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

Xinfeng Yuan[♡], **Siyu Yuan**[♡*], **Yuhan Cui**[♠*], **Tianhe Lin**[♡], **Xintao Wang**[♠], **Rui Xu**[♠], **Jiangjie Chen**[♠], **Deqing Yang**[♡†]
[♡]School of Data Science, Fudan University
[♠]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
{xfyuan23,syyuan21,xtwang21,ruixu21}@m.fudan.edu.cn
{yhcui20,thlin20,jjchen19,yangdeqing}@fudan.edu.cn

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 them with ground truth references and evaluating 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 are available at https://github.com/Joanna0123/character_profiling.

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

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8015–8036
November 12-16, 2024

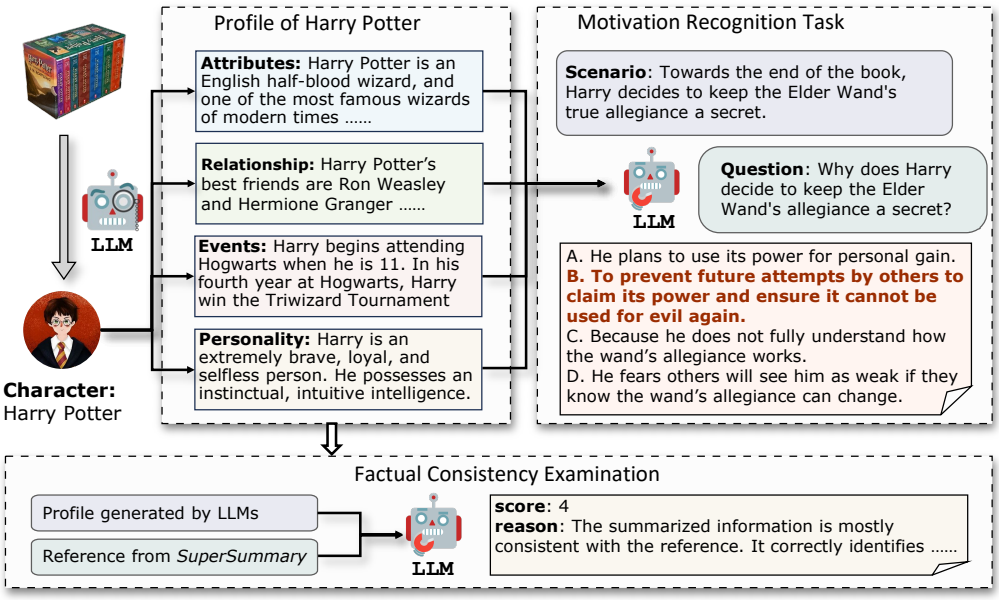


Figure 1: An overview of character profiling with LLMs and the two evaluation tasks we proposed, including factual consistency examination and motivation recognition.

CROSS dataset is sourced from SuperSummary [1], a 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, showcasing 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 employs 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

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

[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). 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, the 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 A 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 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.

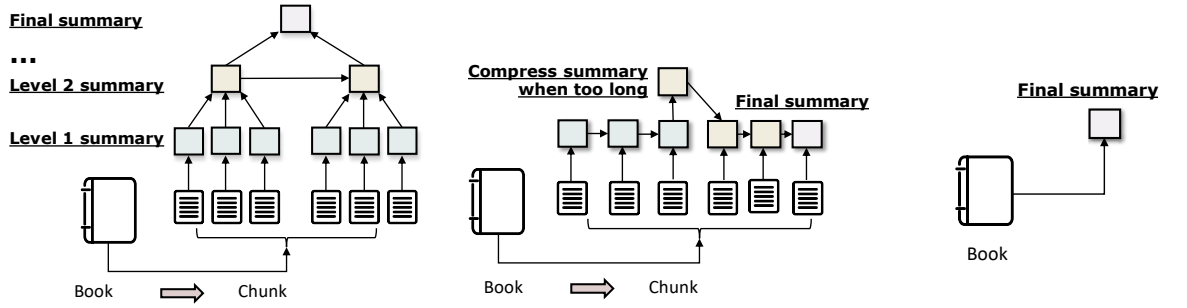
Personality Personality refers to the lasting set of characteristics and behaviors that form an individual's unique way of adapting to life (American Psychological Association, 2018). While well-rounded characters often exhibit complex personalities, their personalities can be analyzed through their 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 $\mathcal{A} = \{a_i\}_{i=1}^{4}$ for $\mathcal{Q}$, the LLMs should determine the answer $\mathcal{Y}$ from $\mathcal{A}$ 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 to improve alignment with human perception and reduce 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 sets 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] In fact, as shown in Appendix B.2, we find that there is no significant correlation between the year of publication and the consistency score for works from the past ten years. 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 analyses from `SuperSummary` lack 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 reorganize character profiles across four main 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

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

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

[6] The detail of human examination is shown in Appendix D.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.

Reasoned by	Ablation Dimension	Acc. %	Std. %
Generated Profile (GPT-4-Turbo + incremental updating)			
GPT-4-Turbo	-	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
Reference Profile in CROSS			
GPT-4-Turbo	-	**63.07**	0.11
human	-	72.58	3.32

Table 3: Results of Motivation Recognition ablation study. The reasoning model by default is GPT-4-Turbo-0409. **Ablation Dimension** refers to omitted dimensions in experiments.

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) and manually select the 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. Each question is designed to include a decision made by the character within a specific scenario, offering four options, the correct option, and justifications for why each option is correct or incorrect. 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

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.

the MR questions dataset.[7]

5 Experiment Settings

Models for Summarization For the incremental and hierarchical methods, 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 both 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 original model and official instruction formats. The temperatures 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. Furthermore, we study human performance in the MR task supported by reference profile in CROSS dataset. We employ two human annotators to answer all the questions and calculate the average accuracy and standard deviation.

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

[7]Further details are shown in Appendix D.

standard deviation of the results.

6 Experiment Results

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

6.1 *Can LLMs generate character profiles from fictional works precisely?*

Experiment result in Table 2 shows that: *1)* LLMs generally exhibit promising performance in generating character profiles from fiction. 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 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 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) *Event 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. The errors are shown in the first examples of *Character misidentification* and *Event Misinterpretation*. Another example occurs in "The Housemaid's Secret", where the model fails to understand the plot twist, which results in an incorrect final summary. This error is shown in the second case of *Event Misinterpretation*.

6.2 *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. Additionally, based on the human annotators' results (72.58%), GPT-4 still shows a performance gap compared to humans in this task.

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.

7 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

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

Risks The CROSS dataset in our experiment is 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 the 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.

Acknowledgements

We thank the anonymous reviewers for their valuable suggestions. This work was supported by the Chinese NSF Major Research Plan (No.92270121).

References

AI@Meta. 2024. Llama 3 model card.