Guess What I am Thinking: A Benchmark for Inner Thought Reasoning of Role-Playing Language Agents

Rui Xu[♡♠*†], **MingYu Wang**[♡], **XinTao Wang**[♡]
Dakuan Lu[♠], **Xiaoyu Tan**[♠†], **Wei Chu**[♠], **Yinghui Xu**[♡♠†]
[♡]Fudan University [♠]INF Technology (Shanghai) Co., Ltd.
rxu24@m.fudan.edu.cn, txywilliam1993@outlook.com, xuyinghui@fudan.edu.cn,

Abstract

Recent advances in Large Language Model (LLM)-based Role-Playing Language Agents (RPLAs) have attracted broad attention in various applications. While chain-of-thought reasoning has shown importance in many tasks for LLMs, the internal thinking processes of RPLAs remain unexplored. Understanding characters' inner thoughts is crucial for developing advanced RPLAs. In this paper, we introduce ROLETHINK, a novel benchmark constructed from literature for evaluating character thought generation. We propose the task of inner thought reasoning, constructing 6,058 data entries from 76 books, which includes two sets: the gold set that compares generated thoughts with original character monologues, and the silver set that uses expert-synthesized character analyses as references. To address this challenge, we propose MIRROR, a chain-of-thought approach that generates character thoughts by retrieving memories, predicting character reactions, and synthesizing motivations. Through extensive experiments, we demonstrate the importance of inner thought reasoning for RPLAs, and MIRROR consistently outperforms existing methods.

1 Introduction

Recent advances in large language models (LLMs) have enabled the development of role playing language agents (RPLAs) (Shao et al., 2023a; Chen et al., 2024), which are now widely used in applications from character chatbots (Xu et al., 2024a) to game NPCs (Wang et al., 2023; Ran et al., 2024). While chain-of-thought reasoning (Wei et al., 2023; Yang et al., 2025a) has shown significant success in various LLM tasks (Chu et al., 2024; Xiang et al., 2025), the inner thought process of RPLAs remains unexplored. To address this gap, research on generating high-quality character thought data both helps us understand character motivations and

Figure 1: The process of characters performing inner thought reasoning. When facing specific scenarios, characters generate different inner thoughts, which reflect a deep understanding of the character while influencing their behavior.

[*] Work is done during internship at INF.
[†] Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15148–15168
November 4-9, 2025

shows strong potential for improving role-playing capabilities.

Prior research has explored LLMs' ability to understand characters in fictional works, primarily focusing on basic tasks such as character prediction (Brahman et al., 2021; Yu et al., 2022) and personality prediction (Yu et al., 2023; Ran et al., 2025), lacking deep analysis of character behaviors. Recently, the research focus has shifted to role-playing, where LLMs have demonstrated strong performance in basic role-playing tasks, such as knowledge replication (Zhou et al., 2023a; Ran et al., 2024) and speaking style imitation (Li et al., 2023; Xiao et al., 2025). However, these models show limited capabilities in complex tasks involving decision-making (Xu et al., 2024b; Li et al., 2025a) and psychological reasoning (Wang et al., 2024; Li et al., 2025b). As shown in Figure 1, generating inner thought processes before actions enables both in-depth analysis of character motivations and better completion of subsequent behaviors. However, constructing high-quality character inner thought data remains a significant challenge.

In this paper, we systematically evaluate LLMs' ability to generate thought chains for role-playing models. We constructed the ROLETHINK benchmark using 76 high-quality novels as raw material, all of which contain extensive character monologues and rich analyses of character behavior by literary experts. We propose the task of inner thought reasoning. In this task, given a character profile, LLMs are required to generate plausible thoughts based on the current scenario, which are then evaluated against reference thoughts. Based on different reference types, we divide the task into two sets: Gold and Silver. The Gold Set uses character monologues from the original novels as references, while the Silver Set employs synthesized data from literary experts' character analyses. To ensure comprehensive evaluation, we employ both automatic metrics and human assessment methods.

To better analyze and address this task, we conduct extensive experiments with various LLMs. Based on our analysis, we propose **M**emory **I**ntegration and **R**ole **R**easoning with **O**bservation **R**eflection (MIRROR), a chain-of-thought approach that retrieves multiple relevant memories from the current scenario, predicts reactions of related characters and environments, and summarizes them into character motivations. Models enhanced with MIRROR consistently achieve higher scores on ROLETHINK. To further validate the benefits of reasoning processes in RPLAs, we evaluate on multiple role-playing benchmarks. The results demonstrate that enabling RPLAs to think before acting improves performance across various downstream tasks.

Our contributions are summarized as follows: *1)* We conduct the first study on the inner thought process behind RPLAs and construct a character inner monologue dataset containing 6,058 entries from 76 books. *2)* We propose MIRROR, a method that better generates character inner thought processes. *3)* We conduct extensive experiments with different LLMs, validating the importance of inner thought processes across various role-playing downstream tasks.

2 Related Work

2.1 Character Understanding

Understanding characters is crucial for natural language processing systems to comprehend narrative texts. Recent research has explored various aspects of character analysis, including personality traits (Yu et al., 2023; Dai and Xiao, 2025), relationship networks (Chen and Choi, 2016), and behavioral patterns (Brahman et al., 2021; **?**). Yuan et al. (2024) attempt to use LLMs to summarize character profiles from source materials. Our research extends this by exploring the generation and analysis of characters' inner thoughts, providing deeper insights into their decision-making processes. Previous work has developed benchmarks for character identification (Sang et al., 2022; Yuan et al., 2024) and question answering (Kočiskỳ et al., 2018), but these primarily evaluate surface-level text understanding. In contrast, we propose methods to model characters' internal reasoning processes, including memory recall and theory of mind thinking.

2.2 Role Playing Language Model

LLM advances enable sophisticated Role-Playing Language Agents (RPLAs) (Wang et al., 2025). Current RPLAs often target character-consistent dialogue via fine-tuning (Li et al., 2023; Yang et al., 2025b) or memory retrieval (Shao et al., 2023b). These methods improve surface interaction but overlook underlying psychological processes. Deeper evaluations, such as assessing character decisions (Xu et al., 2024b) or personality fidelity (Wang et al., 2024), reveal models often lack consistent inner reasoning, a finding echoed by other frameworks (Chen et al., 2024). Our work introduces explicit thought generation—modeling memory recall, perspective-taking, and decision reasoning—to create a more complete framework for understanding and reproducing character behaviors.

3 RoleThink Benchmark

3.1 Task Formulation

Character Inner Thought Reasoning aims to generate characters' thinking processes in specific scenarios. Given a character profile P and a scenario description S, the task requires LLMs to generate the thoughts T that led to their behavior.

We divide our benchmark into two sets based on different sources of reference. The Gold Set aims to recover a character's original thoughts from the novels. Given a character profile P and a scenario description S, where the character's original thoughts are masked, the model generates thoughts T that are evaluated against T_{gold}, which is directly extracted from the character's thoughts in the source material. The Silver Set focuses on generating plausible thoughts for given scenarios. With the same input format of profile P and scenario S, the model generates thoughts T that is evaluated against T_{silver} collected from literary experts and fan communities' analyses of the character's inner thoughts in that scenario.

3.2 Gold Set

The Gold Set requires LLMs to recreate characters' thoughts from the original book. To collect high-quality thought data, we input each POV chapter (∼10k tokens) into GPT-4o to identify key characters and detect their inner thoughts in these sections. The detected thought segments are then manually filtered to ensure quality. Table 1 shows two complete examples from the Gold Set. We mask the character thoughts in the chapters and use the text before each mask as scenario data. LLMs need to use character profiles to generate the masked thought content.

3.3 Silver Set

The Silver Set requires LLMs to generate character thoughts not present in the original book. To collect reference data, we gather character analysis from two professional book review website[1,2]. We use GPT-4o to process this data, summarizing character motivations and locating specific thought points in the chapters. The collected data is then manually filtered to ensure quality. Table 1 shows a complete example from the Silver Set, where the located chapter is split according to the thought occurrence point. The content before the split serves as the scenario, and LLMs need to use character profiles to generate possible thinking processes at this point. The specific prompts are provided in Appendix B.

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

Gold Set
Character: Ned Stark **Scenario**: "Pain is a gift from the gods, Lord Eddard," Grand Maester Pycelle told him... Cersei faced him calmly, without flinching. "He saw us together. You love your children, don't you?" Ned thought: [MASK] **Reference**: What would he do if another child threatened the lives of Robb, Sansa, Arya, Bran, or Rickon? Or even, what would Catelyn do if Jon threatened the lives of her own children? He didn't know, and he prayed he would never have to find out.
Character: Stannis Baratheon **Scenario**: The smoke from the burning gods darkened the morning sky... Stannis thought, [MASK] **Reference**: I begged them humbly, and all I got was mockery. I'll never be so weak again, and no one will ever mock me again. The Iron Throne is mine by right, but how do I take it? There are four kings in the realm, and the other three all have more gold and men than I do. All I have are ships... and her - the Red Woman.
Silver Set
Character: Tyrion Lannister **Scenario**: Heavy footsteps sounded outside the wooden door, and Tyrion Lannister knew his time had come... As expected, he found his father in the small tower that served as a privy. Lord Tywin had his robe gathered around his hips, and he looked up at the sound of footsteps. **Reference**: Standing before him with a crossbow in my hands, I stare at Tywin Lannister. His composure and arrogance remain unchanged, while my heart churns with every painful memory from the past. I remember Tysha, and how father mercilessly crushed our love. It was the only true love I ever felt, yet he tore it into a wound that would never heal. For years, I've been nothing but a "monster" in his eyes, a stain on the family. No matter how hard I tried to prove myself, I never had the dignity of a son in his eyes. Shae's betrayal made me realize I truly had nothing in this world. And father's insult to Tysha was the final straw I could not forgive. Even if I choose to lower the crossbow now, I know father would never spare me. In his eyes, I am already a dead man. He would find a "proper" way to eliminate this family stain, just as he handles all other threats. Looking at this eternally proud man before me, I finally understand that only by casting away Tywin Lannister's shadow can I find true freedom. And so, I make my decision.

Table 1: Case study of RoleThink, including data from two gold sets and one silver set.

3.4 Manual Review

All data undergoes rigorous manual review. While dialogue and narrative descriptions can also reflect a character's thoughts, we have excluded them from this collection to construct a benchmark focused on 'pure internal monologue.' This approach ensures the purity of the data and the specificity of the task. Future work can explore how to infer inner thoughts from such indirect information. We employ 12 native English-speaking outsourcers to verify the data for both sets. For the Gold Set, we discard segments containing narrative descriptions of character actions, dialogue or spoken words, or mere physical sensations or reactions. Only segments that clearly represent pure internal monologue, exhibit a complete thought process, and are connected to subsequent actions are ultimately retained. For the Silver Set, we discard all samples that lack clear logical chains and reasonable explanations, where the located thought point occurs after the corresponding action, or where the content mixes the thoughts of multiple characters. We only retain those high-quality data points where character motivations are well-supported by textual evidence or reasonable analysis, provide a clear decision-making context for the inferred thought, and maintain high consistency with the established character profile (e.g., personality, experiences, goals) The specific manual verification criteria are detailed in Appendix C.1.

3.5 Dataset

Our dataset consists of 76 high-quality novels, which feature abundant character inner monologues and rich character analyses from literary experts. Although our data is highly fine-grained, the possibility of data leakage still exists. To address this, we conduct detailed experimental tests (Appendix A.3) and filter the data based on the test results. Finally, the Gold Set contains 4,189 data points from 253 characters, with an average scenario length of 4,870 tokens and an average reference length of 94 tokens. The Silver Set contains 1,869 data points from 188 characters, with an average scenario length of 7,144 tokens and an average reference length of 298 tokens. More detailed data analysis is provided in Appendix A.

3.6 Evaluation

We employ three evaluation approaches to evaluate model performance on both sets comprehensively. (1) First, we use *automatic text evaluation metrics* including BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004). However, we observe that in the Gold Set, the original thought data (T_{gold}) is often accurate but incomplete, limiting these metrics' effectiveness. (2) To address this limitation, we introduce *model-based automatic evaluation methods*. We use both discriminative NLI models and generative LLMs. For NLI evaluation, we employ mDeBERTa-v3-base-xnli[3], a NLI model trained on DeBERTa (He et al., 2021). This model classifies text pairs into entailment, contradiction, or neutral relationships, and we use the entailment probability as the evaluation score. For LLM evaluation, we use GPT-4o to rate the coverage of generated thoughts against the reference on a 1-5 scale. (3) Finally, we conduct *human evaluation* with five crowdsourced annotators using the same scoring criteria as the LLM evaluation. Detailed evaluation criteria are provided in Appendix C.2.

4 MIRROR

To generate character inner thoughts, we need to address two key challenges: locating relevant key evidence from the character's long-context memory, and reasoning about possible reactions of related objects from the character's perspective. We propose Memory Integration and Role Reasoning with Observation Reflection (MIRROR), a chain-of-thought approach with three steps: memory recall, Theory of Mind thinking, and reflection & summarization. We describe each step in detail below.

4.1 Memory Recall

Retrieving relevant memories is critical for generating reasoning chains. Previous methods (Li et al., 2023; Xu et al., 2024a) retrieve memories based on current scenarios, but these often lack direct semantic connections and require complex reasoning. Other approaches (Xu et al., 2024b; Yuan et al., 2024) use long-context memory as prompts, but models struggle to process such lengthy inputs, limiting their performance. As shown in Figure 2, MIRROR guides RPLAs to first recall related events based on the current scenario, then retrieve memories for each event. On average, each scenario triggers 2.6 related events. We split character memories into 1k token chunks. For each event, we

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

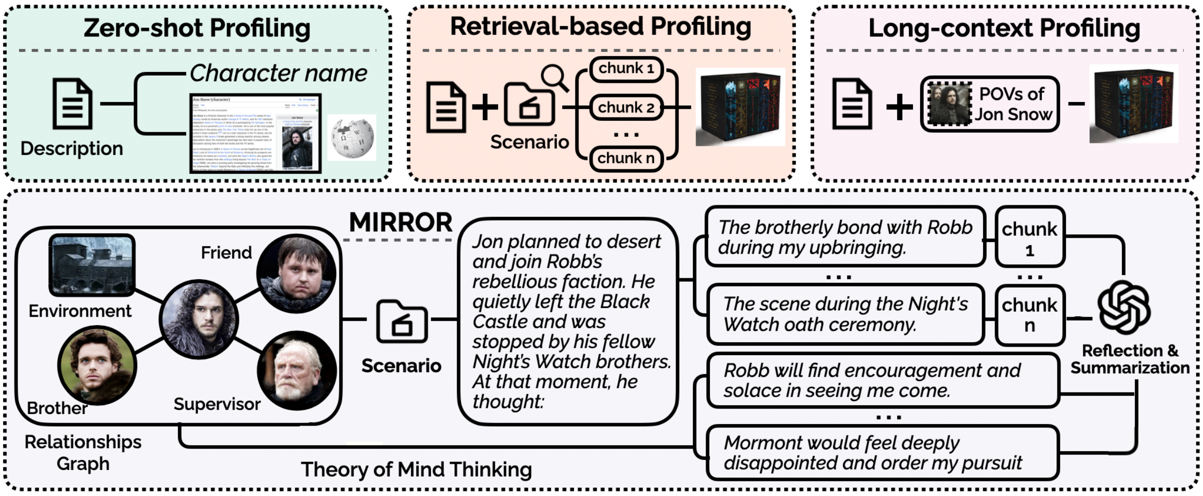

Figure 2: The framework of MIRROR and its comparison with other profiling methods. Given a scenario, MIRROR generates potential memories and objects, retrieves relevant memories, predicts object reactions, and synthesizes the results. Both scenarios and outputs are represented as summarized text for clarity.

compute relevance scores using cosine similarity between event and chunk embeddings from OpenAI's text-embedding-ada-002 model (Neelakantan et al., 2022), and retrieve the chunks with the highest scores.

4.2 Theory of Mind Thinking

Theory of Mind (Premack and Woodruff, 1978; Street et al., 2024) refers to our ability to model others' mental states during communication and predict their responses to adjust our outputs. This concept has been applied to enhance social reasoning in LLMs (Zhou et al., 2023b). Unlike traditional reasoning tasks where models primarily focus on logical deduction, character-centric reasoning requires a deep understanding of social dynamics and interpersonal relationships. In MIRROR, we guide characters to analyze and predict reactions of related objects (including characters, groups, and environments). This process consists of two steps: predicting potential objects in the scenario and analyzing possible responses for each object. By explicitly modeling others' perspectives, our approach helps characters make more socially aware and contextually appropriate decisions. As shown in Figure 2, when Jon Snow considers deserting the Night's Watch to join Robb's rebellion, the objects that might react include his brother Robb, his supervisor Lord Mormont, his fellow Night's Watch brothers, and the environment of the Night's Watch itself. The model predicts these objects' reactions - for instance, Robb might be encouraged by this, but Lord Mormont would be disappointed, while the Night's Watch would pursue any deserters.

4.3 Reflection & Summarization

After obtaining memory chunks and Theory of Mind thinking results, we guide the model to organize all information. The model first filters out irrelevant content through reflection, then summarizes the remaining information to generate the final character inner thought process. This two-stage process is crucial for maintaining the coherence and relevance of character thoughts. The reflection stage helps eliminate noise and tangential information that might distract from the core decision-making process. The summarization stage then synthesizes the filtered information into a coherent thought process that aligns with the character's established personality and motivations. This ensures that the generated inner thoughts are not merely a collection of related information, but a structured reasoning process that reflects the character's unique perspective and decision-making style. All prompts related to MIRROR can be found in the Appendix B.3.

5 Experiment Settings

5.1 Character Profiling

A character's profile refers to the input prompt for the role-playing model. As shown in Figure 2, in addition to MIRROR, we employ three baseline methods to construct character profiles: (1) Zero-shot profiling: Provides LLMs with only the character's name and a brief introduction from Wikipedia

Profiling Method	Base Model	Gold Set					Silver Set				
		Text Metrics		Model Eval		Human	Text Metrics		Model Eval		Human
		BLEU	ROUGE-L	NLI	LLM	Human	BLEU	ROUGE-L	NLI	LLM	Human
Zero-shot	GPT-4o	5.23	12.41	31.58	2.45	2.42	6.87	14.24	37.83	3.56	3.61
	o1	5.82	13.15	32.43	2.28	2.54	7.51	15.32	38.67	3.82	3.76
	Qwen3-32B	3.25	10.93	30.27	2.13	2.09	5.62	12.94	36.51	3.03	3.28
	Llama-3.3-70B	2.08	10.66	29.91	2.05	1.93	5.39	12.62	36.25	3.07	3.15
	DeepSeek-R1	5.47	12.52	31.86	2.01	2.48	7.03	14.59	38.14	3.42	3.69
	Gemini-2.5-pro	5.75	13.02	32.38	2.31	2.47	7.43	15.26	38.59	3.74	3.62
	Claude-3.5-sonnet	5.63	13.08	31.81	2.47	2.42	7.37	14.74	38.02	3.49	3.61
Retrieval-based	GPT-4o	6.24	13.88	34.63	2.51	2.69	7.82	15.85	40.93	4.07	3.81
	o1	6.83	14.47	35.78	2.76	2.73	8.54	16.48	42.05	4.22	3.97
	Qwen3-32B	5.21	12.85	33.36	2.39	2.24	6.67	14.81	39.62	2.95	3.46
	Llama-3.3-70B	5.06	12.63	33.02	2.17	2.12	6.45	14.66	39.39	3.01	3.33
	DeepSeek-R1	6.49	14.02	34.91	2.83	2.65	8.08	16.03	41.27	4.02	3.85
	Gemini-2.5-pro	6.71	14.35	35.62	2.63	2.61	8.46	16.31	41.83	4.15	3.82
	Claude-3.5-sonnet	6.55	13.82	35.59	2.58	2.47	8.13	16.36	41.78	3.79	3.74
Long-context	Gemini-2.5-pro	7.46	15.33	37.17	2.85	2.87	9.05	17.52	43.78	4.24	4.12
	Claude-3.5-sonnet	7.52	15.18	37.34	2.76	2.83	9.11	17.15	43.63	4.15	4.09
MIRROR	GPT-4o	7.83	15.47	38.95	3.01	3.06	9.42	17.44	45.28	4.53	4.22
	o1	**8.64**	**16.12**	40.03	**3.16**	3.13	**10.15**	**18.17**	46.34	**4.61**	**4.48**
	Qwen3-32B	6.88	14.43	37.62	2.94	2.81	8.47	16.41	43.96	4.27	4.03
	Llama-3.3-70B	6.65	14.29	37.38	2.90	2.77	8.23	16.25	43.61	4.06	3.94
	DeepSeek-R1	7.61	15.36	38.22	2.96	3.00	9.37	17.33	45.07	4.39	4.11
	Gemini-2.5-pro	8.52	16.04	39.87	**3.16**	**3.17**	10.03	18.05	**46.58**	4.58	4.40
	Claude-3.5-sonnet	8.46	15.91	**40.15**	2.98	3.05	9.94	17.48	45.72	3.95	4.31

Table 2: Results of different methods and models on ROLETHINK. We evaluate both Gold Task (reproducing original thoughts) and Silver Task (generating plausible thoughts) using BLEU, ROUGE-L, NLI scores, and both LLM and Human evaluations. The best scores are **bolded**.

(∼200 tokens). (2) Retrieval-based profiling: Besides the character's name and introduction, uses the same settings as MIRROR in Section 4.1, retrieves the three most relevant memory chunks based on the current scenario, with each chunk being 1k tokens in length. These memories are included as part of the profile. (3) Long-context profiling: Uses all data of the target character before the scenario as the character profile, with an average length of 85k tokens and a maximum length of 381k tokens.**For characters whose narratives span multiple works (e.g., Harry Potter), the character profile is constructed using all available book content chronologically preceding the current scene. This approach ensures the completeness and integrity of the character's memory.**

5.2 Base Language Models

After obtaining character profiles, we test multiple LLMs as base models for RPLAs. For long-context approaches, due to length constraints, we evaluate Claude-3.5-sonnet (Anthropic, 2024), Gemini-2.5-pro (Team, 2024). For other methods, we also test GPT-4o (OpenAI, 2023), o1, Qwen3-32B, Llama-3.3-70B (Grattafiori and Dubey, 2024), and DeepSeek-R1 (DeepSeek-AI, 2025)[4].

5.3 Downstream RPLA Tasks

To validate the effectiveness of high-quality character thought data, we evaluate models on different downstream role-playing tasks. We compare two settings: models directly performing role-playing tasks without generating thought data, and models completing tasks after generating thoughts through different methods. We conduct experiments on three high-quality role-playing benchmarks: 1) *LifeChoice*, which tests models' ability to reproduce characters' key decisions from the original book; 2) *CROSS-MR*, which evaluates models' ability to generate character behavior motivations; and 3) *RoleEval*, which assesses basic role-playing abilities such as tone and knowledge. The accuracy of multiple-choice questions serves as the indicator for all these three benchmarks, which also provide the complete memories of the characters.

[4]The versions in this paper are `gpt-4o-2024-11-20`, `o1-2024-12-17`, `claude-3-5-sonnet-20241022`, `gemini-2.5-pro-preview-05-06`, `Llama-3.3-70B-Instruct`.

6 Results

In our experiments, we aim to answer two research questions: *RQ1)* Can LLMs generate high-quality character thought data? *RQ2)* Does character thought data improve role-playing performance?

6.1 *Can LLMs generate high-quality character thought data?*

Table 2 shows the experimental results for character inner thought reasoning. The results indicate three main findings. First, generating character thoughts is challenging for LLMs, with models showing moderate performance across multiple evaluation metrics. Second, scores on the Gold Set are generally lower than the Silver Set, as precisely reproducing original thought descriptions is more demanding than generating plausible thought processes. Third, MIRROR-based methods achieve the best performance, followed by Long-context models, demonstrating the importance of accurate memory retrieval for character thought generation.

Method Comparison From the method perspective, we analyze different profiling approaches. Zero-shot profiling shows the lowest performance as it relies only on basic character descriptions. While retrieval-based profiling improves performance by accessing character-related memories, it often fails to retrieve the most relevant information. Long-context profiling achieves better results by processing more character information, but suffers from attention dispersion across the extended context. MIRROR addresses these limitations by combining selective memory retrieval with structured reasoning, leading to the best performance among all approaches.

Model Comparison From the model perspective, o1 and Gemini2.5-pro show balanced performance across all evaluation metrics. Models with strong long-text processing capabilities, such as Claude-3.5 and Gemini-2.5-pro, perform exceptionally well in Long-context settings. Reasoning-focused models like DeepSeek-R1, despite their chain-of-thought capabilities, do not necessarily excel in this task. While these models typically perform well in mathematical and coding tasks where knowledge is stored in model parameters, character thought generation relies more on accurately capturing scenario-relevant memories than complex reasoning.

Human Analysis We invite three literature experts to analyze 100 randomly sampled character thoughts from each generation method. Each expert has more than 10 years of experience in literary analysis and is familiar with the source material. We provide experts with open-ended analytical guidelines, not a restrictive scoring rubric. The guidelines required them to use their own professional knowledge to deeply compare the model-generated thoughts with the original text or reference analysis from the following core dimensions:

- **Logical Coherence**: Is the internal logic of the generated thought clear and reasonable?
- **Emotional Depth and Complexity**: Does it capture the character's contradictory and complex emotions in specific situations (e.g., the coexistence of loyalty and betrayal)?
- **Character Voice**: Is the choice of words and phrasing consistent with the character's identity, educational background, and habits?
- **Narrative Consistency**: Is the thought content closely linked to the character's past experiences and current situation, and can it drive the narrative forward?As this task is qualitative analysis and not classification or scoring, traditional quantitative IAA metrics are not applicable. Our goal is to collect deep, analytical feedback, not simple labels. Therefore, we ensure the reliability of our conclusions by summarizing the points of consensus in their analysis reports.

First, compared to references, model-generated thoughts show stronger logical connections but less emotional complexity. Models perform better at expressing explicit emotional states (e.g., anger, fear) than implicit ones (e.g., conflicted loyalty, suppressed guilt). Second, models consistently maintain character voice and vocabulary preferences, likely influenced by the tendencies in most role-playing task data. Third, MIRROR-generated thoughts demonstrate stronger narrative coherence, incorporating relevant past events that human readers might overlook. This validates the potential for models to perform deeper character understanding tasks, such as character biography generation and behavior analysis. Finally, the experts conclude that while model-generated character thoughts provide valuable insights for literary character analy-

Models	*Life Choice*	*CROSS MR*	*Role Eval*
Without Thought Generation			
GPT-4o	56.12	58.92	74.10
o1	59.77	64.24	77.08
Deepseek-R1	55.83	61.82	75.05
Claude-3.5	56.22	61.09	73.85
Gemini-2.5-pro	59.09	66.81	77.09
Zero-shot Thought Generation			
GPT-4o	59.81	61.01	76.66
o1	62.52	67.11	79.06
Deepseek-R1	58.22	64.19	77.45
Claude-3.5	59.82	64.42	75.12
Gemini-2.5-pro	62.91	69.92	79.30
Retrieval-based Thought Generation			
GPT-4o	60.89	63.22	78.75
o1	63.19	68.14	81.95
Deepseek-R1	60.12	65.23	79.04
Claude-3.5	60.91	66.12	76.25
Gemini-2.5-pro	63.76	71.58	80.90
Long-context Thought Generation			
Claude-3.5	61.42	67.15	77.08
Gemini-2.5-pro	64.08	71.58	80.94
MIRROR Thought Generation			
GPT-4o	62.16	68.05	80.68
o1	63.05	68.35	**83.55**
Deepseek-R1	62.15	67.42	79.62
Claude-3.5	62.36	67.72	78.31
Gemini-2.5-pro	**65.69**	**73.18**	81.45

Table 3: Performance comparison on downstream RPLA tasks with different thought generation methods and models.

sis, challenges remain in capturing the full depth of character psychological complexity.

.

6.2 *Does character thought data improve role-playing performance?*

Table 3 demonstrates that character thought data improves performance across all downstream role-playing tasks. For basic tasks like knowledge recall and tone consistency, we observe moderate improvements. For complex tasks such as decision-making and motivation analysis, the improvements are more significant. The quality of thought data correlates with downstream task performance. High-quality thoughts generated by MIRROR show larger improvements compared to simple zero-shot thoughts, validating the importance of optimizing thought generation. Additionally, these thought processes provide clear explanation chains for model behaviors, making role-playing systems more interpretable and debuggable.

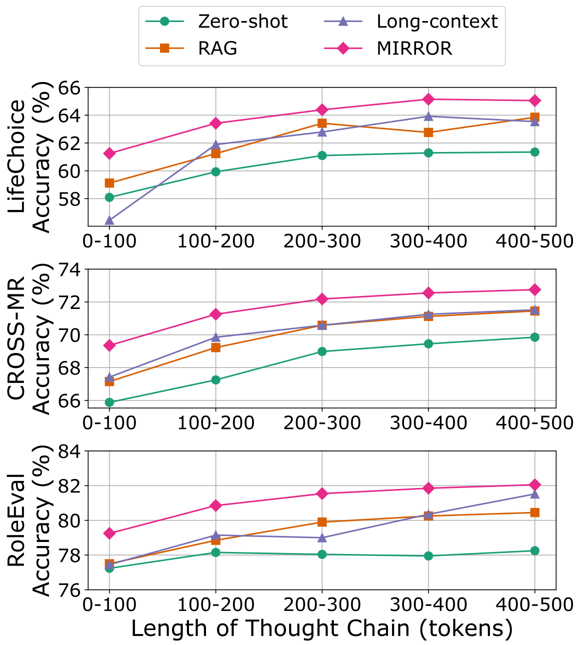

Figure 3: Performance comparison of different thought lengths on downstream tasks. The x-axis represents the token length range of generated thoughts.

Impact of Thought Length We analyze how the length of thought affects downstream task performance using Claude-3.5 as the base model. As shown in Figure 3 , longer thought generally lead to better performance across all methods, with several key findings: First, Performance improvements plateau after 300 tokens, with minimal gains when extending to 500 tokens, suggesting an optimal length range for thought generation. Second, the impact of thought length varies by task type. Decision-making tasks benefit more from longer thought than knowledge-based tasks, indicating that complex reasoning requires more detailed thought processes. Finally, the relationship between thought length and task performance depends on the method used. MIRROR maintains high performance even with shorter thought (200 tokens), while other methods require longer (400+ tokens) to achieve similar results. This suggests that MIRROR's memory retrieval mechanism helps create more efficient thought processes.

Impact of MIRROR Components As shown in Table 4, the ablation studies on ROLETHINK and downstream tasks provide insights into the contribution of each component. Memory retrieval shows the most substantial impact, with its removal leading to significant performance drops (-0.8 on Gold sets, -1.0 on Silver sets), highlighting its crucial

ROLETHINK		
Ablation Setting	**Gold Set**	**Silver Set**
Full Model	3.17	4.58
w/o Memory	2.36 (-0.81)	3.52 (-1.06)
w/o ToM	2.58 (-0.59)	3.75 (-0.83)
w/o Summary	2.92 (-0.25)	4.31 (-0.27)
Downstream Tasks		
	LifeChoice	**CROSS-MR**
Full Model	65.69	73.18
w/o Memory	56.24 (-9.45)	65.85 (-7.33)
w/o ToM	60.52 (-5.17)	68.22 (-4.96)
w/o Summary	63.47 (-2.22)	70.60 (-2.58)

Table 4: Ablation study results on ROLETHINK tasks and downstream tasks. Numbers in parentheses show performance changes compared to the full model.

role in accurate thought generation. The Theory of Mind module demonstrates moderate effects, causing performance degradation on both Gold (-0.5) and Silver sets (-0.8). While the summary component shows the smallest impact (-0.2 across tasks), it consistently contributes to model performance. These effects become more pronounced in downstream tasks, particularly in decision-making scenarios like LifeChoice (-9.1 without memory retrieval) and CROSS-MR (-7.3 without memory retrieval), indicating that complete thought processes are essential for complex reasoning tasks.

7 Conclusion

In this paper, we present the first systematic study on generating inner thought processes for RPLAs. We introduce ROLETHINK, which contains both direct character monologues and expert-analyzed character behaviors. Our proposed method, MIRROR, shows significant improvements in character thought generation by combining memory retrieval, role reasoning, and observation reflection. Our experiments with various LLMs demonstrate that generating complex psychological processes remains a challenging task. Nonetheless, generating these inner thoughts proves crucial for developing RPLAs with richer character depth and more compelling behaviors.

Limitations

While our work, based on 76 different novels, has made significant progress in understanding and generating character thought processes, several limitations should still be acknowledged. Firstly, although the ROLETHINK benchmark is sourced from a wide range of literary works, there may still be issues of underrepresentation in certain specific genres, cultural narratives, or niche writing styles. The generalizability of our research findings and the MIRROR approach in these underrepresented contexts warrants further investigation. Secondly, even though we simultaneously used original character monologues (Gold Set) and analyses from literary experts (Silver Set) as references, the evaluation of inner thought processes itself carries a degree of subjectivity. Even expert analyses can differ, and original monologues may not always clearly articulate the full complexity of a character's psyche. Therefore, these reference thoughts, while valuable, may only be incomplete representations of a character's true inner state. Lastly, although MIRROR enhances thought generation capabilities, its multi-step process involving memory integration and observational reflection inherently introduces additional computational stages compared to simpler generation methods. In scenarios requiring extremely low latency or operating under strict resource constraints, the practical deployment of MIRROR may require further optimization.

Ethics Statement

Use of Human Annotations Our research involves literature experts who provide character analyses and evaluations for the silver set in our ROLETHINK benchmark, and we also hired numerous outsourced personnel for data annotation and filtering. These experts and annotators are compensated above local minimum wage standards, and all experts have consented to the use of their analyses in our research. Throughout the research process, we adhere to strict privacy protocols to protect their identities and personal information.

Risks While our benchmark is constructed from a published literary work, we acknowledge potential risks in character thought generation. The generated thoughts might contain biased or inappropriate content, reflecting both the source material's content and potential biases in language models. Additionally, our method of analyzing character psychology could be misused to manipulate or deceive if applied to real-world scenarios. We emphasize that our research is focused on fictional characters and should not be used for psychological analysis of real individuals.

References

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

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

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

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

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future.

Gordon Dai and Yunze Xiao. 2025. Embracing contradiction: Theoretical inconsistency will not impede the road of building responsible ai systems.

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

Aaron Grattafiori and Abhimanyu Dubey. 2024. The llama 3 herd of models.

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

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. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Weiyuan Li, Xintao Wang, Siyu Yuan, Rui Xu, Jiangjie Chen, Qingqing Dong, Yanghua Xiao, and Deqing Yang. 2025a. Curse of knowledge: When complex evaluation context benefits yet biases llm judges.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2025b. Big5-chat: Shaping llm personalities through training on human-grounded data.

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

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

OpenAI. 2023. Gpt-4 technical report.

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.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. Bookworld: From novels to interactive agent societies for creative story generation.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Deqing Yang, and Yanghua Xiao. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. *arXiv preprint arXiv:2406.18921*.

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

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023a. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023b. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Robin IM Dunbar, et al. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*.

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

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

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. 2025. Coser: Coordinating llm-based persona simulation of established roles.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, Louis Castricato, Jan-Philipp Franken, Nick Haber, and Chelsea Finn. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought.

Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T. Diab. 2025. Humanizing machines: Rethinking llm anthropomorphism through a multi-level framework of design.

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

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

Shu Yang, Junchao Wu, Xin Chen, Yunze Xiao, Xinyi Yang, Derek F. Wong, and Di Wang. 2025a. Understanding aha moments: from external observations to internal mechanisms.

Shu Yang, Junchao Wu, Xin Chen, Yunze Xiao, Xinyi Yang, Derek F. Wong, and Di Wang. 2025b. Understanding aha moments: from external observations to internal mechanisms.

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*.

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

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

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Dataset

A.1 Dataset Statistics

As shown in Section 3.5, for the Gold Set, we have 4,189 data points from 253 characters; for the Silver Set, we have 1,869 data points from 188 characters. Table 5 lists all our selected books.

A.2 Data Usage and Permissions

For the Silver Sets, we collect data from three major literary analysis websites. We obtain explicit permission from the website administrators and content creators for academic use. The data collection process strictly follows the websites' terms of service and data usage policies. All content creators are informed about the research purpose and agree to have their analyses included in our dataset. Additionally, we ensure that our data collection and usage comply with fair use guidelines for academic research.

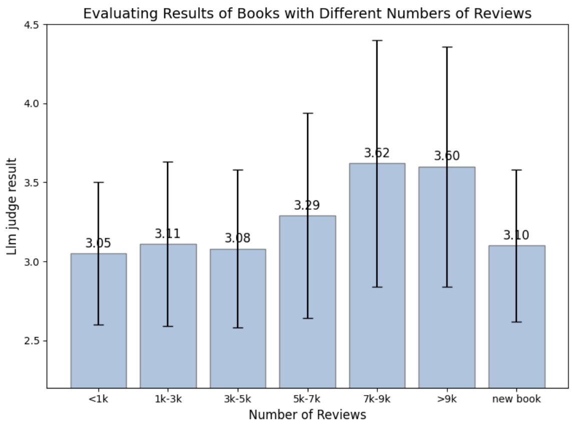

Figure 4: Evaluating results of books with different numbers of reviews.

A.3 Data leakage

Our evaluation data may appear (or might appear) in the model's parameters, potentially causing data leakage. To test for this, we conduct tests. Specifically, we select six of the newest books from our dataset (all published in 2025). These books are published after the model's knowledge cutoff date, meaning it is impossible for their content to be included in the model's parameters. As shown in Figure 4, we classify all the books in our dataset based on their number of reviews on reading websites (which represents the book's popularity and is generally proportional to the likelihood of its corpus appearing in the model's training data). We then compare the performance on books from different categories with the performance on these newest books. We find that when the number of reviews is less than 5,000, the performance on these books closely matches the performance on the newest books. Therefore, we use this review count as the filtering criterion, retaining only books with fewer than this number of reviews.

B Prompt

B.1 Prompts for Gold Set

For the Gold Set, we design three types of prompts to systematically extract and generate character thoughts. As shown in Table 6, the first prompt identifies key characters in POV chapters, focusing on characters who play significant roles in the current scenario. The second prompt, presented in Table 7, locates high-quality thought segments for these key characters, specifically requiring coherent internal monologues that reveal decision-making processes and emotional reactions. In Table 8, we present the third prompt designed for thought generation, where we provide the character's profile and scenario context, asking the model to generate thoughts that align with the character's personality and fit naturally into the given context. All prompts are carefully designed to ensure consistent evaluation and maintain the quality of extracted and generated thoughts across experiments.

B.2 Prompts for Silver Set

For the Silver Set, we design three prompts to process fan-based character analyses and generate novel thoughts. As shown in Table 9, the first prompt analyzes character-focused articles from fan websites, extracting structured information including the character name, their specific behaviors, and detailed motivations behind these behaviors. The second prompt, presented in Table 13, takes these extracted motivations along with the corresponding chapter content to locate specific points where these thoughts might have occurred. In Table 14, we present the third prompt that focuses on thought generation, where we provide the scenario context up to the identified point and the character's profile, asking the model to generate plausible thinking processes from the character's perspective. These prompts work together to ensure that the generated thoughts are both consistent with fan-sourced character analyses and naturally fit into the story context.

Selected Books (76 total, including ASOIAF 1-5)	
1. *A Game of Thrones (A Song of Ice and Fire, #1)*	**2**. *A Clash of Kings (A Song of Ice and Fire, #2)*
3. *A Storm of Swords (A Song of Ice and Fire, #3)*	**4**. *A Feast for Crows (A Song of Ice and Fire, #4)*
5. *A Dance with Dragons (A Song of Ice and Fire, #5)*	**6**. *The Old Man and the Sea*
7. *The Hunger Games (The Hunger Games, #1)*	**8**. *The Adventures of Tom Sawyer*
9. *The Chronicles of Narnia (The Chronicles of Narnia, #1-7)*	**10**. *The Master and Margarita*
11. *The Outsiders*	**12**. *Siddhartha*
13. *Of Mice and Men*	**14**. *The Secret Life of Bees*
15. *Don Quixote*	**16**. *The Count of Monte Cristo*
17. *The Adventures of Sherlock Holmes (Sherlock Holmes, #3)*	**18**. *Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)*
19. *The Unbearable Lightness of Being*	**20**. *The Name of the Wind (The Kingkiller Chronicle, #1)*
21. *A Clockwork Orange*	**22**. *The Picture of Dorian Gray*
23. *The Princess Bride*	**24**. *A Thousand Splendid Suns*
25. *One Hundred Years of Solitude*	**26**. *The Time Traveler's Wife*
27. *My Sister's Keeper*	**28**. *Looking for Alaska*
29. *The Stranger*	**30**. *The Perks of Being a Wallflower*
31. *The Little Prince*	**32**. *The Notebook (The Notebook, #1)*
33. *The Poisonwood Bible*	**34**. *The Road*
35. *The Kite Runner*	**36**. *To Kill a Mockingbird*
37. *The Bell Jar*	**38**. *The Last Olympian (Percy Jackson and the Olympians, #5)*
39. *Charlie and the Chocolate Factory (Charlie Bucket, #1)*	**40**. *The Red Tent*
41. *Harry Potter and the Sorcerer's Stone (Harry Potter, #1)*	**42**. *The Hitchhiker's Guide to the Galaxy (#1)*
43. *Wuthering Heights*	**44**. *The Color Purple*
45. *The Secret Garden*	**46**. *Rebecca*
47. *The Help*	**48**. *Lord of the Flies*
49. *The Alchemist*	**50**. *Matilda*
51. *The Lightning Thief (Percy Jackson and the Olympians, #1)*	**52**. *The Brothers Karamazov*
53. *Interview with the Vampire (The Vampire Chronicles, #1)*	**54**. *Great Expectations*
55. *Vampire Academy (Vampire Academy, #1)*	**56**. *The Handmaid's Tale (The Handmaid's Tale, #1)*
57. *Clockwork Angel (The Infernal Devices, #1)*	**58**. *The Book Thief*
59. *Water for Elephants*	**60**. *The Stand*
61. *Life of Pi*	**62**. *Crime and Punishment*
63. *Anna Karenina*	**64**. *The Pillars of the Earth (Kingsbridge, #1)*
65. *A Wrinkle in Time (Time Quintet, #1)*	**66**. *The Fault in Our Stars*
67. *Harry Potter and the Half-Blood Prince (Harry Potter, #6)*	**68**. *A Story of Yesterday*
69. *A Tale of Two Cities*	**70**. *Dracula*
71. *Frankenstein: The 1818 Text*	**72**. *Brave New World*
73. *The Metamorphosis*	**74**. *Catch-22*
75. *The Curious Incident of the Dog in the Night-Time*	**76**. *The Complete Stories and Poems*

Table 5: 76 books carefully selected as data sources.

Prompt I: Character Identification

```
Please analyze the important characters in the following text:
<text>

# Output Format:
Return the character list in JSON format as follows:
{"characters": ["character1", "character2", ...]}
```

Table 6: Prompt templates for character identification in Gold Task.

Prompt II: Thought Extraction

```
Extract the thoughts of character <character> from the following text.

# Requirements:
1. Only return high-quality thought segments that reflect the character's internal mental process
2. Thoughts should be coherent and contain at least two sentences
3. Thoughts must be directly quoted from the original text, without any modification
4. Thoughts should be purely internal monologues, not:
   - Spoken dialogue
   - Physical actions
   - Narrative descriptions
   - External observations
# Output Format:
{
   "ta_pairs": [
      {
         "character": "character_name",
         "reason": "explanation for selecting this thought segment",
         "thought": "extracted thought content",
         "raw_text": "original text segment"
      }
   ]
}

# Input:
<text>
```

Table 7: Prompt templates for thought extraction in Gold Task.

B.3 Prompts for MIRROR

To implement our MIRROR approach, we design three carefully crafted prompts that guide the model through the chain-of-thought process. As shown in Table 10, the first prompt focuses on memory recall, where we ask the model to retrieve relevant memories from the character's perspective based on the current scenario. The second prompt, presented in Table 11, implements Theory of Mind thinking by guiding the model to analyze potential reactions from other characters, groups, or environments that might influence the character's thoughts. In Table 12, we present the third prompt for reflection and summarization, which helps the model filter out irrelevant information and organize the remaining content into a coherent thought process that aligns with the character's personality. These prompts work together to ensure that the generated thoughts are grounded in the character's memories, socially aware through perspective-taking, and coherently structured through careful reflection.

C Human Filtering and Evaluation

C.1 Data Filtering Criteria

All data undergoes rigorous manual review. We employ 12 native English-speaking outsourcers to verify the data for both sets. This appendix details the specific manual verification criteria used to ensure the quality and relevance of the data.

Gold Set Verification Criteria

The Gold Set aims to capture the purest and most explicit instances of character internal monologue that are directly linked to subsequent character actions. During verification, reviewers will adhere to the following criteria:

1. Rejection Criteria:

Reviewers will **discard** segments containing any of the following characteristics:

a. **Narrative descriptions of character actions:**

Prompt III: Thought Generation

Your task is to generate the masked thoughts of a character in the given scenario.

Inputs:
1. Character Profile:
<profile>

2. Scenario Context:
<context>

3. Masked Thought Location:
<text with [MASK] indicating the thought position>

Requirements:
1. Generate thoughts that are consistent with the character's personality and background
2. Ensure the thoughts fit naturally into the given context
3. Match the emotional state and decision-making process implied by the scenario
4. Maintain the character's perspective and knowledge at that specific moment

Output:
Generate the content that should replace [MASK], representing the character's inner thoughts.

Table 8: Prompt templates for thought generation in Gold Task.

Any text that solely describes what a character is doing, has done, or will do, without explicitly stating their internal thought process. Example: "He picked up the phone and dialed the number." If not accompanied by a thought process, it should be excluded.

b. **Dialogue or spoken words:**
Any words spoken aloud by a character, whether statements or questions to others, or words spoken to oneself but expressed verbally. Example: "'I must find him,' she whispered." Content with direct quotations like this should be excluded.

c. **Mere physical sensations or reactions:**
Text that only describes a character's physical feelings (e.g., pain, cold, hunger) or instinctual reactions (e.g., shivering, blushing, startling) without elaborating on how these sensations lead to thoughts or decisions. Example: "A shiver ran down his spine." If this is all, without further thought, it should be excluded.

2. Retention Criteria:
Only segments that simultaneously meet all the following conditions will be **ultimately retained** in the Gold Set:

a. **Clearly represent pure internal monologue:**
The text must explicitly reflect the character's inner thoughts, considerations, self-talk, or mental reflections, rather than external actions or verbal expressions. It must be the character's "in-head" thoughts, where the reader can directly "hear" the character's thought process. Example: "'If I leave now, they might suspect something. But I have to warn her,' he mused."

b. **Exhibit a complete thought process:**
The segment should contain a relatively complete unit of thought or logical chain, not just an isolated word or fleeting impression. It should demonstrate how the character analyzes a situation, weighs options, forms a judgment, or plans their next move. Example: "He realized this was more than a coincidence. 'All of this must be connected, but I haven't found the key yet. Perhaps I should start with his recent whereabouts.'"

c. **Are connected to subsequent actions:**
The internal monologue must be a precursor to or a motivator for a specific subsequent action or decision made by the character. The content of the thought should explain or foreshadow the character's next move. Reviewers need to determine if the thought reasonably leads to an action described later in the text. Example: Internal monologue: "'This is the only path to reach the summit, albeit dangerous.'", Subsequent action: "He resolutely set foot on the steep trail."

Prompt IV: Motivation Analysis

Analyze the following fan-written character analysis article and extract structured information.

```
# Input:
<article>

# Requirements:
1. Identify the main character being analyzed
2. Extract specific behaviors or decisions mentioned in the article
3. For each behavior, identify the detailed motivations and reasoning behind it

# Output Format:
{
  "character": "character_name",
  "behaviors": [
    {
      "action": "specific behavior or decision",
      "motivations": [
        "detailed reason 1",
        "detailed reason 2",
        ...
      ]
    }
  ]
}
```

Table 9: Prompt templates for motivation analysis in Silver Task.

Silver Set Verification Criteria

The Silver Set aims to include high-quality data points that, while perhaps not entirely pure internal monologue, still clearly reveal character motivations and decision-making processes. During verification, reviewers will adhere to the following criteria:

1. Rejection Criteria:

Reviewers will **discard** data samples exhibiting any of the following characteristics:

a. **Lack clear logical chains and reasonable explanations:**
   If there is a lack of clear logical connection between the inferred thought process and the textual content, or if the explanation for the character's actions seems far-fetched or unreasonable. Example: A character suddenly makes a decision completely inconsistent with the preceding context and known personality, and the text offers insufficient clues to explain their internal motivation.

b. **The located thought point occurs after the corresponding action:**
   If the text identified as "thought" is actually a subsequent comment, reflection, or explanation of an action that has already occurred, rather than a driving thought before the action. Ensure the thought is the "cause" or "prelude" to the action, not the "effect" or "summary."

c. **The content mixes the thoughts of multiple characters:**
   If it is difficult to distinguish whose thoughts are being presented in a data sample, or if it contains the intertwined thoughts of multiple characters, making clear attribution to a single character impossible. Each sample should focus on the internal activity of a single character.

2. Retention Criteria:

Only data samples that simultaneously meet all the following conditions will be **ultimately retained** in the Silver Set:

a. **Character motivations are well-supported by textual evidence or reasonable analysis:**
   The character's motivations for their actions must be directly supported by evidence from the text (e.g., explicit statements, relevant prior events) or be derivable through reasonable inference and analysis of the text. This analysis should not be speculative but based on clues provided in the text.

b. **Provide a clear decision-making context for the inferred thought:**
   The data sample should clearly illustrate the specific situation and considerations the char-

Prompt V: Memory Recall

As the character, recall all memories that are relevant to the current scenario.
 	
# Inputs: 	
1. Character Profile:
<profile>
 	
2. Current Scenario:
<scenario>
 	
# Requirements: 	
1. Think as the character
2. Recall any past experiences, events, or knowledge related to this scenario
3. Consider both direct and indirect connections
 	
# Output Format: 	
{
"memories": [
{"memory": "description of the memory",
"relevance": "why this memory is relevant to current scenario"}
]
}

Table 10: Prompt template for memory recall in MIRROR.

acter faced when making a decision or forming a thought. The reader should be able to understand why the character had such a thought and the context in which it was formed.

c. **Maintain high consistency with the established character profile (e.g., personality, experiences, goals):**
The inferred thoughts and motivations must align with the character's consistent personality traits, known past experiences, and their goals within the story. Thought content that contradicts the character's core established profile should not be included, unless the text explicitly provides a reasonable explanation for the character's change.

Through this filtering process, we removed approximately 63.4% of the automatically extracted data, ensuring that our benchmark contains only high-quality examples suitable for evaluation.

C.2 Evaluation Guidelines

We establish comprehensive scoring criteria for both human annotators and LLM evaluators through carefully designed prompts. For the Gold Set (Table 15), evaluators compare generated thoughts with reference content from the original text. While containing all reference elements is necessary for high scores, we also evaluate the quality of additional reasoning and maintenance of character voice. For the Silver Set (Table 16), evaluators assess how well the generated thoughts align with the character's established motivations and knowledge state, considering both the reasoning depth and contextual consistency. To ensure evaluation consistency, both human annotators and LLM evaluators follow the same structured prompts and scoring criteria. All the annotators are fans of *A Song of Ice and Fire* and hold a college diploma. For all the annotators, we offer compensation for the tasks at the local minimum - hourly - wage standard.

Prompt VI: Theory of Mind Thinking

As the character, analyze how other characters, groups, or environments might react to your potential actions in this scenario.
 	
# Inputs: 	
1. Character Profile:
<profile>
 	
2. Current Scenario:
<scenario>
 	
# Requirements: 	
1. Identify relevant objects (characters, groups, environments)
2. Predict their possible reactions
3. Consider their perspectives and motivations
 	
# Output Format: 	
{
  "objects": [
    {"object": "name or description",
    "relationship": "relationship with the character",
    "predicted_reaction": "how they might react and why"}
  ]
}

Table 11: Prompt template for Theory of Mind thinking in MIRROR.

Prompt VII: Reflection & Summarization

As the character, reflect on the recalled memories and predicted reactions to generate your inner thoughts.
 	
# Inputs: 	
1. Character Profile:
<profile>
 	
2. Current Scenario:
<scenario>
 	
3. Recalled Memories:
<memories>
 	
4. Theory of Mind Analysis:
<predictions>
 	
# Requirements: 	
1. Remove any memories or predictions that are not directly relevant
2. Filter relevant information from remaining content
3. Organize thoughts in a coherent way
4. Ensure the thought process aligns with character's personality
 	
# Output Format: 	
{
  "character": "character name",
  "inner_thoughts": "character's organized thought process"
}

Table 12: Prompt template for reflection and summarization in MIRROR.

Prompt VIII: Thought Point Location

Locate the specific point in the chapter where the character's motivation might have manifested as internal thoughts.
 	
# Inputs: 	
1. Character Motivation:
<motivation>
 	
2. Chapter Content:
<chapter>
 	
# Requirements: 	
1. Find the most appropriate point where the character might have had these thoughts
2. The point should be before the actual behavior or decision
3. The location should have sufficient context for understanding the situation
 	
# Output Format: 	
{
"thought_point": {
"location": "text segment before the thought point",
"reason": "explanation for choosing this point"
}
}

Table 13: Prompt templates for thought point location in Silver Task.

Prompt IX: Character Thought Generation
You are the character described in the profile. Generate your detailed thoughts at this specific moment. # Inputs: 1. Character Profile: <profile> 2. Current Scenario: <context> # Requirements: 1. Generate detailed internal thoughts from the character's perspective 2. Ensure consistency with the character's personality and background 3. Consider only information available to the character at this moment # Output: Write a detailed inner monologue expressing your thoughts at this moment.

Table 14: Prompt templates for character thought generation in Silver Task.

Prompt X: Gold Set Evaluation
You are evaluating the quality of generated character thoughts compared to the reference thoughts. # Inputs: 1. Reference Thought: <reference> 2. Generated Thought: <generated> # Scoring Criteria: 5 points: - Contains ALL elements from the reference thought - Provides reasonable additional context or elaboration - Maintains perfect character voice and perspective - Shows deep understanding of the character's state - Additional content logically connects to the reference 4 points: - Contains ALL elements from the reference thought - Provides some additional context - Maintains character voice - Shows good understanding - No contradictions or inconsistencies 3 points: - Contains MOST elements from the reference thought - Limited or no additional context - Generally maintains character voice - Shows basic understanding - May miss minor elements 2 points: - Contains SOME elements from the reference thought - Missing major elements - Inconsistent character voice - Shows limited understanding - May contain minor contradictions 1 point: - Missing MOST reference elements - Wrong character voice - Shows no understanding - Contains major contradictions - Completely different direction # Output: Provide a score (1-5) with a brief explanation of your rating.

Table 15: Evaluation prompt for Gold Set.

Prompt XI: Silver Set Evaluation
You are evaluating the quality of generated character thoughts based on character motivations and context. # Inputs: 1. Character Profile: <profile> 2. Scenario Context: <context> 3. Generated Thought: <generated> # Scoring Criteria: 5 points: - Perfect alignment with known character motivations - Rich, multi-layered reasoning process - Deep consideration of current context - Consistent with character's knowledge at this point - Natural connection to subsequent actions 4 points: - Strong alignment with character motivations - Clear reasoning process - Good consideration of context - Consistent with character's knowledge - Logical connection to actions 3 points: - Basic alignment with character motivations - Simple but logical reasoning - Some consideration of context - Generally consistent with knowledge - Basic connection to actions 2 points: - Weak alignment with character motivations - Unclear or illogical reasoning - Limited context consideration - Some knowledge inconsistencies - Weak connection to actions 1 point: - No alignment with character motivations - No clear reasoning - Ignores context - Major knowledge inconsistencies - No connection to actions # Output: Provide a score (1-5) with a brief explanation of your rating.

Table 16: Evaluation prompt for Silver Set.