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

Anonymous ACL submission

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 1,462 characters' decision points from 388 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, achieving 5.03% increase in accuracy. We will make our dataset and code publicly available.

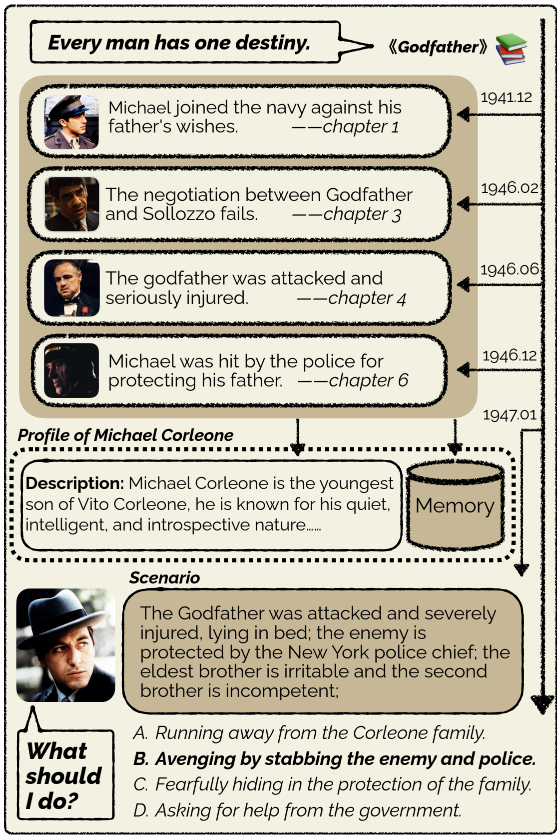

Figure 1: An example of LIFECHOICE. Given a character, a decision point and the preceding context, RPLAs are expected to reproduce the original decision. Typically, RPLAs are constructed by parsing the context into the character's description and memory.

1 Introduction

The fault, dear Brutus, is not in our stars,
but in ourselves, that we are underlings.
– *Julius Caesar*. Act 1, Scene 2.

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). 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 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 1,462 character decisions from 388 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, our questions and decisions are well-designed and closely aligned with the personas, since they are sourced from well-crafted narratives. Hence, our data establish solid ground truth for simulating characters' persona-driven decisions. 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 need to identify 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.

We conduct extensive experiments to evaluate RPLAs on LIFECHOICE. Our experiments cover various LLMs and different RPLA frameworks, including memory-enhanced agents, long-context LLMs, and our proposed method CHARMAP towards better simulation of persona-driven decisions. The results demonstrate that existing RPLAs have shown a promising accuracy of up to 62.92% on LIFECHOICE. Furthermore, CHARMAP significantly enhances RPLAs on this task, achieving an accuracy of 67.95%, which exceeds previous baselines by 5.03%. However, compared to the human performance of 92.01%, there is still significant room for improvement. Meanwhile, we observe that both well-summarized character descriptions and accurate memory retrieval are crucial for 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 Early research on character-related studies focuses on character understanding. Brahman et al. (2021) attempts to predict a specific character through the text of the novel. Yu et al. (2022) provides dialogues from movie scripts for the model to examine and then asks it to identify the character who speaks each passage. With the enhancement of model abilities, some work attempts to make the model simulate complex role-playing. Li et al. (2023) analyzes 32 anime characters using 54k dialogues and personality traits. They use sentence embeddings for dialogue selection and evaluation. Zhou et al. (2023) uses identity, interests, and relationships, collecting AI behaviors for imitation and using character data for fine-tuning. They evaluate model consistency and linguistic style. Wang et al. (2023) creates a dataset for script characters and evaluates role-playing quality based on speaking style imitation and role-specific knowledge. These studies make a chatbot for a certain character, but they focus more on imitating the character from the perspective of dialogue, which is a shallow imitation. We aim to role-play from the perspective of behavior and decision-making. This form tests the model's understanding of the role more.

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 can model the user's profile and preferences through the user's historical data (Gurrin et al., 2014; Dodge and Kitchin, 2007), such as extracting personality from the user's record text (Majumder et al., 2017; Štajner and Yenikent, 2020), reading emotions from the user's image data (Jaiswal et al., 2020; Zad et al., 2021), modeling preferences from historical interaction information (Tang et al., 2019; Li et al., 2018), and pushing notifications from smart phones (Li et al., 2018). These memories can enhance the model's decision-making and reasoning, bringing a better personal experience for users. However, obtaining real user memory data is difficult and violates privacy. We model characters from historical data in high-quality novel texts, allowing the model to restore the real choices in the storyline based on the previous text, providing the first benchmark for the wide testing of personal intelligent agents.

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*, *full-text and 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 of being him 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

[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.

books that are overly popular, 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. We analyze each key character's life choice decision points and the corresponding gold motivations. Additionally, we have GPT-4 identify the corresponding chapters 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 Question 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. The specific annotation

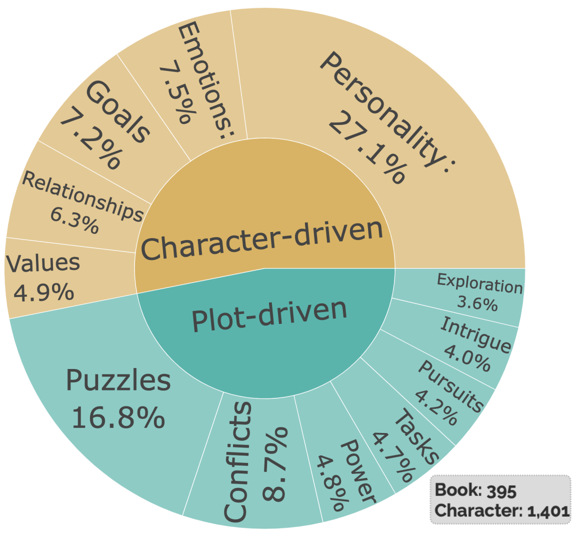

Figure 2: Statistics of motivation types in LIFECHOICE, with the first words for each motivation type.

rules are available in Appendix B.1.

Ultimately, we collect 1,401 characters from 396 books and their corresponding life choices. Table 1 shows a complete data example.

3.2 Dataset Analysis

We refer to the drama theory of Aristophanes (Sommerstein, 2013; Silk, 2002) as the system prompt and use GPT-4 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 A.2.

3.3 Task Setups

This task can be formulated as $P(y|x)$. Given the input $x = (p, s, c, q, a)$, the RPLA needs to identify the correct choice y that aligns with the character's 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. Below, I will detail the methods for constructing these two parts:

Description Construction We adopt two automatic methods to construct character descriptions: (1) Hierarchical 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) Incremental updating 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 human-written descriptions.

Profile Construction	**Role-Playing Model**	**ACC**	***+motivation***
Description Construction			
Hierarchical merging	LLaMA-3	42.10	83.09
	GPT-3.5	39.85	80.00
	GPT-4	45.43	85.24
Incremental updating	LLaMA-3	43.82	83.21
	GPT-3.5	41.06	81.63
	GPT-4	47.02	86.47
Human Description	LLaMA-3	52.51	87.28
	GPT-3.5	52.04	86.33
	GPT-4	55.17	90.23
Memory Retrieval			
BM25	GPT-4	26.08	75.88
Embedding	GPT-4	35.66	78.24
Description & Memory			
Direct concatenation	LLaMA-3	57.02	92.04
	Mixtral	58.56	91.75
	Claude-3	59.85	93.45
	Gemini-1.5-pro	57.16	91.38
	GPT-3.5	55.62	90.39
	GPT-4	62.92	95.46
CHARMAP	LLaMA-3	63.72	95.93
	Mixtral	65.02	92.05
	Claude-3	65.13	93.61
	Gemini-1.5-pro	63.94	91.39
	GPT-3.5	61.62	90.95
	GPT-4	**67.95**	**96.87**

Table 3: Results of different LLMs on LIFECHOICE. **ACC** refers to the decision accuracy. ***+motivation*** refers to the results after providing the motivations behind character decisions, which are extracted from expert analyses by GPT-4.

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.

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 meth-

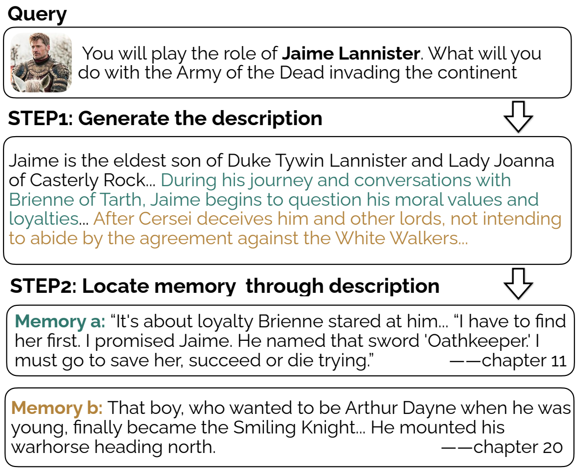

Figure 3: An overview of CharMap, a two-step scenario-specific character profile building approach.

ods to form the character's profile. We adopt two methods: (1) Direct concatenation: This method concatenates the results from both approaches by prompting the user 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 GPT-3.5, GPT-4, and LLaMA-3(Team, 2024b). For methods using both, we also include Claude-3(Anthropic, 2024), Gemini(Team, 2024a), and Mixtral (Jiang et al., 2024). For all these models, we adopt the official instruction formats where available [2].

[2]The versions in this paper are `gpt-3.5-turbo-1106`, `gpt-4-1106-preview`, `Llama-3-70B-Instruct`, `Claude-3-Sonnet`, `Gemini-1.5-pro` and `Mixtral-8x7B-v0.1` respectively.

	Raw text	Concat.	CharMap
GPT-4	-	65.92	71.99
human	92.01	66.82	74.78

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 three research questions: *RQ1)* Can LLMs make decisions based on historical data? *RQ2)* What influences the decision-making of LLMs?

5.1 Can LLMs make decisions based on historical data?

Analysis of Model Results Table 3 presents the accuracy results of different RPLA methods on the LifeChoice. Additionally, we evaluate the results when the model is provided with gold motivation, and several observations can be made: First, 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. Second, when gold motivation is provided, the accuracy consistently exceeds 80%, indicating the rationality of these motivations in the data. Third, the performance gap among different LLMs is not significant 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 5.03%, 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 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 sub-

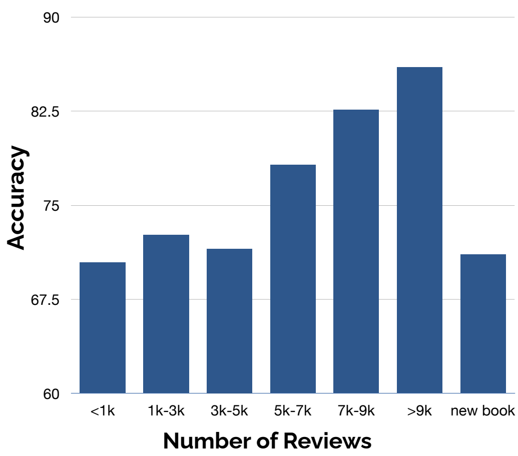

Figure 4: The impact of the number of book reviews on accuracy in LIFECHOICE, with *new books* being those not present in the training corpus of LLMs.

tle 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.

Mitigation and Analysis of Data Leakage Data leakage is a significant challenge since our data might appear in the model's pre-training corpus. During the data collection phases in section 3.1, we adopt various preventive measures. For evaluation, we employ an entity replacement strategy, substituting character names, locations, and other entities with placeholders. We believe data leakage relates to the amount of relevant corpus used during LLM pre-training, with more popular books having more related corpus. To verify this, we use the number of reviews on the book review website[3] to indicate a book's popularity and evaluate the results of books with different review counts on LIFECHOICE. We use CHARMAP to build profiles and GPT-4 as the role-playing model, sampling thirty books with different numbers of reviews, including thirty books not in the LLMs' corpus (published after November 6 for gpt-4-1106-preview). As shown in Figure 4, the model's accuracy significantly improves when the number of reviews exceeds 5,000. In contrast, books with fewer than 5,000 reviews show slight fluctuation and results similar to those not in the LLMs' corpus. Therefore, it can be considered that for books with a low number of reviews, data leakage has little impact on CHARMAP. In section 3.1, we use 5,000 reviews as a threshold to filter the books.

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

LLMs	Method	Accuracy
Claude3	long-context	64.95
Claude3	CHARMAP	68.13
Kimi-chat	long-context	61.14
Kimi-chat	CHARMAP	64.01

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

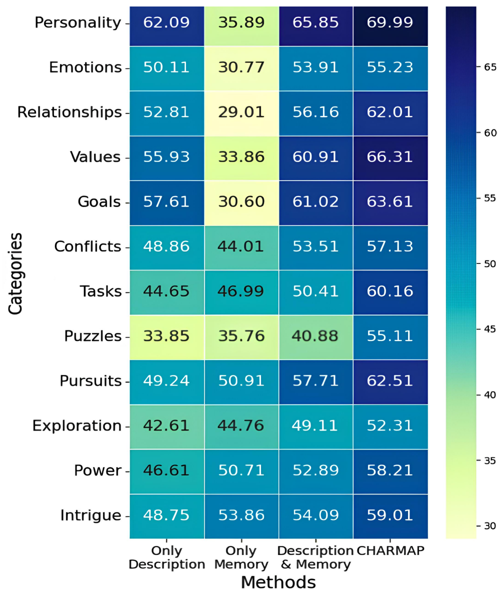

Figure 5: Heatmap of the impact of motivation types on the results. The results are predicted from the Incremental updating, the embedding-retrieved memory, the direct concatenation of both, and CHARMAP. The role-playing model uses GPT-4.

Analysis of Long-Context LLMs Long context is an essential feature of LIFECHOICE, and directly using long-context models for role-playing is an exciting topic. Making decisions based on extensive context tests a model's ability to understand global data and reason from a character's perspective. We evaluate two long-context models: Claude3-sonnect and kimi-chat. As shown in Table 5, although the performance of long-context models is not as strong as CHARMAP, they still demonstrate potential in role-playing. LIFECHOICE, as a task requiring multiple reasoning points and an overall understanding of the context, can also serve as a vital benchmark for evaluating long-context models.

5.2 What influences the decision-making of LLMs?

The Impact of Motivation Types In line with the motivation types presented in Section 3.2, we examine how different types of motivation influ-

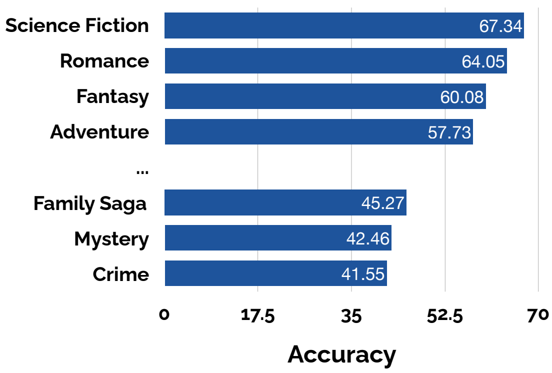

Figure 6: The result of the impact of different novel genres on accuracy.

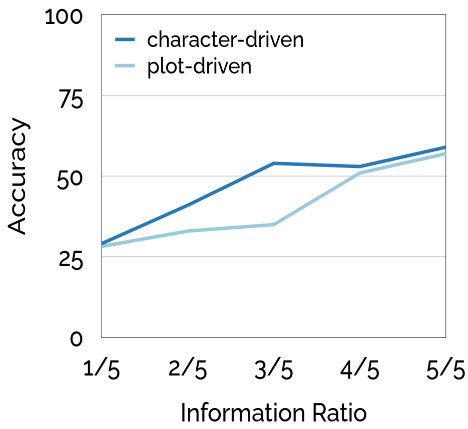

Figure 7: Analysis of whether character selection will change. The x-axis represents the input length relative to the point truncation.

ence 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 use the genre tags from novels on the website to analyze the accuracy of character selection across different genres. We conduct experiments on the the direct concatenation of description and memories, and the role-playing model using GPT-4. As depicted in Figure 6, the accuracy of science fiction, fantasy novels, and romance novels is quite high. This could be because the characters in these novels are often stylized or have fixed creative patterns and archetypes. In contrast, crime and mystery novels perform poorly, which might be because they involve complex logical chains, and characters in these novels frequently take abnormal actions. Further details about each genre and the complete table can be found in Appendix A.1.

The Impact of Temporal Data If faced with the decisions of years past at this moment, would you make the same choices? We conduct a study on this matter. 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 + embedding-retrieved memories, and the role-playing 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.

6 Conclusion

In this work, we propose the first task to evaluate the decision-making of RPLAs, testing whether LLMs can accurately reconstruct storylines using historical data. We construct LIFECHOICE, which includes 1,462 characters from 388 books and their life choices. Extensive experiments on LIFECHOICE demonstrate the promising performance of RPLAs in decision simulation. Additionally, we propose CHARMAP, which uses persona-based memory retrieval to enhance decision-making. We hope this work provides better evaluation benchmarks for RPLAs and directs the future development of personal LLM assistants.

Limitations

The partial evaluation method we proposed is dependent on GPT-4, which could be biased towards GPT-4 generations. Finally, our dataset is constructed through the decision of high-quality novel characters. However, compared to human choice, this part of the data is not sparse or challenging enough. We hope to construct real human decision-making data while ensuring privacy.

References

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

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. " let your characters tell their story": A dataset for character-centric narrative understanding. *arXiv preprint arXiv:2109.05438*.

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.

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.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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.

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

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.

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.

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. 2023. Gpt-4 technical report.

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.

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

Meta LLaMA Team. 2024b. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2023-10-03.

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.

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.

Mo Yu, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, 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 preprint arXiv:2211.04684*.

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 Dateset Details

A.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.

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 magical elements do not exist in real life, but they are perfectly normal in the realm of magical realism.

A.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.

B Manual Annotation

For all individuals involved in the annotation, we provide compensation based on the local minimum hourly wage.

B.1 Manual Examination Rules

This is a supplement to Section 3.1. After constructing the multiple-choice question data using GPT-4, we perform manual examination.

For each annotator, we provide novel summaries and character analyses written by human literature experts on the Supersummary website. Each annotator is asked to score the questions constructed by GPT-4 based on the following evaluation criteria:

1. Comprehensiveness

Rule 1.1: 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:

2 points: The question is detailed and comprehensive, aligning perfectly with the character's background and motivation.

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

0 points: The question significantly misaligns with the character's background or motivation.

2. Logical Consistency

Rule 2.1: 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:

2 points: The question is entirely consistent with the character's known decisions and the structure of the plot.

1 point: The question is generally consistent but has minor inconsistencies in detail.

0 points: The question is logically inconsistent with the character's known decisions or the structure of the plot.

3. Challenge Level

Rule 3.1: 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:

2 points: All incorrect options are highly plausible, convincingly misleading.

1 point: Most incorrect options are reasonable, but one or two lack plausibility.

0 points: Incorrect options are obviously illogical and lack the ability to mislead.

4. Alignment with Character Motivation

Rule 4.1: 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:

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

1 point: 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.

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

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.

We evaluated the scores of each annotator and only retained the data with an average score of more than 6 points.