Evaluating Character Understanding of Large Language Models via Character Profiling from Fictional Works

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated impressive performance and spurred numerous AI applications, in which role-playing agents (RPAs) are particularly popular, especially for fictional characters. The prerequisite for these RPAs lies in the capability of LLMs to understand characters from fictional works. Previous efforts have evaluated this capability via basic classification tasks or characteristic imitation, failing to capture the nuanced character understanding with LLMs. In this paper, we propose evaluating LLMs' character understanding capability via the character profiling task, *i.e.*, summarizing character profiles from corresponding materials, a widely adopted yet understudied practice for RPA development. Specifically, we construct the CROSS dataset from literature experts and assess the generated profiles by comparing ground truth references and their applicability in downstream tasks. Our experiments, which cover various summarization methods and LLMs, have yielded promising results. These results strongly validate the character understanding capability of LLMs. Resources of this paper will be released upon publication.

1 Introduction

The recent progress in large language models (LLMs) (OpenAI, 2023; Anthropic, 2024) has catalyzed numerous AI applications, among which role-playing agents (RPAs) have attracted a wide range of audiences. RPAs are interactive AI systems that simulate various personas for applications, including chatbots of fictional characters (Wang et al., 2023c), AI none player characters in video games (Wang et al., 2023a), and digital replicas of real humans (Gao et al., 2023a). In practice, LLMs are generally prompted with character profiles to role-play fictional characters (Wang et al., 2023b; Zhao et al., 2023), and these profiles are typically generated through the automatic summarization of corresponding literature using advanced LLMs (Wang et al., 2023c; Li et al., 2023a).

Previous efforts have studied LLMs' capabilities of understanding characters from fictional works. The research on character understanding mainly concentrates on basic classification tasks, such as character prediction (Brahman et al., 2021; Yu et al., 2022; Li et al., 2023b) and personality prediction (Yu et al., 2023), which aims at recognizing characters or predicting their traits from given contexts correspondingly. Recently, the research focus has shifted to character role-playing, primarily focusing on the imitation of characteristics such as knowledge (Tang et al., 2024; Shen et al., 2023) and linguistic style (Zhou et al., 2023; Wang et al., 2023c). Hence, these tasks fail to capture the nuanced character understanding of LLMs.

In this paper, we systematically evaluate LLMs' capability on the **character profiling** task, *i.e.*, summarizing profiles for characters from fictional works. For research, character profiling is indeed the first task to explore the depth of LLMs' character understanding via generation. This is more challenging than previous classification tasks, contributing to a more nuanced comprehension of how LLMs understand the character. In practice, the character profiles generated by LLMs have been widely adopted for RPA development (Wang et al., 2023c; Li et al., 2023a; Xu et al., 2024), and have the potential to facilitate human understanding of characters, but their effectiveness remains significantly understudied. Our work in this paper aims to evaluate LLMs' performance on character profiling, of which the challenges mainly include the absence of high-quality datasets and evaluation protocols.

To address these challenges, we construct the CROSS (**C**haracter **Pro**files from ***SuperSummary***) dataset for character profiling, and propose two tasks to evaluate the generated profiles. The CROSS dataset is sourced from `SuperSummary` [1], a

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

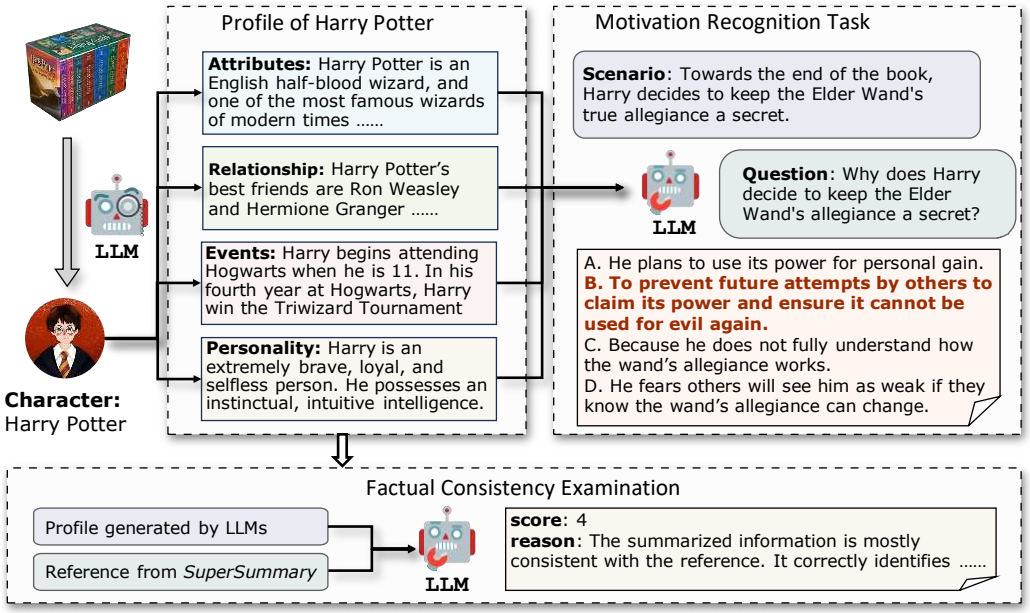


Figure 1: An overview of character profiling with LLMs and the two evaluation tasks we proposed, including factual consistency examination and motivation recognition.

platform providing summaries for books and characters contributed by literature experts. Our evaluation distinguishes four essential dimensions for character profiles: attributes, relationships, events, and personality. We parse the character profiles from SuperSummary into these dimensions by GPT-4, as the ground truth references. Then, the generated profiles are evaluated in either an intrinsic or extrinsic way. The intrinsic evaluation directly employs Llama-3-70B (AI@Meta, 2024) to compare the generated profiles with the references. For extrinsic evaluation, we propose the *Motivation Recognition* task and measure whether the generated profiles can support LLMs in this task, *i.e.*, identifying the motivations behind characters' decision-making.

Our experiments cover various summarization methods, including *Hierarchical Merging*, *Incremental Updating* and *Summarizing in One Go*, implemented on numerous LLMs. The results reveal that character profiles generated by LLMs are satisfactory but leave space for further improvement. This suggests the potential information loss in RPAs built on these profiles. Additionally, the results of *Motivation Recognition* demonstrate the importance of each of the four dimensions for character profiles.

Our contributions are summarized as follows: *1)* We present the first work to evaluate LLMs' capability of character profiling and propose an evaluation framework with detailed dimensions, tasks and metrics. *2)* We introduce CROSS, a high-quality dataset valuable for character profiling tasks, which is sourced from literature experts. *3)* We conduct extensive experiments with different summarization methods and LLMs and showcase the promising effectiveness of using LLMs for character profiling.

2 Related Work

Character Role-Playing Recent advancements in LLMs have significantly enhanced the capabilities of role-playing agents (RPAs) across various aspects. Currently, many role-playing tasks require interactive AI systems to act as assigned personas, including celebrities and fictional characters. In these studies, researchers have utilized various methods to develop RPAs, which can be divided into three categories: *1) Manual Construction* (Chen et al., 2023; Zhou et al., 2023), which employed book fans or professional annotators to label information related to characters; *2) Online Resource Collection* (Shao et al., 2023; Tu et al., 2024), which collects character profiles from online resources, *e.g.*, Wikipedia [2], and Baidu Baike [3]; *3) Automatic Extraction* (Li et al., 2023a; Zhao et al., 2023), which utilizes LLMs to extract character dialogues from origin books or scripts. In this

[2] https://en.wikipedia.org/wiki/Main_Page
[3] https://baike.baidu.com/

paper, we explore the capabilities of LLMs in generating character profiles for RPAs construction.

Motivation Analysis & Character Understanding Motivation is a fundamental concept, which is shaped by personality traits and the immediate surroundings (Young, 1961; Atkinson, 1964; Kleinginna Jr and Kleinginna, 1981). In narrative texts, the motivation of a character can reveal their inner traits and their relationship with the external world. Thus, understanding the motivation of characters strongly aligns with the LLMs' ability to comprehend characters. Previous studies typically propose benchmarks in character identification (Chen and Choi, 2016; Brahman et al., 2021; Sang et al., 2022; Yu et al., 2022), situated personality prediction (Yu et al., 2023), question answering (Kočiskỳ et al., 2018; Anthropic, 2024). Despite these efforts, prior research has not focused on assessing a character's motivation based on character profiles. To bridge this gap, we propose the motivation recognition task. This task aims to directly evaluate whether LLMs can grasp a character's essence by identifying the motivations behind each decision within a story.

3 Character Profiling Framework

3.1 Task Formulation

Character profiling aims to generate profiles for fictional characters from corresponding literature. Given the input character name $\mathcal{N}$ and the original content $\mathcal{B}$ of a fictional work, the LLM should output the character profile $\mathcal{P}$ which covers the core information about the character. Specifically, in this paper, $\mathcal{P} = (\mathcal{P}_{attributes}, \mathcal{P}_{relationships}, \mathcal{P}_{events}, \mathcal{P}_{personality})$ is structured in four dimensions, as detailed in Section 3.2. An example of a character profile is presented in Figure 1.

3.2 Character Profile Dimensions

For a character, his/her profile should be highly complex and multi-faceted, embodying diverse information. Drawing inspiration from previous studies and current developments in persona products (Zhao et al., 2023; Baichuan, 2023), we define four main profile dimensions for LLMs to summarize, which are commonly examined in literary studies (Yu et al., 2023; Zhao et al., 2024; Shen et al., 2023). Please refer to Appendix B for a further comparison.

Attributes The basic attributes of a character encompass gender, skills, talents, objectives, and background.

Relationships A character's interpersonal relationships are a vital aspect of their profile, which are intimately connected to the character's experiences and their personality. Moreover, these relationships can serve as a foundation for constructing fictional character relationship diagrams.

Events Events cover the experiences that characters have been part of or impacted by, marking a critical profile dimension. Due to the complexity of certain narratives, such as alternating timelines and showcasing events from diverse worlds or different perspectives, we require the model to rearrange events and order them chronologically.

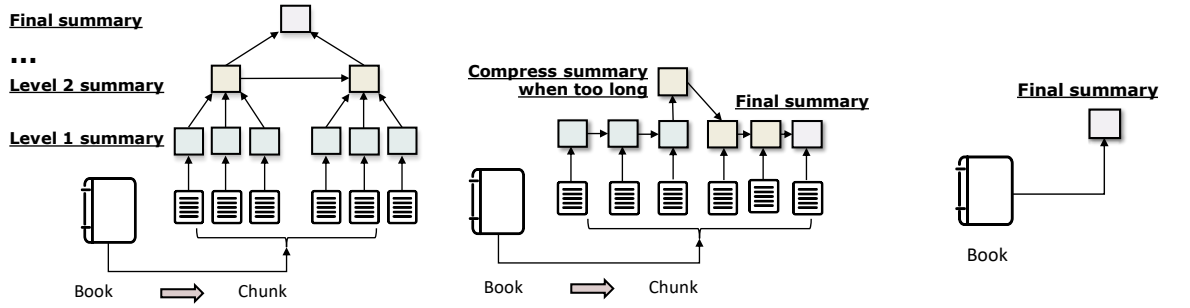
Personality Personality refers to the lasting set of characteristics and behaviors that form an individual's unique way of adapting to life (American Psychological Association, 2018). A well-rounded character often exhibits a complex personality. It can analyze a character's personality through their actions, choices, and interactions with others.

3.3 Summarization Methods

Book-length texts often comprise over 100,000 tokens, surpassing the context window limitations of many current LLMs. As a result, the primary framework for long context processing involves segmenting books into manageable segments for LLMs, followed by subsequent comprehensive processing. As illustrated in Figure 2a and Figure 2b, we inherit two methods for book summarization (Chang et al., 2023), *i.e.*, hierarchical merging and incremental updating. Additionally, for models that can handle long context windows, we explore the method of summarizing in one go, as shown in Figure 2c.

Hierarchical Merging The hierarchical merging approach (Wu et al., 2021) employs a simple, zero-shot prompt technique. It begins by summarizing information from segments within a book, generating the summaries at level 1. Then, several summaries are combined to establish the initial context at level 2. Subsequently, it merges the following summaries with context iteratively. The merging process continues at the next level until a final summary is generated.

Incremental Updating One major issue with hierarchical methods lies in constructing summaries



(a) Hierarchical Merging (b) Incremental Updating (c) Summarizing in One Go

Figure 2: The three methods of long context processing for LLM-based character profiling.

at level 1. As shown in Figure 2a, the provided text only contains novel content from the current segment without any background information from earlier segments. Thus, this absence of context may increase the risk of misinterpreting information in later segments.

In response, Chang et al. (2023) introduces incremental updating. This method leverages background information from the preceding text to enhance summary quality. The process of incremental updating consists of three phases: First, it starts by summarizing the book's opening segment. Then, this summary is refined and updated by incorporating details from the following segments recursively. Throughout this process, to ensure conciseness and relevance, the summary is periodically condensed to comply with a set pre-defined maximum length. By following these steps, the method seeks to promote a more integrated and coherent comprehension of the entire text.

Summarizing in One Go Recent developments in LLMs have introduced models capable of processing over 100,000 tokens. For example, GPT-4-Turbo (OpenAI, 2023) supports a context window of up to 128,000 tokens. This advancement enables us to explore a method for inputting the full content of a book into the model in one step. For this investigation, we select books from our dataset that contain fewer than 120,000 tokens.

Character: Nora Stephens **Character Profile**: Attributes:Nora Stephens is a tall, thin, ambitious female literary agent with dyed platinum blonde hair, known for her pragmatic approach and dedication to her career. She has a history of troubled. . . **Scenario**: Nora decides not to pursue a long-distance relationship with Charlie after the summer ends. **Question**: Why does Nora make the decision to part ways with Charlie? **Options**: *A. Because she fears long-distance relationships are doomed to fail.* B. Because she believes she needs to focus on her personal growth and independence. C. Because she feels their goals and aspirations are no longer aligned. D. Because she worries that their frequent arguments are harming her well-being.
Model Reasoning Output
Choice: "A", **Reason**: Nora's decision . . . is primarily motivated by her fear that long-distance relationships are doomed to fail. Given her history of being dumped and her protective nature due to her family's past, Nora is likely cautious about entering a relationship that has inherent challenges and uncertainties. . .

Table 1: A toy example of MR task. A complete set of data includes character name, character profile, scenario, question, options, correct answer, and reason. The reasoning model is GPT-4-Turbo-0409 [1].

4 Evaluation Protocol

4.1 Evaluation Tasks

Intrinsic Evaluation: Factual Consistency Examination (FCE) To generate character profiles from books, we implement the three methods previously described. Throughout the summarization process, we require the model to produce four distinct sections, each detailing one dimension of a character's profile. An excellent profile should accurately cover all the important information about the character across these four dimensions. Therefore, we evaluate factual consistency by comparing the model-summarized profile with the reference profile. The metrics for this examination are intro-

[1] In this paper, unless otherwise specified, we adopt the version of `GPT-4-Turbo-0125` throughout.

duced in Section 4.2.

Extrinsic Evaluation: Motivation Recognition (MR) As shown in Table 1, to thoroughly evaluate whether the summarized profiles enhance models' understanding of a character's essence, we introduce a *Motivation Recognition* task for downstream evaluation. This task investigates if the character profiles generated by the model effectively aid in comprehending the characters, particularly in recognizing the motivations behind their decisions.

Given the input $\mathcal{X} = (\mathcal{N}, \mathcal{P}, \mathcal{D}, \mathcal{Q}, \mathcal{A})$, which includes the character name $\mathcal{N}$, the character profile $\mathcal{P}$ defined by four dimensions, the character's decision $\mathcal{D}$, a question $\mathcal{Q}$ about the motivations behind the decision, and a set of potential answer $A = \{a_i\}_{i=1}^4$ for $\mathcal{Q}$, the LLMs should determine the answer $\mathcal{Y}$ from A that correctly reflects the character's motivation. Details of MR dataset construction are provided in Section 4.3.

4.2 Evaluation Metrics

Metric for FCE: Consistency Score As demonstrated in a previous study (Goyal et al., 2022), current reference-based automatic metrics like ROUGE metric (Lin, 2004) exhibit a significantly low correlation with human judgment for summaries generated by GPT-3. Therefore, we adopt the evaluation method used in recent research (Liu et al., 2023; Gao et al., 2023b; Li et al., 2024), utilizing an LLM as an evaluator for improved alignment with human perception and reduced cost. [4] Specifically, we introduce **Consistency Score**, which is the degree of factual consistency between the reference profiles and the summaries generated by LLMs, evaluated by Llama-3-70B. We ask Llama-3-70B to assign a score on a scale from 1 to 5, reflecting the accuracy of the summaries in capturing the essential factual details. A higher score indicates a closer match to the factual content.

To evaluate the quality of the LLM evaluation, we randomly select 50 samples for human evaluation. We calculate the Pearson Correlation Coefficient (Cohen et al., 2009) between the consistency score result of human annotators and Llama-3-70B. The coefficient value of 0.752 with the p-value $= 4.3\text{e}{-12} < 0.05$ suggests that these two set of results have a significant correlation. This validates that the evaluation capabilities of Llama-3-70B for this task are comparable to those of humans.

Metric for MR: Accuracy Multiple-choice questions can be easily evaluated by examining the choice of models. We define `Acc` as the accuracy across the entire question dataset.

4.3 CROSS Dataset Construction

Book Dataset To reduce the confounding effect of book memorization on the results, we select 126 high-quality novels published in 2022 and 2023.[5] For each novel, we concentrate solely on its main character. We manually remove sections not pertinent to the novel's original content, such as prefaces, acknowledgments, and author introductions. Additionally, we select 47 books within CROSS containing fewer than 120,000 tokens for the summarizing-in-one-go method.

Golden Character Profile Extraction The golden character profiles are gathered from the `SuperSummary` website, known for its high-quality plot summaries and character analyses conducted by literary experts. With permission from the site, we utilize their book summaries, chapter summaries, and character analyses. The original character analysis from `SuperSummary` lacks a standardized format and predefined profile dimensions. Therefore, we utilize GPT-4 to reorganize the original summaries.

Given the original plot summaries and character analyses, we require the model to reconstruct character profiles across four key dimensions while ensuring no critical details are overlooked. To guarantee the quality of the reorganized profiles, two annotators evaluate whether the reorganized profiles adequately retained the essential information from the original text. The assessment reveals that all results exhibit a high level of informational integrity and consistency, confirming the credibility of the reorganized profiles.[6]

MR Dataset Construction Using resources from the `SuperSummary` website, we develop motivation recognition questions for key characters in CROSS. The process involves four main steps: First, we utilize GPT-4 to generate several motivation recognition multiple-choice questions (MCQs)

[4] The result of existing evaluation metrics is provided in Appendix F.

[5] Details on the construction process and integrity verification experiments of the CROSS dataset can be found in Appendix C.

[6] The detail of human filtering is shown in Appendix E.1.

Summarization Method	Summarization Model	Consistency Score					MR Acc.
		Attr	Rela	Even	Pers	Avg.	
CROSS *(Full dataset)*							
Incremental Updating	Mistral-7B-Instruct-v0.2	2.75	2.20	1.88	3.89	2.68	48.31
	Mixtral-8x7B-MoE	2.75	2.58	2.28	4.02	2.91	52.13
	vicuna-7b-v1.5-16k	2.44	1.72	1.45	3.17	2.20	42.70
	vicuna-13b-v1.5-16k	2.79	2.22	1.76	3.56	2.58	46.29
	Qwen1.5-7B-Chat	2.35	1.98	1.58	3.75	2.42	44.49
	Qwen1.5-14B-Chat	2.39	2.18	1.41	3.74	2.43	47.42
	Qwen1.5-72B-Chat	3.33	2.71	2.45	4.08	3.14	52.36
	GPT-3.5-Turbo	3.49	2.57	1.95	3.95	2.99	49.44
	GPT-4-Turbo	3.72	3.24	**3.58**	3.87	3.60	**57.75**
Hierarchical Merging	Mistral-7B-Instruct-v0.2	3.07	2.20	1.98	3.83	2.77	50.56
	Mixtral-8x7B-MoE	3.17	2.59	2.03	3.93	2.93	48.09
	vicuna-7b-v1.5-16k	2.40	1.77	1.40	3.08	2.16	44.94
	vicuna-13b-v1.5-16k	2.91	2.12	1.54	3.27	2.46	45.39
	Qwen1.5-7B-Chat	3.05	2.37	1.88	3.83	2.78	44.04
	Qwen1.5-14B-Chat	3.29	2.70	2.21	4.04	3.06	47.42
	Qwen1.5-72B-Chat	3.67	2.97	2.98	4.21	3.46	54.61
	GPT-3.5-Turbo	3.29	2.87	2.17	3.90	3.06	51.69
	GPT-4-Turbo	**3.81**	**3.48**	3.36	**4.23**	**3.72**	53.71
CROSS *(Short subset)*							
Sum-in-One-Go	GPT-4-Turbo	**3.98**	**3.83**	**3.72**	**4.28**	**3.95**	56.79
	Claude3-Sonnet	3.81	3.32	3.57	4.11	3.70	**61.11**
Incremental	GPT-4-Turbo	3.66	3.47	3.62	3.72	3.62	**61.11**
Hierarchical	GPT-4-Turbo	3.66	3.62	3.38	4.09	3.69	51.85

Table 2: Results of different LLMs performance on character profiling and motivation recognition. The abbreviations used in this table stand for the following terms: 'Attr' represents 'Attributes'; 'Rela' stands for 'Relationships'; 'Even' denotes 'Events'; 'Pers' indicates 'Personality'; 'Avg.' refers to the mean values for the scores across the four dimensions. The best scores are **bolded** and the second best scores are underlined.

Profile Method	Ablation Dimension	Acc. %	Std. %
Reference Profile			
CROSS	-	**63.07**	0.11
Generated Profile (GPT-4-Turbo)			
Incremental Updating	-	57.75	0.32
	Attr	57.38	0.11
	Rela	57.30	0.37
	Even	48.54	0.32
	Pers	57.08	0.31
	Attr&Rela	56.93	0.28
	Attr&Rela&Even	42.62	0.56
	Attr&Rela&Even&Pers	40.90	0.73

Table 3: Results of Motivation Recognition Ablations study. **Ablation Dimension** refers to omitted dimensions in experiments.

and manually select the best top 10 examples. Second, we identify a primary character from each of the 126 books and formulate questions related to them. Given the character's name, chapter summaries from the `SuperSummary`, and the 10 examples, GPT-4 is instructed to generate a set of motivation recognition multiple-choice questions in a few-shot scenario. Each question is designed to include a decision made by the character within a specific scenario, offering four options, the correct answer, and justifications for the correctness or incorrectness of each option. Through this process, GPT-4 generates a total of 641 questions for the 126 characters. Moreover, we find that some questions can be easily answered using commonsense knowledge or grammatical structure, Thus, given a question and the correct answer, we ask GPT-4 to provide three likely motivations behind the decision in the question that differ from the correct answer. These options, meant to confuse, are similar to the correct answer in sentence structure. We replace the incorrect options generated in the previous step with these three motivations.

To maintain the quality of MR questions, two annotators are assigned to filter them, with Fleiss's $\kappa = 0.91$ (Fleiss et al., 1981). According to the annotation results, 445 out of the 641 questions meet the established criteria, guaranteeing the quality of the MR questions dataset.[7]

[7]Further details are shown in Appendix E.

Error type	Generated Profile	Golden Profile
Character Misidentification	Benjamin's relationships are complex and multi-faceted. He is *married to Mildred*, a woman of delicate health and refined tastes · · ·	Rask *marries Helen Brevoort*, a woman from an old-money New York family with a similarly reserved personality· · ·
Relationship Misidentification	Benjamin's life takes a dramatic turn when he saves *his grandson, Waldo*, during an unexpected home birth	Benjamin's role as a caregiver extends beyond his family when he helps deliver *Waldo Shenkman, his neighbor's son*, in a dramatic home birth· · ·
Omission of Key Information	Bobby Western's relationships are complex, featuring camaraderie with colleagues like Oiler and Red, a controversial bond with his sister, and deep connections with *figures such as Heaven, Asher, Granellen*· · ·	Bobby's most significant relationships are with his sister Alicia, who suffers from schizophrenia and eventually dies by suicide, and *his father, a renowned physicist*· · ·
	Avery continues her work, focusing on *helping clients like Marissa and Matthew Bishop navigate their marital issues* · · · Avery encounters various challenges, including dealing with Skylar's unexpected visit · · ·	Matthew is orchestrating these events as part of a revenge plot against Marissa and her affair partner, Skip, whom Avery briefly dated· · · it's orchestrated by a pharmaceutical company, Acelia, seeking *retribution against Avery for whistleblowing* · · ·
Event Misinterpretation	In the wake of Mildred's death, Benjamin's life takes a turn towards solitude and reflection. He *begins to work on his autobiography* with the help of Ida Partenza, a young secretary· · ·	Returning to New York, Rask realizes his wife's death has little impact on his life. He *continues investing* but never replicates his earlier success, returning to the solitary, dispassionate life· · ·
	Millie's history with Enzo and her relationship with Brock add complexity as she aids Wendy in escaping Douglas's control, *accidentally killing Douglas* in the process· · ·	Millie ends up shooting a man *she believes to be Douglas* during a violent altercation, only to *discover later that the man was actually Russell Simonds*· · ·
Character Misinterpretation	Ava is introspective, self-aware, and *morally driven*, with a strong desire for acceptance. She's empathetic but guarded, resourceful in adversity, and adept at navigating complex social situations· · ·	Ava is adept at manipulating situations to her advantage, portraying herself as vulnerable to deceive others while secretly harboring a willingness to *commit fraud to achieve her goals*· · ·
	June Hayward is introspective, ambitious, and somewhat cynical. She navigates her literary career with *determination and vulnerability, showing resilience* in the face of criticism and a deep appreciation for her moments of success· · ·	June Hayward is characterized by her intense jealousy, ambition, and insecurity. She is *manipulative, willing to betray* close relationships and ethical boundaries to achieve literary success· · ·

Table 4: A case study on common errors generated by models in the character profiling task.

5 Experiment Results

The important details of our experimental settings are provided in Appendix A.

In the experiments, we wish to answer two research questions: *RQ1)* Can LLMs generate character profiles from fictions precisely? *RQ2)* Can LLMs recognize the character's motivation for a specific decision based on the character profile?

5.1 *Can LLMs generate character profiles from fictions precisely?*

Experiment result in Table 2 shows that: *1)* LLMs generally exhibit promising performance in generating character profiles from fictions. Among all models, GPT-4 consistently outperforms other models across various methods, exhibiting the advanced capability of LLMs to accurately summarize character profiles. *2)* Despite GPT-4, larger and more complex LLMs, such as Qwen1.5-72B-Chat, tend to achieve higher consistency scores. *3)* There are variations in model performance across different dimensions. For example, LLMs typically achieve higher consistency scores in capturing personality traits but are less effective at summarizing event-related information.

Summarization Method Comparison We compare the outcomes of the incremental and hierarchical methods across the full CROSS. For 47 books containing fewer than 120,000 tokens in CROSS, we include the summarizing-in-one-go method.

The results in Table 2 show that the summarizing-in-one-go method achieves the highest consistency scores in all dimensions, surpassing methods that process content in segments. We believe this success stems from processing the entire content of a book at once, which maintains the narrative's coherence and minimizes information loss. Additionally, since character details are unevenly distributed

throughout a fiction, summarizing the text in one step allows the model to focus more effectively on the essential elements of the narrative.

The incremental updating method, while slightly lagging in average consistency, performs better in events than hierarchical summarizing. This performance can be attributed to its iterative updating nature, which allows the model to refine and update its understanding as more information becomes available or as errors are corrected in subsequent passes. This finding aligns with those reported by Chang et al. (2023), which indicate that book summaries generated by the incremental method surpass those produced by the hierarchical method in terms of detail.

Error Analysis We conduct a case study to further investigate why LLMs fail to generate the correct character profile. We define five types of errors, *i.e.*, 1) *Character misidentification*, which occurs when characters are mistaken for one another, leading to confusion about their actions or roles. 2) *Relationship Misidentification*, an error where the type of relationship between characters is inaccurately represented. 3) *Omission of Key Information*, a common error where the significant relationships or events are overlooked while less important information is described in excessive detail. 4) *Events Misinterpretation*, events are incorrectly interpreted, or earlier interpretations are not adequately revised in light of subsequent revelations. 5) *Character Misinterpretation*, where the motives or traits of a character are incorrectly summarized, resulting in a cognitive bias in the understanding of a character's overall image.

As shown in Table 4, a key finding is that the model often becomes confused and generates illusions when faced with complex narrative structures. For example, in the book "Trust", the character Benjamin Rask is a figure in the novel "Bonds" which is part of "Trust". The prototype for Rask is another character, Andrew Bevel, from "Trust". Due to frequent shifts in narrative perspective, the model confuses Rask with Bevel, mistakenly attributing Bevel's traits to Rask. Another example occurs in "The Housemaid's Secret", where the model fails to understand the plot twist, which results in an incorrect final summary.

5.2 *Can LLMs recognize the character's motivation for a specific decision?*

Overall Performance As shown in Table 2, profiles generated by GPT-4 through incremental method enable the model to achieve the highest accuracy (57.75%), which is slightly lower than that of the reference profiles (63.07%) shown in Table 3, indicating the effectiveness of the generated profiles in enhancing character comprehension.

Moreover, a strong positive correlation is observed between the consistency scores and the MR accuracy of the profiles summarized by the model. This finding supports the validity of character profiling, suggesting that accurate character profiles help models better understand the motivations behind a character's behavior.

Among the three summarization methods, profiles from hierarchical merging exhibit relatively low accuracy on the MR task. It is also found that despite high scores in other dimensions, the consistency score for the events obtained through the hierarchical method is relatively low. This indirectly suggests that the quality of events has a more significant influence on the MR task.

Ablation Study on MR As Table 3 demonstrates, the results of the ablation experiments reveal that each of the four dimensions within the profile contributes to the downstream task. Among these, the dimension of the event is the most critical. Excluding this dimension alone leads to a notable decrease in accuracy (-9.21%). The rationale behind this is that events contain substantial plot-related information, which assists the model in grasping the background knowledge pertinent to the characters' decision-making processes. Additionally, events integrate elements from the other dimensions, offering a holistic depiction of character personas. However, omitting the other dimensions has a less pronounced impact. We also observe that reducing the amount of information in the profile correlates with greater variance in experimental outcomes, suggesting that the model becomes less stable as it processes less detailed profiles.

6 Conclusion

We introduce the first task for assessing the character profiling ability of large language models (LLMs), using a dataset of 126 character profiles from novels. Our evaluation, which includes the *Factual Consistency Examination* and *Motivation Recognition*, reveals that LLMs generally perform well. However, even the most advanced models occasionally generate hallucinations and errors, particularly with complex narratives, highlighting the need for further improvement.

Limitations

In this paper, we only explore four common dimensions for character profiles, thus leaving other potential dimensions unexplored. This limitation suggests that future work could expand the scope to include a wider range of dimensions and investigate their effects on downstream tasks.

Another limitation of our work stems from potential biases in the evaluation process. Despite selecting highly contemporaneous data to prevent data leakage, it is still possible that some models might have been trained on these specific books. Besides, the evaluation metrics used in this paper rely on the evaluator LLMs, potentially compromising the accuracy of the results due to errors inherent in these models, which could result in a biased estimation of profile consistency. Moreover, while we test the three most popular summarization methods, we acknowledge that there is potential for improvement in the design of these methods to maximize the character profiling capabilities of LLMs.

Ethics Statement

We acknowledge that all authors are informed about and adhere to the ACL Code of Ethics and the Code of Conduct.

Use of Human Annotations Our institution recruit 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 the use of motivation recognition data processed by them for research purposes. Appendix E provides further details on the annotations.

Risks The CROSS dataset in our experiment are sourced from publicly available sources. However, we cannot guarantee that they are devoid of socially harmful or toxic language. Furthermore, evaluating the data quality of motivation recognition dataset is based on common sense, which can vary among individuals from diverse backgrounds. We use ChatGPT (OpenAI, 2022) to correct grammatical errors in this paper.

References

AI@Meta. 2024. Llama 3 model card.

American Psychological Association. 2018. Personality. https://dictionary.apa.org/personality. Retrieved April 9, 2024.

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

John William Atkinson. 1964. An introduction to motivation.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baichuan. 2023. Baichuan-npc.

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*.

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

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.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

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

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023b. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Paul R Kleinginna Jr and Anne M Kleinginna. 1981. A categorized list of motivation definitions, with a suggestion for a consensual definition. *Motivation and emotion*, 5(3):263–291.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

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

Dawei Li, Hengyuan Zhang, Yanran Li, and Shiping Yang. 2023b. Multi-level contrastive learning for script-based character understanding. *arXiv preprint arXiv:2310.13231*.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

OpenAI. 2022. Chatgpt.

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.

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*.

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

Charles Spearman. 1961. The proof and measurement of association between two things.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and gpt-3: Automated extraction of character roles without training data. *arXiv preprint arXiv:2205.07557*.

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

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

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv: Arxiv-2305.16291*.

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

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

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny: Can large language models simulate persona-driven decisions in role-playing? *arXiv preprint arXiv:2404.12138*.

Paul Thomas Young. 1961. Motivation and emotion: A survey of the determinants of human and animal activity.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023. Personality understanding of fictional characters during book reading. *arXiv preprint arXiv:2305.10156*.

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*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. 2023. Narrativeplay: Interactive narrative understanding. *arXiv preprint arXiv:2310.01459*.

Runcong Zhao, Qinglin Zhu, Hainiu Xu, Jiazheng Li, Yuxiang Zhou, Yulan He, and Lin Gui. 2024. Large language models fall short: Understanding complex relationships in detective narratives. *arXiv preprint arXiv:2402.11051*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

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 Experimental Setting

Models for Summarization For the incremental and hierarchical method, we experiment with the following LLMs: Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Mixtral-8x7B-MoE (Jiang et al., 2024), Qwen1.5-7B-Chat, Qwen1.5-14B-Chat, Qwen1.5-72B-Chat (Bai et al., 2023), vicuna-7b-v1.5-16k, vicuna-13b-v1.5-16k (Zheng et al., 2024), GPT-3.5-Turbo-0125 and GPT-4-Turbo-0125. We set the chunk size to 3000 tokens for all methods. We require that the complete profile generated by the model contain no more than 1200 words. For the summarizing-in-one-go method, we experiment with the GPT-4-Turbo-0125 and Claude-3-Sonnet (Anthropic, 2024). For all these models, we all adopt the origin model and official instruction formats. The temperature of all these models are set to 0 in our experiments.

MR Task Setting We assess the quality of profiles summarized under different models and methods through the accuracy rate on MR tasks. We uniformly employ GPT-4-Turbo-0409 as the reasoning model for this specific task.

Dimension Ablation Study To further explore the impact of different dimensions of character information on the MR task, we conduct an analysis through ablation experiments as shown in Table 3, using character profiles summarized via the incremental method with GPT-4. Each experiment is repeated three times, and we report the average and standard deviation of the results.

B Profiling v.s. Other Character Centric Summaries

B.1 Dimension

Some previous work also summarizes certain information about characters from books or scripts. However, these studies concerning character understanding focus on one specific aspect of character, such as role (Stammbach et al., 2022), relationship (Zhao et al., 2024), personality (Yu et al., 2023), mental states (Yu et al., 2022). Furthermore, although these studies offer valuable insights into character understanding, this focused approach may not capture the multifaceted nature of character.Although recent RPA works have managed to summarize character information in a multi-dimensional approach, including attributes, appearance, relationship, storyline etc (Zhao et al., 2023; Li et al., 2023a; Wang et al., 2023c), there is a lack of systematic assessment of the quality of these summaries.

B.2 Evaluation

Our evaluation framework of character profiling covers a wide range of information related to characters, helping to understand characters from various dimensions. Although we explicitly requested models to consider four dimensions, many dimensions of information are included in our framework or can be easily derived from the summarized profile. Specifically, in contrast to key mental states (Yu et al., 2022), our framework inherently encompasses critical mental information, For example, a character's objectives, part of their attributes profile defined in Section 3.2, reflect their desires and intentions. Key emotions and beliefs are often revealed through their reactions and behaviors in the events profile. Our work also includes factual details, like relationships with other characters and interactions with the external world.

B.3 Application

We believe that the extensive character summary can provide necessary and valuable information for various downstream applications, *e.g.* chatbots of fictional characters (Chen et al., 2023), interactive narratives (Zhao et al., 2023), and study guides for human readers. However, since we prompt the model to limit the total word count of the entire profile, some dimensions may be more concise and not as detailed as summaries that focus solely on that dimension.

C CROSS Dataset

C.1 Dataset Construction

We select 126 books to construct our dataset. For each book, we collect the book's epub format and transform it into TXT format, and then process the texts into chunks of content with the required chunk size. All 126 books are fictional novels with an average token count of 134412. Among these books, 47 books are less than 120k tokens in length, and the average token count of these books is 101885.

In order to minimize the potential for data leakage, we exclusively restrict our book selection to those published within the years 2022 and 2023. Additionally, we ensure that the selected books are either not sequels or, if they are sequels, can be

regarded as independent works. Our selection focuses on works of fiction, specifically excluding biographical novels and other works based on real historical figures.

For the evaluation of our work, we obtain permission from the developer of the SuperSummary website to use the summaries and character analyses of these books written by experts. All book summaries and character analyses are intended for academic research, and to protect the copyright of the website developers, we will not release the original summaries.

C.2 Integrity Verification of CROSS dataset

To confirm that our datasets and task genuinely evaluate understanding capability rather than simply testing the recovery of LLM training data, we design experiments to show that data leakage is not a significant concern with our dataset. In the following four datasets, all profiles are generated by GPT-4 through the incremental updating method.

For Publication Years We count the number of books in CROSS dataset published in the years 2022 and 2023, and their average consistency scores are shown in Table 5. The average consistency scores of the books from 2022 and 2023 are very close (3.60 vs 3.62).

For Different Sales Volumes We collect the collections tag in SuperSummary for books in CROSS dataset. The results of books in the "New York Times Best Sellers" collection and those are not are shown in Table 6. These two consistency scores are also very close (3.58 v.s. 3.60). The two experiments above demonstrate that within CROSS dataset, publication year and sales volume do not significantly affect task performance. Furthermore, we collect a small set of highly canonized texts for comparison with our dataset.

For Classic Works in the 20th Century We gather the top 10 books (excluding series) from the "Best Books of the 20th Century" list on the *goodreads*[8] website. This collection includes well-known classics like "The Little Prince", "1984", and "One Hundred Years of Solitude". The results of this set are shown in Table 7, where the consistency score is much higher than that of CROSS dataset. This finding suggests that our selection of data effectively reduces the impact of data leakage compared with choosing classic works. We

[8]https://www.goodreads.com/

Year	Count	Consistency Score				
		Attr	Rela	Even	Pers	Avg.
2022	93	3.70	3.23	3.71	3.75	3.60
2023	33	3.79	3.27	3.21	4.21	3.62

Table 5: Results of character profiling on books published in different year.

Bestseller	Count	Consistency Score				
		Attr	Rela	Even	Pers	Avg.
Yes	43	3.67	3.28	3.44	3.91	3.58
No	83	3.75	3.22	3.65	3.86	3.62

Table 6: Results of character profiling on books in "New York Times Best Sellers" collection in SuperSummary and those are not.

believe that high performance is due to the accumulation of time. The training set contains a large number of related corpora, such as Wikipedia entries, literary analyses, fan creations, etc., which deepen the model's understanding of these books and characters.

For Books Over Last Ten Years In order to test the impact of publication year on task performance in more recent books, we collect books from the "Best Books in {#the year}" list on the *goodreads* website. We gather five books each year for the last ten years. The results of the average consistency score over different years are shown in Figure 3, and the detailed score on different dimensions is shown in Table 8. We conduct a Spearman Rank Correlation Coefficient Test(Spearman, 1961) on this set. The coefficient of -0.037 with p-value of 0.799 (>0.05) indicates no significant correlation between the year of publication and the average consistency score over these 50 samples. This result suggests that even though the texts of books from a few years ago may well have been trained by the model, there are not enough related corpora to allow the model to perform well on this task solely based on memory.

Based on the above analysis, we reasonably speculate that, at least for the next few years, our dataset will remain effective for updated LLMs. Moreover, this work does not only focus on the dataset itself but, importantly, on a feasible framework designed to continually update and expand this dataset. Furthermore, we will keep updating the dataset and evaluating the performance of new LLMs in our future works.

Count	Consistency Score				
	Attr	Rela	Even	Pers	Avg.
Best Books of the 20th century					
10	4.7	4.1	4.2	4.8	4.45

Table 7: Results of character profiling on 10 books in "Best Books of the 20th Century" list in *goodreads* .

Year	Count	Consistency Score				
		Attr	Rela	Even	Pers	Avg.
2023	5	3.4	3.8	3.2	4.2	3.65
2022	5	4.2	3.2	3.6	4.4	3.85
2021	5	4.0	3.4	3.6	4.2	3.80
2020	5	4.4	4.0	3.8	4.4	4.15
2019	5	4.0	3.8	4.0	4.6	4.10
2018	5	3.8	3.6	4.0	3.8	3.80
2017	5	3.4	4.4	3.2	4.2	3.80
2016	5	4.4	3.6	4.0	4.0	4.00
2015	5	4.0	3.6	3.4	4.0	3.75
2014	5	4.0	2.8	3.8	4.0	3.65

Table 8: Results of character profiling on books in "Best Books in {#the year}" list in *goodreads*.

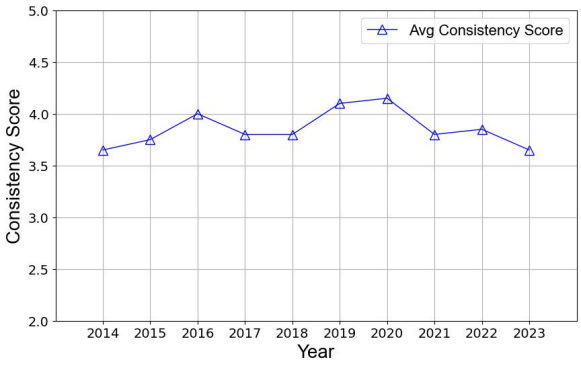


Figure 3: Average Consistency Score of books in "Best Books in {#the year}" list in *goodreads* in different years.

D Detailed Information of Summarization Method

Given a book B with length L, for chunk-based method, we split B into independent chunk $c_1, c_2, \cdots, c_{\lceil L/C \rceil}$ with chunk size $C = 3000$. We fix the context window $W = 8096$ and the maximum summary length $M = 1200$.

D.1 Incremental Updating

The progress of incremental updating is listed as follows:

- Step 1: Given the first chunk c_1, the model outputs the initial summary s_1.
- Step 2: Given the chunk content c_2, and the summary s_1, the model outputs summary s_2 which contains content of the first two chunk.
- The summary is iteratively updated within the next chunk through step 2 until the final summary $s_{\lceil L/C \rceil}$ is obtained.
- If the summary exceeds M in these steps, the model is required to compress the summary into the required length.

D.2 Hierarchical Merging

The progress of hierarchical merging is listed as follows:

- Step 1: Given the chunks $c_1, c_2, \cdots, c_{\lceil L/C \rceil}$, the model outputs the level 1 summaries for each chunk.
- Step 2: Merge as many consecutive level 1 summaries as possible with the limit that the total length of the summaries and the prompt is less than W. Given these summaries, the model outputs the first level 2 summary, which serves as the context for next merging.
- Step 3: Merge as many remaining level 1 summaries as possible with the limit that the total length of these summaries and the prompt and the context is less than W. Given this content, the model outputs the next level 2 summary, which also serves as the context for next merging. This process is iteratively conducted within the remaining summaries.
- Merge the level 2 summaries by repeating steps 2 and 3 until a final summary is obtained.

D.3 Summarizing in One Go

We first ensure the total length of the selected book and the summarizing prompt is less than the context window limit of GPT-4 and Claude-3-Sonnet. Given the whole content of the book, the model outputs the final summary at once.

E Manual Annotation

We invite two native English-speaking college students as human annotators for manual evaluation in our work. These annotators receive compensation exceeding the local minimum wage. They also have consented to the use of motivation recognition data filtered by them for research purposes.

E.1 Reference Profile Examination

To examine the correctness of the character profile parsed by GPT-4 from the original book summary and character analysis, we employ two annotators

to check the consistency between the reorganized profile and the original content. The annotators are given the origin plot summary, character analysis, and reorganized character profile. Then they are required to determine whether the reorganized profile is consistent with the original information. The two annotators' result shows that the profiles of all samples in CROSS dataset do not contain plot inconsistencies and misjudgments of the character's traits. This result indicates that the quality of the profile can be used as a golden profile.

E.2 Manual Evaluation

In order to examine the quality of Llama-3-70B evaluator result, we sample 50 pieces in our dataset and invite two annotators to evaluate the generated profile in consistency score. We provide the annotators and Llama-3-70B with the same scoring prompt. For the metric consistency score, the Pearson Correlation Coefficient between the average human result and Llama-3-70B scoring is 0.752 with p-value $= 4.3\text{e}{-12}$. The p-value < 0.05 demonstrates that these two sets of results have a significant correlation. The coefficient result indicates that the Llama-3-70B evaluation ability is comparable with human annotators on the assessing character profile.

E.3 Motivation Recognition MCQs filtering

To ensure the quality of the MR question dataset, we employ two annotators for conducting manual filtering. The annotators are provided with reference character profiles, generated questions, and the following criteria:

- **The decision must be made by the selected character.** Each question must feature a decision and the scenario, with the focus character as the decision-maker.
- **Questions should ask directly or indirectly about the character's motivation for making the decision.** Each question must directly or indirectly inquire about the character's motivation for making their decision, avoiding irrelevant information.
- **The decision must be meaningful within the story context.** The decision in the question must contribute meaningfully to the storyline. It should reflect a conscious choice by the character that holds importance in the narrative, rather than representing a mundane or routine decision.

Init Feedback (Incremental)
If there is no information about character {} in the beginning part of a story, just output 'None' in each section. Do not apologize. Just output in the required format.
Init Feedback (Hierachical)
If there is no information about character {} in this part of the story, just output 'None' in each section. Do not apologize. Just output in the required format.
Update Feedback (Incremental)
If there is no information about character {} in this excerpt, just output the origin summary of the character {} of the story up until this point. Do not apologize. Just output in the required format.

Table 9: The additional prompt for the GPT-4 model.

- **Leaking questions is prohibited.** Scenarios and questions must not include the motivation behind the characters' decisions.

We require the annotators to determine if the question meets the criteria. By filtering the dataset, we finally get 445 high-quality motivation recognition multiple-choice questions with Fleiss's $\kappa = 0.91$. We also adjust the arrangement of the options to ensure a fair distribution of correct answers.

F Traditional Metrics on Generated Profiles

In our evaluation protocol, traditional metrics for text summarization like ROUGE(Lin, 2004), BLEU(Papineni et al., 2002), and BERTScore(Zhang et al., 2019) are not used because they have been shown to be unreliable for measuring summary quality of GPT-3 generated summaries compared to human evaluations(Goyal et al., 2022). However, to provide a comprehensive perspective, we present the results of these three traditional metrics in Table 10 and Table 11 for reference.

G Prompts

For summarization, we mainly adopt the prompt structure from Chang et al. (2023).

G.1 Summarizing in One Go

In our experiment, we have found that the long-context capabilities of Claude-3-Sonnet are limited. Consequently, the model occasionally forgets the instructions and generates a simplistic summary instead of organizing the output into four distinct sections when the task prompt precedes the novel's content. Therefore, we choose to put

Summarization Method	Summarization Model	ROUGE-L %					BLEU %				
		Attr	Rela	Even	Pers	Avg.	Attr	Rela	Even	Pers	Avg.
CROSS *(Full dataset)*											
Incremental Updating	Mistral-7B-Instruct-v0.2	25.35	25.87	20.98	19.20	22.85	7.16	5.11	2.74	2.15	4.29
	Mixtral-8x7B-MoE	24.97	25.96	20.14	19.78	22.71	7.29	5.36	3.05	3.79	4.87
	vicuna-7b-v1.5-16k	28.23	25.63	20.45	23.21	24.38	8.52	4.82	2.13	4.20	4.92
	vicuna-13b-v1.5-16k	29.14	26.72	21.66	21.05	24.64	9.03	5.31	2.91	3.19	5.11
	Qwen1.5-7B-Chat	23.47	24.27	20.90	24.41	23.26	5.82	4.37	2.59	4.87	4.41
	Qwen1.5-14B-Chat	25.40	23.47	19.16	24.56	23.15	7.78	4.82	1.51	4.36	4.62
	Qwen1.5-72B-Chat	29.23	27.99	21.34	26.97	26.38	9.58	7.06	2.60	6.64	6.47
	GPT-3.5-Turbo	29.40	26.43	21.78	23.46	25.27	9.25	5.53	3.59	4.98	5.84
	GPT-4-Turbo	30.90	28.24	22.95	26.57	27.17	9.65	6.98	3.74	5.66	6.51
Hierarchical Merging	Mistral-7B-Instruct-v0.2	26.62	25.55	21.70	21.96	23.96	7.14	5.79	3.42	3.29	4.91
	Mixtral-8x7B-MoE	26.66	26.77	23.44	20.78	24.41	7.44	5.75	3.91	3.29	5.10
	vicuna-7b-v1.5-16k	26.56	24.74	20.02	21.48	23.20	6.22	3.82	1.83	3.43	3.83
	vicuna-13b-v1.5-16k	27.00	26.19	20.96	21.34	23.87	7.74	4.52	2.34	3.27	4.47
	Qwen1.5-7B-Chat	26.10	27.15	20.91	26.31	25.12	6.76	5.86	2.27	5.93	5.21
	Qwen1.5-14B-Chat	26.17	25.81	20.90	25.30	24.55	8.11	5.99	1.68	5.19	5.24
	Qwen1.5-72B-Chat	30.50	29.60	25.27	27.57	28.24	10.58	8.05	4.76	7.75	7.79
	GPT-3.5-Turbo	31.63	27.92	22.31	28.52	27.60	10.91	6.97	3.52	8.75	7.54
	GPT-4-Turbo	**32.03**	**29.92**	**26.16**	**30.20**	**29.58**	**12.18**	**10.33**	**6.10**	**9.26**	**9.47**
CROSS *(Short subset)*											
Sum-in-One-Go	GPT-4-Turbo	**35.07**	**36.34**	**29.04**	**33.54**	**33.50**	**14.45**	**14.39**	5.83	**11.81**	**11.62**
	Claude3-Sonnet	30.50	29.60	25.27	27.57	28.24	10.58	8.05	4.76	7.75	7.79
Incremental	GPT-4-Turbo	31.91	29.52	22.59	26.68	27.68	10.79	7.73	3.34	5.42	6.82
Hierarchical	GPT-4-Turbo	32.69	30.80	26.09	30.25	29.96	12.52	11.25	**5.88**	9.52	9.79

Table 10: Metric ROUGE-L and BLEU of different LLMs performance on character profiling. The best scores are **bolded** and the second best scores are underlined.

the task prompt after the content of the novel. The prompts for summarizing-in-one-go method can be found in Table 12.

G.2 Incremental Updating

The prompts for incremental updating can be found in Table 13.

We have found that the GPT-4 model will provide an apology if there is no information available about the designated character in the current excerpt, instead of outputting in the required format. So we add an additional prompt for the GPT-4 model and regenerate, if the response starts with apology. The feedback prompt can be found in Table 9.

G.3 Hierarchical Summarizing

Likewise, we add a feedback prompt for the GPT-4 model if the response starts with an apology. The prompts for hierarchical summarizing can be found in Table 14.

G.4 Factual Consistency Examination

For evaluation, we mainly adopt the prompt structure from Liu et al. (2023). The prompt template is shown in Table 15.

G.5 Motivation Recognition

The prompt template of MR task is shown in Table 16.

Summarization Method	Summarization Model	BertScore %				
		Attr	Rela	Even	Pers	Avg.
CROSS *(Full dataset)*						
Incremental Updating	Mistral-7B-Instruct-v0.2	87.43	88.01	84.84	86.29	87.06
	Mixtral-8x7B-MoE	87.34	85.43	75.98	77.67	81.61
	vicuna-7b-v1.5-16k	87.57	87.72	84.71	85.77	86.44
	vicuna-13b-v1.5-16k	87.62	87.61	84.72	83.20	85.79
	Qwen1.5-7B-Chat	87.19	87.77	85.18	88.08	87.06
	Qwen1.5-14B-Chat	87.78	88.03	85.21	88.83	87.46
	Qwen1.5-72B-Chat	88.63	88.93	85.36	89.19	88.03
	GPT-3.5-Turbo	88.32	88.24	85.35	88.23	87.54
	GPT-4-Turbo	88.71	88.7	<u>85.86</u>	88.89	88.04
Hierarchical Merging	Mistral-7B-Instruct-v0.2	87.79	88.15	83.92	87.13	86.75
	Mixtral-8x7B-MoE	87.54	87.51	84.81	86.12	86.50
	vicuna-7b-v1.5-16k	86.14	86.42	84.37	85.31	85.56
	vicuna-13b-v1.5-16k	87.22	87.42	84.91	86.15	86.43
	Qwen1.5-7B-Chat	87.66	88.27	85.37	88.37	87.42
	Qwen1.5-14B-Chat	88.47	88.62	85.95	88.76	87.95
	Qwen1.5-72B-Chat	88.97	89.24	86.63	89.23	<u>88.52</u>
	GPT-3.5-Turbo	<u>89.13</u>	<u>89.31</u>	85.81	<u>89.57</u>	88.46
	GPT-4-Turbo	**89.31**	**89.60**	**86.82**	**89.86**	**88.90**
CROSS *(Short subset)*						
Sum-in-One-Go	GPT-4-Turbo	**90.05**	**90.69**	**87.68**	**90.56**	**89.75**
	Claude3-Sonnet	88.97	89.24	86.63	89.23	88.52
Incremental	GPT-4-Turbo	89.00	88.81	85.80	88.72	88.08
Hierarchical	GPT-4-Turbo	<u>89.33</u>	<u>89.89</u>	<u>86.86</u>	<u>89.80</u>	<u>88.97</u>

Table 11: Metric BERTScore of different LLMs performance on character profiling . The best scores are **bolded** and the second best scores are <u>underlined</u>.

/ Data */*
Below is the content of the novel:

- - -
{}
- - -
/ Task prompt */*
You are a character persona extraction assistant. Your task is to write a summary for the character {} in this novel. You must briefly introduce characters, places, and other major elements if they are being mentioned for the first time in the summary. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative. The summary must be within {} words and could include multiple paragraphs.
/ Output Format */*
Output your summary in four specific sections, using the following titles as paragraph headers:

Attributes: // Briefly identify the character's gender, skill, talents, objectives, and background within {} words.
Relationships: // Briefly describe the character's relationships with other characters within {} words.
Events: // Organize the main events the character experiences or is involved in chronological order within {} words.
Personality: // Briefly identify the character's personality within {} words.

Ensure that each section explicitly starts with the specified title, followed by the content and that there is a clear separation (a newline) between each section.

Summary:
Attributes:
Margot Davies is a determined and skilled female reporter with...

Relationships:
Margot has a close and loving relationship with her uncle...

Events:
Margot returns to her hometown of Wakarusa to care for her ailing uncle...

Personality:
Margot is tenacious, intelligent, and compassionate....

Table 12: Prompt templates for summarizing-in-one-go method. Generated texts by a LLM are *highlighted*.

I: Init

/ Data */*
Below is the beginning part of a story:

- - -
{}
- - -
/ Task prompt */*
We are going over segments of a story sequentially to gradually update one comprehensive summary of the character {}. Write a summary for the excerpt provided above, make sure to include vital information related to gender, skills, talents, objectives, background, relationships, key events, and personality of this character. You must briefly introduce characters, places, and other major elements if they are being mentioned for the first time in the summary. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative. Despite this step-by-step process of updating the summary, you need to create a summary that seems as though it is written in one go. The summary must be within {} words and could include multiple paragraphs.
/ Output Format */*
Output your summary into four specific sections, ...

Summary:

II: Update

/ Data */*
Below is a segment from a story:

- - -
{}
- - -

Below is a summary of the character {} of the story up until this point:

- - -
{}
- - -
/ Task prompt */*
We are going over segments of a story sequentially to gradually update one comprehensive summary of the character {}. You are required to update the summary to incorporate any new vital information in the current excerpt. This information may relate to gender, skills, talents, objectives, background, relationships, key events, and personality of this character. You must briefly introduce characters, places, and other major elements if they are being mentioned for the first time in the summary. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative. Despite this step-by-step process of updating the summary, you need to create a summary that seems as though it is written in one go. The updated summary must be within {} words and could include multiple paragraphs.
/ Output Format */*
Output your summary into four specific sections, ...

Updated summary:

III: Compress

/ Data */*
Below is a segment from a story:

- - -
{}
- - -
/ Task prompt */*
Currently, this summary contains {} words. Your task is to condense it to less than {} words. The condensed summary should remain clear, overarching, and fluid while being brief. Whenever feasible, maintain details about gender, skills, talents, objectives, background, relationships, key events, and personality about this character - but express these elements more succinctly. Make sure to provide a brief introduction to characters, places, and other major components during their first mention in the condensed summary. Remove insignificant details that do not add much to the character portrayal. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative.
/ Output Format */*
Output your summary into four specific sections, ...

Condensed summary (to be within {} words):

Table 13: Prompt templates for incremental updating.

I: Init

/ Data */*
Below is a part of a story:

- - -
{}
- - -
/ Task prompt */*
We are creating one comprehensive summary for the character {} by recursively merging summaries of its chunks. Now, write a summary for the excerpt provided above, make sure to include vital information related to gender, skills, talents, objectives, background, relationships, key events, and personality of this character. You must briefly introduce characters, places, and other major elements if they are being mentioned for the first time in the summary. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative. Despite this recursive merging process, you need to create a summary that seems as though it is written in one go. The summary must be within {} words and could include multiple paragraphs.
/ Output Format */*
Output your summary into four specific sections, ...

Summary:

II: Merge

/ Data */*
Below are several summaries of the character {} from consecutive parts of a story:

- - -
{}
- - -
/ Task prompt */*
We are creating one comprehensive summary for the character {} by recursively merging summaries of its chunks. Now, merge the given summaries into one single summary, make sure to include vital information related to gender, skills, talents, objectives, background, relationships, key events, and personality of this character. You must briefly introduce characters, places, and other major elements if they are being mentioned for the first time in the summary. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative. Despite this recursive merging process, you need to create a summary that seems as though it is written in one go. The summary must be within {} words and could include multiple paragraphs.
/ Output Format */*
Output your summary into four specific sections, ...

Summary:

III: Merge Context

/ Data */*
Below is a summary of the context about the character {} preceding some parts of a story:

- - -
{}
- - -

Below are several summaries of the character {} from consecutive parts of the story:

- - -
{}
- - -
/ Task prompt */*
We are creating one comprehensive summary of the character {} by recursively merging summaries of its chunks. Now, merge the preceding context and the summaries into one single summary, make sure to include vital information related to gender, skills, talents, objectives, background, relationships, key events, and personality of this character. You must briefly introduce characters, places, and other major elements if they are being mentioned for the first time in the summary. The story may feature non-linear narratives, flashbacks, switches between alternate worlds or viewpoints, etc. Therefore, you should organize the summary so it presents a consistent and chronological narrative. Despite this recursive merging process, you need to create a summary that seems as though it is written in one go. The summary must be within {} words and could include multiple paragraphs.
/ Output Format */*
Output your summary into four specific sections, ...

Summary:

Table 14: Prompt templates for hierarchical merging.

I: Consistency Score

/ Task prompt */*
You are a character extraction performance comparison assistant. You will be given the golden information about character {}'s {dimension} in a novel. You will then be given the summarized information about character {} extracted by a model from the origin novel.
Your task is to rate the summarized information on one metric.
Please make sure you read and understand these instructions carefully.

Evaluation Criteria:
Consistency (1-5) - the factual alignment between the golden and the summarized information. A score of 1 indicates significant discrepancies, while a score of 5 signifies a high level of factual consistency.

Evaluation Steps:
1. Read the golden information carefully and identify the main facts and details it presents.
2. Read the summarized information and compare it to the golden information. Check if the summary contains any factual errors or lacks necessary foundational facts. If the summarized one includes information not mentioned in the golden information, please ignore it, as the summary is extracted from the original book and may contain more extraneous information.
3. Assign a score for consistency based on the Evaluation Criteria and explain the reason. Your output should be structured as the following schema: {{"score": int // A score range from 1 to 5, "reason": string // The reason of evaluation result}}
/ Data */*
Golden information:
{}
Summarized information:
{}
/ Output Format */*
Evaluation Form (Please output the result in JSON format. Do not output anything except for the evaluation result. All output must be in JSON format and follow the schema specified above.):
- Consistency:
{
"score": 3,
"reason": "The summarized information is partially consistent with the golden information, ..."
}

II: Win-win Rate

/ Task prompt */*
You are a character extraction performance comparison assistant. You will be given the golden information about character {}'s {dimension} in a novel. You will then be given the summarized information about character {} extracted by two different models from the origin novel.
Your task is to rank the models based on which summarization has a higher consistency with the golden information.
Please make sure you read and understand these instructions carefully.

Ranking Steps:
1. Read the golden information carefully and identify the main facts and details it presents.
2. Read the outputs of the models and compare them to the golden information. Check if the summary contains any factual errors or lacks necessary foundational facts.
3. Choose a model whose output has a higher factual alignment with the golden information and explain the reason. Your output should be structured as the following schema: {{"model_name": str // The model name with higher rank, if these models have the same level of performance, output "Equilibrium", "reason": string // The reason of ranking result}}
/ Data */*
Golden information:
{}

Outputs of the models:
"model_name": "model_1",
"summarization":{}
"model_name": "model_2",
"summarization": {}
/ Output Format */*
Ranking Form (Please output the result in JSON format. Do not output anything except for the evaluation result. All output must be in JSON format and follow the schema specified above.):
- Consistency:
{
"model_name": "model_1",
"reason": "Model 1's summarization is more consistent ..."
}

Table 15: Prompt templates for factual consistency examination. Generated texts by GPT-4 are *highlighted*.

I: Normal

/ Task prompt */*
You are a helpful assistant proficient in analyzing the motivation for the character's decision in novels. You will be given the profile about character {} in a novel. Your task is to choose the most accurate primary motivation for the character's decision according to the character's profile. You also need to provide reasons, the reasons should be related to the character's basic attributes, experiences, relationships, or personality, of this character.
Your output should be structured as the following schema:
{{"Choice": str // "A"/"B"/"C"/"D", "Reason": string // The reason of the choice}}
/ Data */*
Character Profile:
name: {}
Summary of this character: {}

Question:
{}
/ Output Format */*
Output (All output must be in JSON format and follow the schema specified above.):
{
"Choice": "A",
"Reason": "Margot's primary motivation for ..."
}

II: Ablate All Dimensions

/ Task prompt */*
You are a helpful assistant proficient in analyzing the motivation for the character's decision in novels. Your task is to choose the most accurate primary motivation for the character's decision according to the character's profile. Since you are not given the character analysis, you are supposed to choose the most reasonable motivation based on the provided information in the question.
Your output should be structured as the following schema:
{{"Choice": str // "A"/"B"/"C"/"D", "Reason": string // The reason of the choice}}
/ Data */*
Character Profile:
name: {}

Question:
{}
/ Output Format */*
Output (All output must be in JSON format and follow the schema specified above.):
{
"Choice": "A",
"Reason": "Given the lack of specific information about Margot, ..."
}

Table 16: Prompt templates for motivation recognition. Generated texts by GPT-4 are *highlighted*.