whether change is truly needed. This section begins to examine Racial Injustice and the Legal System. The text reveals the limits that the legal system has when placed in the hands of racist people. Despite what the law says, people continue to perpetuate racism, both directly and indirectly. This reveals that the law struggles without the support of its people. David Baldacci raises the stakes surrounding the trial to build suspense. For example, after registering himself as Jerome's lawyer, Jack receives a frightening phone call. The call reinforces his belief that he is doing the right thing using his legal skills, and that he is fighting back against injustice. It foreshadows the danger that will surround the case for both Jack and the people in his life.
...
Chapters 76-93
This section continues to examine The Importance of Family and Community Support. Like with Lucy's death, the end of the trial and Jerome's murder act as catalysts, where people of different races come together. For instance, Jerome's funeral is attended by many white people who did not even know him. Additionally, Jack notes that people agree with his speech after the trial. In this way, the novel implies hope for the future, and suggests that racial injustice can be overcome. This section continues to examine Racial Injustice and the Legal System. Jack and DuBose prove several things throughout the trial: Several of the witnesses for the prosecution were pressured into giving false statements, Pearl could not have helped with the crime, and Jerome could not have committed the murder due to his injury. Despite all of this, they are still not able to get the case thrown out and are forced to consider the best plea deal Battle can offer—which still involves Jerome going to prison. This blatant injustice reflects just how unfair the legal system was for Black people in the 1960s—and how little it did to defend their rights, even when the law is on their side. When Jerome is killed, the injustice of the legal system is further illuminated. As Jerome lies dying, several policemen do not stop the shooter with force, and instead try to talk to him. Even after the shooter raises his gun to shoot Pearl, Jeff, a civilian, shoots him. As he does so, the policeman angrily asks: "Why in the hell did you shoot him?". The police's rage and inaction reflects how little the legal system does to enforce even the laws in place. This scene establishes the importance of Overcoming Personal Bias. Jerome's death makes it clear that even new laws and courtroom triumphs are not enough for true change when people act on their own prejudice. Individual people need to change if racism is going to be overcome. The novel again presents youth as a solution to combatting personal bias. As Jack gives his speech at the conclusion of the trial, he notes how a woman in the crowd "was looking angrily" at him, but that "her boy's expression was more muted; he actually appeared to be listening". In this way, the novel shows that the youth are amenable to change. Although the grown woman is angry at the outcome of the trial, there is hope for the future—her son seems to be internalizing what Jack is saying, forming his own opinions instead of perpetuating his mother's bias. Jack and DuBose overcome their own personal bias and agree to start a romantic relationship. Their hesitancy has been two-fold. First, they both reflect throughout the novel on how complicated it would be to be with someone of a different race, with DuBose scolding herself for considering it. Second, their hesitancy comes from the fact that they come from such different backgrounds. However, Jack acknowledges that, just like with the trial, there are things that are worth the fight, and he believes that their relationship is one. He tells DuBose that he is "far tougher than [he] thought [he] would be", reflecting his newfound internal strength and transformation.

Table 9: Data examples of book analysis written by literary experts, sourced from the 2024 novel *"A Calamity of Souls."*

Prompt I

Your task is to help me identify a significant decision point for a character in a book. I will provide you with some information, and you need to return outputs as required.

# Requirements:
1. The decision must be a life choice of the character, which can reflect the character's personality, past experiences, and interpersonal relationships.
2. The decision must have a rationale that can be found earlier in the text, which might be determined by the character's overall personality or by a subtle hint.
3. The decision is determined by earlier text, not revealed by reasons in later sections.

Below are the inputs I will provide and the outputs you need to return.
# Inputs:
1. Input 1: A character description written by a human literature expert
<description>
2. Input 2: The book divided into chapter summaries
<chapter>
3. Input 3: A book analysis written by a human literature expert
<analysis>

Below is the content you need to output.
# Outputs:
1. Output 1: The location of the character's decision point. Please answer with the original text from the chapter summaries (Input 2).
Output format: {"summary_location":<content>}
2. Output 2: The motivation for the character's decision.
Output format: {"motivation":<content>}
3. Output 3: Chapter numbers related to this decision.
Output format: {"related_chapter":["chapter_1",...]}

# Execution Steps:
1. Read all inputs.
2. Consider the life choice that best reflects the character's personality, past experiences, and interpersonal relationships.
3. Output the location of the character's decision point, sourced from the chapter summaries (Input 2).
4. Output the motivation for the character's decision.
5. Based on the motivation for the character's decision, find the relevant chapters and output the chapter numbers related to this decision.

Table 10: Prompt templates for selecting decision points.

Prompt II

Your task is to locate the position of a given segment of text within the original text. The text segment I provide to you comes from a summary of the original text. I will provide you with the summary and the original text, and you need to find where this segment occurs in the original text. This position must be the exact wording from the original text.

# Inputs:
1. Input 1: Summary of the original text
<summary>
2. Input 2: Text from the summary of the original text
<text>
3. Input 3: Original text
<original>

# Outputs:
Text from the summary of the original text:

.

Table 11: Prompt templates for locate the position of the node in the original book.

Prompt III

Your task is to create a multiple-choice question where the correct answer is a decision made by a character in a book. You need to design three incorrect answers. Below are the detailed requirements:

# Requirements:
1. Provide the scenario in which the character is making the decision.
2. Design three incorrect answers that are reasonable and could be choices the character might make, but are not the optimal choice.
3. Ensure that there is no data leakage in any of the outputs.

# Inputs:
1. Input 1: Character description written by a human literature expert
<description>
2. Input 2: Summary of the entire book divided by chapters
<chapter>
3. Input 3: Book analysis written by a human literature expert
<analysis>
4. Input 4: Location of the character's decision point
<location>
5. Input 5: Motivation for the character's decision
<motivation>
6. Input 6: Original text from chapters related to the decision
<original>

# Outputs:
1. Output 1: Scenario in which the character is situated.
Output format: {"scenario":<content>}
2. Output 2: Multiple-choice question.
Output format: {"question":<q>,"options":{"A":<o1>,"B":<o2>,"C":<o3>,"D":<o4>}}

# Execution Steps:
1. Read all inputs.
2. Output the scenario in which the character faces this decision.
3. Output the multiple-choice question, ensuring that the incorrect options are also reasonable.

Table 12: Prompt templates for constructing multiple-choice questions.

Prompt IV

Please play the role of <Character A> based on the <Profile> and make your life choice under the <Scenario> regarding <Question>. Return the option letter (A, B, C, or D) that your character should most appropriately choose in the current scenario. The <Profile> consists of <Description> and <Memory>, where <Description> is an overall description of the character, and <Memory> consists of specific events the character has experienced.

# Inputs:
1. Profile:
1.1. Description
<description>
1.2. Memory
<memory>
2. Scenario:
<scenario>
3. Question:
<question>
4. Options:
<option>

# Outputs:
Your choice(A, B, C, or D):

Table 13: Prompt templates for role-playing as the character.

Prompt V

Your task is to find segments within a character's <Description> that may relate to the content of the <Scenario> and <Question>. The <Scenario> describes the situation the character is in, and the <Question> asks what choice the character should make. The segments you need to find could influence the character's motivation for making their choice, including aspects that shape the character's personality and foreshadowing related to the decision scenario.

# Inputs:
1. Description:
<description>
2. Scenario:
<scenario>
3. Question:
<question>

# Outputs:
Segments that may influence the character's choice:

Table 14: Prompt templates for CHARMAP.

Manual Examination Rules

1. Comprehensiveness

Rule:

Evaluators must ensure that each multiple-choice question fully considers the character's background, context, and motivation. The questions should reflect the true decisions and experiences of the character within the narrative.

Scoring Guide:

Score 2 (Excellent): The question is detailed and comprehensive, aligning perfectly with the character's background and motivation.

Score 1 (Average): The question aligns generally but is missing key aspects of the character's background information or motivational nuances.

Score 1 (Poor): The question significantly misaligns with the character's background or motivation.

2. Logical Consistency

Rule:

Evaluators should assess the internal consistency and plausibility of the question within the narrative thread. The content and structure of the multiple-choice question must be consistent with the plot and the character's logical decision-making process.

Scoring Guide:

Score 2 (Excellent): The question is entirely consistent with the character's known decisions and the structure of the plot.

Score 1 (Average): The question is generally consistent but has minor inconsistencies in detail.

Score 0 (Poor): The question is logically inconsistent with the character's known decisions or the structure of the plot.

3. Challenge Level

Rule:

Evaluators need to assess the plausibility of the incorrect options. Wrong options should be reasonably believable and attractive within the constraints of the character's background and motivations, making the questions sufficiently challenging.

Scoring Guide:

Score 2 (Excellent): All incorrect options are highly plausible and convincingly misleading.

Score 1 (Average): Most incorrect options are reasonable, but one or two lack plausibility.

Score 0 (Poor): Incorrect options are obviously illogical and lack the ability to mislead.

4. Alignment with Character Motivation

Rule:

Evaluators must assess whether the question correctly guides the testing model to step into the role and make a choice, i.e., testing if the model can replicate the real storyline's choices. It is crucial that the character's motivations, as articulated by literary experts, are a central component reflected in these questions.

Scoring Guide:

Score 2 (Excellent): The question unambiguously points to a specific character decision point, accurately testing the model's ability to role-play.

Score 1 (Average): The question points to a character decision point to some extent, but the indicators are not clear enough, potentially reducing the accuracy of the model's role-playing test.

Score 0 (Poor): The question fails to clearly define the character decision point, unable to test the model's role-playing ability effectively.

Additional Notes:

1. Before starting the evaluation, each evaluator must understand the core motives and development axes of the character by reading summaries and analyses of the novels created by literary experts.
2. Ensure that evaluators are familiar with all background material before scoring any questions.
3. Evaluators should reference the analyses by literary experts of the characters to evaluate each of GPT-4's multiple-choice questions, maintaining consistency of standards.
4. Application of the evaluation rules should be flexible and adapted to the specific context; scoring standards may be adjusted for special cases.

Table 15: Guidelines for Manual Examination of Multiple-Choice Questions in Literary Analysis.