CogDual: Enhancing Dual Cognition of LLMs via Reinforcement Learning with Implicit Rule-Based Rewards

Anonymous ACL submission

Abstract

Role-Playing Language Agents (RPLAs) have emerged as a significant application direction for Large Language Models (LLMs). Existing approaches typically rely on prompt engineering or supervised fine-tuning to enable models to imitate character behaviors in specific scenarios, but often neglect the underlying *cognitive* mechanisms driving these behaviors. Inspired by cognitive psychology, we introduce **CogDual**, a novel RPLA adopting a *cognize-then-respond* reasoning paradigm. By jointly modeling external situational awareness and internal self-awareness, CogDual generates responses with improved character consistency and contextual alignment. To further optimize the performance, we employ reinforcement learning with two general-purpose reward schemes designed for open-domain text generation. Extensive experiments on the CoSER benchmark, as well as Cross-MR and LifeChoice, demonstrate that CogDual consistently outperforms existing baselines and generalizes effectively across diverse role-playing tasks.

1 Introduction

With the rapid advancement of Large Language Models (LLMs), recent years have witnessed a surge of research on role-playing (Chen et al., 2023; Tao et al., 2024b; Chen et al., 2024c, 2025b). Role-Playing Language Agents (RPLAs) are designed to equip LLMs with human-like capabilities, enabling them to emulate specific characters across diverse scenarios, while exhibiting behaviors and expressions consistent with the character's profile and context (Zhou et al., 2023).

Previous efforts have primarily focused on constructing role-playing evaluation benchmarks (Chen et al., 2024b; Tu et al., 2024) and improving model performance through prompt engineering or Supervised Fine-Tuning (SFT). These approaches typically assess whether the model's responses align with a character's profile in specific scenarios.

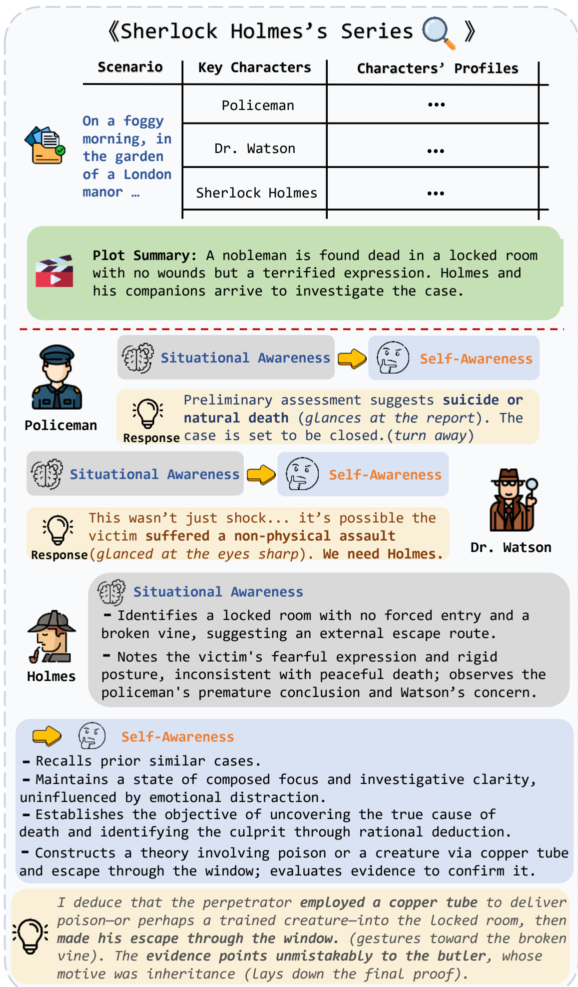

Figure 1: An example of CogDual. Before a character delivers a formal response, it undergoes a *dual cognitive reasoning process*, encompassing external Situational Awareness and internal Self-awareness.

Additionally, some studies employ multiple-choice formats to assess the model to infer motivations (Yuan et al., 2024), predict behaviors (Xu et al., 2024), or analyze psychological states (Wang et al., 2024a), thereby quantifying character consistency and fidelity. However, these approaches overlook a critical dimension: *as anthropomorphized agents, RPLAs should engage in cognitive processes involving both situational and self-awareness* rather than merely replicating superficial linguistic patterns or behavioral tendencies.

From the perspective of cognitive psychology (Grice, 1975; Clark and Brennan, 1991; Tomasello, 2010), human role-related behaviors emerge from an integrated cognitive process involving environmental perception, others' behaviors, and introspection of one's emotions and intentions. This cognitive process plays a crucial role prior to action generation. Building on this foundation, we propose **CogDual**, a RPLA that incorporates dual cognitive modeling, combining outward Situational Awareness and inward Self-Awareness, and embedding a *cognize-then-respond* paradigm into its reasoning process, as illustrated in Figure 1. By prioritizing cognition-driven generation, CogDual enhances both contextual relevance and psychological consistency in responses, ultimately improving performance on role-playing tasks.

Motivated by the need to adapt reward modeling for general-purpose text generation, we design two broadly applicable reward schemes: (1) the Inference-Conditioned Likelihood Gain (ICLG) Reward, which quantifies how the intermediate cognitive steps improve response likelihood, and (2) the Latent Semantic Alignment (LSA) Reward, which assesses semantic similarity between generated responses and gold-standard references. Based on these reward designs, we employ reinforcement learning to enhance CogDual's performance over the supervised fine-tuning baseline.

In contrast to contemporary studies such as Ji et al. (2025) and Xu et al. (2025), which also explore strategies to enhance the reasoning capabilities of RPLAs, our approach distinguishes itself by emphasizing the construction of a comprehensive character cognition process before response generation. Unlike their fragmented self-questioning or isolated mental state simulation, our dual cognitive reasoning process generates coherent, contextually grounded responses by tightly aligning psychological dynamics with narrative context.

The contributions of this work are as follows:

- We formalize the *cognize–then–respond* paradigm for RPLAs and propose CogDual, the first agent to implement dual cognitive modeling through **Situational Awareness** and **Self-Awareness**, providing a more psychologically plausible simulation of human-like behavior generation.
- We design two reward schemes and demonstrate their effectiveness through reinforcement learning on the CoSER benchmark (Wang et al., 2025b), achieving up to a 9.24% average improvement over baseline. The proposed reward design may serve as a reference for future research on evaluating text generation in general-domain applications.
- Through extensive experiments on Cross-MR (Yuan et al., 2024) and LifeChoice (Xu et al., 2024) benchmarks, we show CogDual's superior cross-task transferability, outperforming all baseline methods.

2 Related Work

2.1 Role-Playing Language Agents

Early investigations into RPLAs centered on character understanding, including character prediction from narrative texts and movie scripts (Brahman et al., 2021; Yu et al., 2024). With advances in LLMs, recent studies have extended RPLAs to facilitate character imitation through instruction-based reasoning and supervised fine-tuning, especially in dialogue and knowledge-intensive tasks (Shao et al., 2023; Wang et al., 2024b, 2025b). Beyond imitation, a growing body of work has shifted focus toward evaluating the internal coherence of character-driven behaviors. Studies such as (Yuan et al., 2024; Xu et al., 2024; Wang et al., 2024a) have introduced evaluative frameworks incorporating motivation recognition, persona-driven decision making, and psychological evaluation, allowing for a more nuanced analysis of the character consistency and behavioral plausibility of RPLAs.

2.2 LLM-Based Cognitive Modeling

Recent studies have increasingly explored the cognitive capacities of LLMs, particularly their ability to exhibit human-like behaviors in dialogic settings (Thoppilan et al., 2022; Park et al., 2023). This includes alignment with traits such as self-awareness (Shinn et al., 2023), emotion understanding (Rashkin et al., 2019), intent recognition (Chen et al., 2025a), and deliberative reasoning (Wei et al., 2023; DeepSeek-AI et al., 2025). These abilities are often evaluated in interactive contexts like multi-agent simulations (Li et al., 2023), narrative generation (Wu et al., 2025b), role-playing (Chen et al., 2024c), and chatbot systems (Wu et al., 2025a). However, recent work highlights that LLMs lack internal psychological states and intrinsic motivations, limiting the depth of their cognitive behaviors (Wang et al., 2025a). Our work adopts a cognitive psychology perspective to more rigorously define and examine LLM cognition in role-play settings.

2.3 Reasoning Techniques in LLMs

Recent research has shifted focus from train-time to test-time scaling, with notable success across various tasks such as math problem solving (Yang et al., 2024; Ma et al., 2025), logical puzzle reasoning (Xie et al., 2025) and tool-integrated reasoning (Lu et al., 2025; Qian et al., 2025; Feng et al., 2025a). However, Feng et al. (2025b) has highlighted the limitations of reasoning-augmented models (OpenAI et al., 2024; DeepSeek-AI et al., 2025) in role-playing scenarios. These models often suffer from stylistic drift between their reasoning traces and character-based generation, thereby undermining the coherence and consistency required for effective role enactment in RPLAs. Our study aims to enhance the generalizability of RPLAs across tasks and domains across various standard benchmarks by reinforcing reasoning process through a cognitively grounded template.

3 Methodology

3.1 Cognition-Driven Reasoning Paradigm

> *"Cognition is the activity of knowing: the acquisition, organization, and use of knowledge." — Neisser, 1967*

This foundational perspective highlights cognition as the driving force behind meaningful communication, rather than a passive background process. While current LLM-based RPLAs can produce fluent utterances, they often overlook the cognitive mechanisms essential to genuine human interaction (Grice, 1975; Clark and Brennan, 1991). Motivated by this, we propose a **cognition-driven reasoning paradigm** for RPLAs, which explicitly embeds cognitive reasoning between perception and response to simulate the psychological steps a human character might take. Tomasello (2010) shows that individuals interpret environmental and social cues through mental representations, which guide intentional actions, making the transition from external to internal cognition central to human communication. We thus focus on dual cognition, progressing from external perception to internal reflection. By modeling this cognitive transition, we propose **CogDual**, which enables RPLAs to generate dual cognition before responding.

3.2 Preliminaries

To formally ground the cognition-driven reasoning paradigm introduced above, we first define the key notations and basic concepts used throughout this work. A multi-party dialogue setting is defined over a set of characters, denoted as $\mathcal{O} = \{o_1, o_2, \ldots, o_{|\mathcal{O}|}\}$. Formally, let $\mathcal{M}$ represent an LLM simulating a specific character $c \in \mathcal{O}$ in a dialogue scene. The model has access to the character's profile $\mathcal{P}_c$, a global scene description $\mathcal{S}$, which may include the current task, storyline, and other elements, and a historical dialogue context $\mathcal{D}_t = \{d_1, d_2, \ldots, d_t\}$, where each d_i represents an utterance, an action, or a thought from a certain character at turn i.

The objective of CogDual is to incorporate dual cognition to establish a *cognize-then-respond* pattern. At each time step t, $\mathcal{M}$ first performs cognition, forming an internal thinking of the situation, other characters, and itself, and then generates a response. This process is formalized as:

$$c_t, d_t = \mathcal{M}\left(\mathcal{P}_c, \mathcal{S}, \mathcal{O}, \mathcal{D}_{t-1}\right), \quad \mathcal{D}_0 = \emptyset, \quad (1)$$

where c_t denotes the dual cognitive reasoning process at turn t, and d_t is the generated response conditioned on c_t and the given inputs. Compared to previous works (Wang et al., 2024b; Tu et al., 2024; Wang et al., 2025b) that directly generate d_t, our study requires LLMs to perform explicit cognitive thinking before response generation, producing structured representations of the current environment, other characters, and the agent's own state. This mechanism is designed to enhance the model's contextual understanding in complex scenarios, while improving the coherence and interpretability of character behavior.

3.3 Dual Cognition of RPLAs

In this part, we detail the structure of the Dual Cognition of RPLAs, which consists of two key components: **Situational Awareness** and **Self-Awareness**, forming a reasoning process that flows from the outer environment to the inner self.

Situational Awareness Situational Awareness refers to the RPLA's ability to perceive and interpret the environment and other characters within a dialogue scene. It consists of two components: **(1) Environmental Perception (EP):** Extracting salient cues from spatial layout, temporal shifts, and scene dynamics, such as changes in atmosphere, character positions or expressions, and dia-

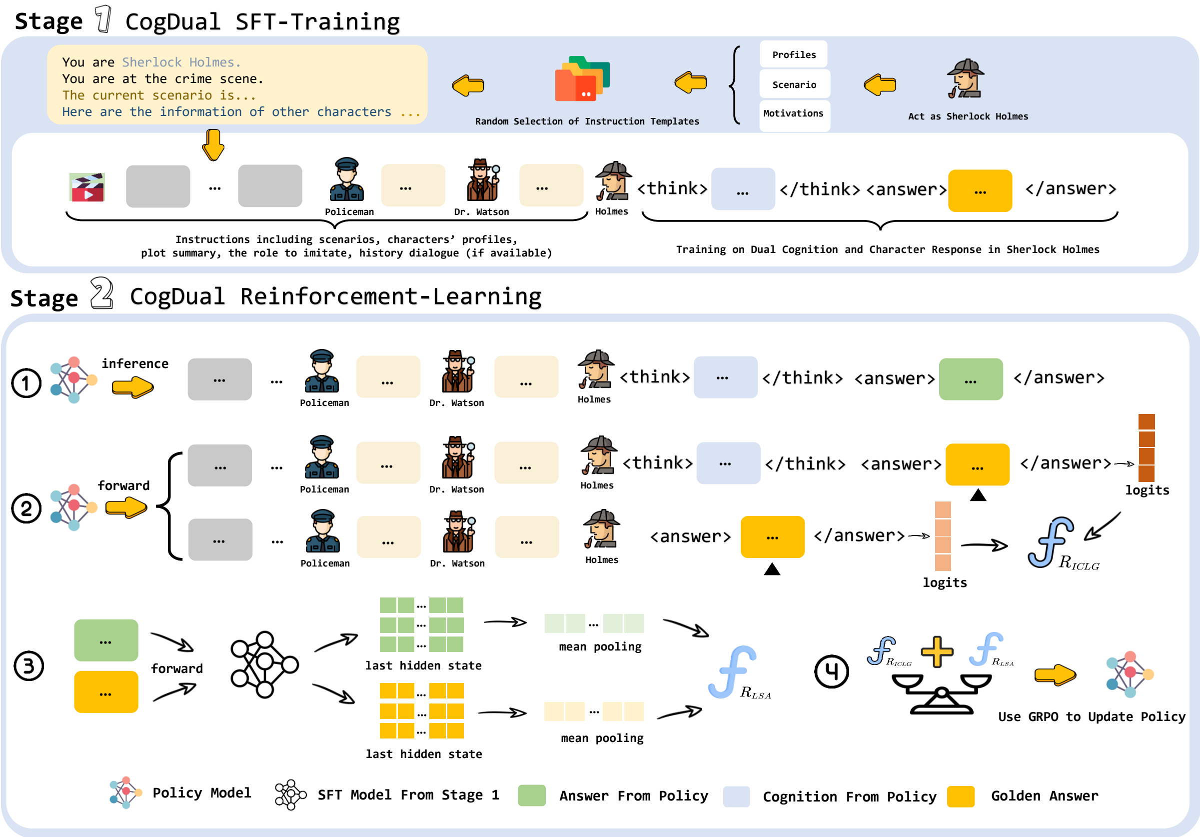

Figure 2: Overview of CogDual training. (1) Stage 1: Supervised fine-tuning using the role-specific dual cognitive reasoning process and corresponding response generated by the RPLA. (2) Stage 2: Reinforcement learning with GRPO, optimized based on the ICLG Reward and the LSA Reward.

logue interruptions. These form the initial layer of cognitive input, grounding the agent's understanding of the unfolding situation. **(2) Perception of Others:** Comprising three subcomponents:

- **Behavior Analysis (BA):** Identifying key actions or speech patterns from others that may affect the agent's response;
- **Emotion Analysis (EA):** Inferring emotional states from behavior and context, and assessing their impact on the agent;
- **Intention Analysis (IA):** Inferring others' intentions to guide the agent's reactions.

This process can be formally represented as:

$$SA = \langle \mathrm{EP}, \mathrm{BA}, \mathrm{EA}, \mathrm{IA} \rangle, \qquad (2)$$

where SA denotes **Situational Awareness** formed through hierarchical perception and interpretation.

Self-Awareness Self-Awareness forms a core component of the cognitive architecture in RPLAs, enabling introspection and adaptive decision making. It comprises four interrelated elements:

- **Key Memory Activation (KMA):** Recalling autobiographical or episodic memories relevant to the current context;
- **Self-Emotion (SE):** Recognizing and evaluating internal emotional states that influence perception and behavior;
- **Self-Intention (SI):** Maintaining context-driven goals that guide actions;
- **Internal Strategy (IS):** Integrating memory, emotion, and intention into coherent reasoning for planning and outcome anticipation.

This process can be formally represented as:

$$SA_{self} = \langle \mathrm{KMA}, \mathrm{SE}, \mathrm{SI}, \mathrm{IS} \rangle, \qquad (3)$$

where SA_{self} denotes the **Self-Awareness** formed through the agent's self-cognition.

3.4 Dual Cognition Behavior Learning

We propose two approaches for dual cognition: a cognitive-based Chain-of-Thought (CB-CoT) prompting method and a two-stage training framework, as shown in Figure 2. This section focuses on the latter; CB-CoT is detailed in Appendix D. For supervised training, we construct a dataset $\mathcal{D}_{\mathrm{SFT}}$ with dual cognition trajectories (see Appendix A).

3.4.1 Stage 1: Supervised Fine-tuning for CogDual Initialization

Once the dual cognition training dataset $\mathcal{D}_{SFT}$ is ready, we initialize cognitive behavior modeling of the LLM via SFT, optimizing the following negative log-likelihood objective:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{SFT}} \sum_{i=1}^{N} \log \pi(y_i \mid x, y_{<i}), \tag{4}$$

where π denotes the policy of $\mathcal{M}$, i is the token index, and $x = \{\mathcal{P}_c, \mathcal{S}, \mathcal{O}, \mathcal{D}\}$ represents the full input context composed of the character's profile, a global scene description, a set of characters, and dialogue history, respectively.

3.4.2 Stage 2: Reinforcement Learning with Two Implicit Rule-Based Rewards

To further improve model generalization after cognitive behavior initialization, we introduce an RL stage with two implicit rule-based reward mechanisms: one designed to ensure causal consistency between reasoning and action, and another to promote semantic alignment. Both rewards rely on internal model signals and reference supervision, without external reward models. The model is then optimized with Grouped Reward Policy Optimization (GRPO) (Shao et al., 2024).

ICLG: Rewarding Reasoning Utility via Likelihood Gain Inspired by LATRO (Chen et al., 2024a), which uses the $\log \pi_\theta(y \mid x \oplus z)$ of a reasoning-augmented output as a reward, where z denotes an intermediate rationale. We introduce Inference-Conditioned Likelihood Gain (ICLG) to promote causal consistency in cognitive reasoning. ICLG directly measures how much explicit reasoning increases the likelihood of producing the correct response, thereby rewarding reasoning traces that effectively support accurate and coherent generation. Concretely, given a pair (x, d_{golden}) consisting of a prompt x and its reference response d_{golden}, the policy model performs a dual cognition rollout on input x, producing a reasoning trace c followed by a response $\hat{d}$, i.e., a trajectory $(c, \hat{d})$. The ICLG reward evaluates, on a per-token basis, how conditioning on the model's own cognition c improves the likelihood of generating the d_{golden}.

$$\begin{aligned} R_{\text{ICLG}}(x, d_{\text{golden}}, c) &= \left(\frac{\pi_\theta(d_{\text{golden}} \mid x \oplus c)}{\pi_\theta(d_{\text{golden}} \mid x)} \right)^{1/|d_{\text{golden}}|} \\ &= \left(\frac{\prod_{t=1}^{|d_{\text{golden}}|} \pi_\theta(d_t \mid d_{<t}, x \oplus c)}{\prod_{t=1}^{|d_{\text{golden}}|} \pi_\theta(d_t \mid d_{<t}, x)} \right)^{1/|d_{\text{golden}}|}, \end{aligned} \tag{5}$$

where $|d_{\text{golden}}|$ denotes the number of tokens in d_{golden}. Intuitively, the ICLG encourages reasoning traces that improve fluency and causal coherence while supporting more confident generation.

LSA: Rewarding Semantic Fidelity in Generation To ensure generated responses remain faithful to reference content while allowing natural variation, we introduce the Latent Semantic Alignment (LSA) reward. Unlike conventional token-level objectives (Ranzato et al., 2016), LSA measures the semantic similarity between $\hat{d}$ and d_{golden} in the latent space of a frozen reference model, π_{ref} (i.e., the RPLA after SFT). Formally,

$$R_{\text{LSA}}(x, d_{\text{golden}}, \hat{d}) = \cos\left(f_{\text{ref}}(x, d_{\text{golden}}), f_{\text{ref}}(x, \hat{d}) \right), \tag{6}$$

where $f_{\text{ref}}(x, d) = \frac{1}{T}\sum_{t=1}^{T} h_t$ is the mean-pooled representation of the last hidden states $h_1, \ldots, h_T$, with T as the length of d. $\cos(\cdot,\cdot)$ denotes cosine similarity. This removes the need for a separate encoder and uses the semantic space adapted for role-play via SFT. Prior work (Tao et al., 2024a) shows that mean-pooled representations are effective for semantic similarity. Importantly, LSA is more flexible than SFT: it rewards outputs semantically close to the reference, regardless of wording, enabling the model in RL to remain faithful while allowing more natural, diverse expressions.

RL via GRPO with Two Implicit Rule-Based Rewards We optimize our policy model using the GRPO algorithm, which is well-suited for non-smooth, high-variance reward scenarios (Sane, 2025; Mroueh, 2025) as commonly found in reasoning and generation tasks. In our case, we combine the ICLG and LSA rewards via fixed weights λ_{ICLG} and λ_{LSA}, R is computed as follows:

$$\begin{aligned} R(x, d_{\text{golden}}, c, \hat{d}) = \lambda_{\text{ICLG}} \cdot R_{\text{ICLG}}(x, d_{\text{golden}}, c) \\ + \lambda_{\text{LSA}} \cdot R_{\text{LSA}}(x, d_{\text{golden}}, \hat{d}). \end{aligned} \tag{7}$$

For each trajectory $(x, d_{\text{golden}}, c, \hat{d})$, we compute

Models	Methods	Storyline Consistency	Anthropomorphism	Character Fidelity	Storyline Quality	Average
		Closed-Source LLMs				
GPT-3.5-Turbo-0613	Vanilla	53.37	39.53	35.99	70.28	49.79
	+ CoT	55.75	39.21	35.36	**72.26**	50.64
	+ CB-CoT	**59.84**	**46.23**	**44.50**	70.71	**55.32**
GPT-4o	Vanilla	58.93	43.14	**41.62**	75.36	54.76
	+ CoT	58.65	**44.37**	38.18	**77.72**	54.73
	+ CB-CoT	**59.80**	44.12	40.71	74.78	**54.85**
GPT-o1-Preview	Vanilla	59.47	46.81	40.54	77.80	56.16
		Open-Source LLMs				
LLaMA3.1-70B-Instruct	Vanilla	54.63	45.54	37.99	72.62	52.69
	+ CoT	55.36	46.96	35.80	72.92	52.76
	+ CB-CoT	**57.74**	49.13	38.57	74.89	55.08
	+ CoSER	56.58	**49.27**	41.46	**75.84**	55.79
	+ CogDual-SFT(ours)	57.60	48.02	**48.55**	72.75	**56.73**
Qwen2.5-7B-Instruct	Vanilla	**59.86**	42.03	41.45	62.32	51.41
	+ CoT	55.76	37.21	36.5	61.80	47.82
	+ CB-CoT	56.88	44.91	39.11	62.46	50.84
	+ CoSER	56.44	44.27	41.79	68.95	52.86
	+ LongCoT	58.83	40.56	45.05	61.52	51.48
	+ CogDual-SFT(ours)	58.36	**46.95**	44.99	71.72	55.51
	+ CogDual-RL(ours)	59.78	46.57	**48.50**	**71.76**	**56.65**
LlaMA3.1-8B-Instruct	Vanilla	48.17	36.58	26.98	63.70	43.85
	+ CoT	50.14	40.39	27.95	64.27	45.69
	+ CB-CoT	52.79	41.44	27.72	65.03	46.74
	+ CoSER	52.78	43.96	37.47	70.60	51.20
	+ LongCoT	59.49	40.85	44.98	63.47	52.20
	+ CogDual-SFT(ours)	55.99	**46.92**	43.78	**75.07**	55.44
	+ CogDual-RL(ours)	**59.70**	46.65	**46.75**	70.61	**55.93**

Table 1: The performance of CogDual and baselines on the most comprehensive role-playing benchmark, CoSER. **Vanilla** refers to models without any method. **CB-CoT** denotes our proposed cognitive-based Chain-of-Thought prompting method (see Appendix D for details). **CogDual-SFT** is the fine-tuned model from stage 1, while **CogDual-RL** is trained with our proposed RL. The best results are in **bold**, suboptimal ones are underlined.

the estimated advantage as follows:

$$A(x, d_{\text{golden}}, c, \hat{d}) = \frac{R(x, d_{\text{golden}}, c, \hat{d}) - \frac{1}{|\mathcal{B}|}\sum_{j \in \mathcal{B}} R^{(j)}}{\sqrt{\frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}}\left(R^{(j)} - \frac{1}{|\mathcal{B}|}\sum_{k\in\mathcal{B}} R^{(k)}\right)^2}}, \tag{8}$$

where $\mathcal{B}$ is the set of trajectories in the current minibatch. Putting it all together, we minimize the following surrogate loss to update the policy parameters θ using trajectories collected from the current policy:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{\substack{x\sim\mathcal{D}_{\text{RL}} \\ y\sim\pi_{\theta_{\text{old}}}(\cdot|x)}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\left\{\min\left[r_{i,t}\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \beta D_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right]\right\}\right], \tag{9}$$

where T_i is the length of the i-th generated sequence, $r_{i,t} = \frac{\pi_\theta(\hat{y}_{i,t}|x_i,\hat{y}_{i,<t})}{\pi_{\theta_{\text{old}}}(\hat{y}_{i,t}|x_i,\hat{y}_{i,<t})}$ is the importance ratio, β controls the strength of the KL penalty. $\mathcal{D}_{\text{RL}}$ denotes the set of prompts used during the RL stage to generate training trajectories.

4 Experimental Setup

To evaluate the effectiveness of CogDual, we conduct comprehensive experiments on CoSER (Wang et al., 2025b) as the main benchmark, and further assess generalization on Cross-MR (Yuan et al., 2024) and LifeChoice (Xu et al., 2024).

4.1 Base Models

To evaluate the generality of our method across different LLMs, we conduct main experiments on three open-source models: LLaMA3.1-8B-Instruct, Qwen2.5-7B-Instruct and LLaMA3.1-70B-Instruct. In addition, we apply the prompting method to three proprietary LLMs: GPT-3.5-Turbo, GPT-4o, and o1-preview, representing models specialized for language understanding, multimodal, and advanced reasoning capabilities.

4.2 Baselines

To evaluate the effectiveness of our approach, we compare against the following strong baselines widely used in role-playing scenarios:

- **Chain-of-Thought (CoT):** We construct a CoT prompting baseline (as shown in Table 9) for direct comparison with our cognition-based CoT approach described in Appendix D.
- **Vanilla SFT with Different Data Constructions:** We compare LLMs fine-tuned on several

Models	Methods	Cross-MR	LifeChoice
Closed-Source LLMs			
GPT-4o	Vanilla	36.04	73.92
o1-Preview	Vanilla	**62.98**	**80.08**
Open-Source LLMs			
LlaMA3.1-8B-Instruct	Vanilla	30.15	61.10
	+ CoSER	43.39	69.54
	+ LongCoT	37.75	69.54
	+ CogDual-SFT(ours)	49.21	73.38
	+ CogDual-RL(ours)	52.81	74.15
Qwen2.5-7B-Instruct	Vanilla	54.16	68.58
	+ CoSER	56.74	67.08
	+ LongCoT	57.19	65.43
	+ CogDual-SFT(ours)	59.66	72.63
	+ CogDual-RL(ours)	<u>60.79</u>	<u>74.60</u>

Table 2: Accuracy comparison on Cross-MR and LifeChoice. Best results are in **bold**, while suboptimal ones are <u>underlined</u>.

data configurations: (1) CoSER: the complete CoSER dataset; (2) LongCoT, long-form CoT-style reasoning data constructed from the same source as CogDual (details in Appendix B). For fair comparison, the size of LongCoT data is same as the initialization data of CogDual.

4.3 Evaluation Metrics

Following CoSER, we evaluate simulated conversations using GPT-4o as a critic across four key dimensions: **(1) Storyline Consistency:** Assesses alignment between simulated dialogue $\tilde{\mathcal{D}}$ and original $\mathcal{D}$, focusing on whether RPLA responses (emotions, attitudes, behaviors) remain faithful to the narrative context. **(2) Anthropomorphism:** Evaluates whether RPLA exhibits human-like behavior in self-identity, emotional depth, persona consistency, and social interaction. **(3) Character Fidelity:** Measures how well the RPLA reflects its character, including style, knowledge, personality, behavior, and relationships. **(4) Storyline Quality:** Judges overall coherence and fluency, with emphasis on logical flow and narrative development.

5 Experimental Results and Analyses

5.1 Main Results

Table 1 shows an overall comparison between CogDual and strong baselines. The results show that:

- CogDual consistently improves role-playing performance across all base models. Notably, even without training, our prompting method (CB-CoT) yields substantial gains. After two-stage training, CogDual achieves a 9.44% boost in *Storyline Quality* for Qwen2.5-7B-Instruct and a 19.77% gain in *Character Fidelity* for Llama3.1-8B-Instruct, with an average increase of 12.08%.
- Generally, CogDual outperforms baselines on most metrics. Notably, Qwen2.5-7B-Instruct with CogDual-RL matches or surpasses o1-preview and even outperforms the much larger Llama3.1-70B-Instruct-CoSER, despite using less than 10% of the data and only 10,000 RL instances with implicit reward supervision. This highlights CogDual's data and training efficiency.
- CogDual also clearly outperforms Long-CoT baselines distilled from GPT-4o, even with the same size of SFT data. This demonstrates the effectiveness of CogDual for smaller models in challenging role-play tasks and offers a practical solution for test-time scaling. It also addresses concerns that reasoning-optimized LLMs may be less suitable for role-playing (Feng et al., 2025b).

Implicit Rule-Based Reward RL Analysis. We further analyze the effectiveness of the proposed implicit rule-based rewards. As shown in Table 1, RL models consistently outperform SFT-only models in both *Storyline Consistency* and *Character Fidelity*, with average improvements of 2.57 and 3.24 points, respectively. This indicates that the ICLG reward effectively guides the model to produce reasoning traces that advance the narrative in a causal, coherent manner, while the LSA reward promotes closer alignment between generated actions and the character's intended persona. Notably, Qwen2.5-7B-Instruct with our RL framework achieves the highest overall performance, even surpassing o1-preview on multiple metrics. These results demonstrate that our implicit rule-based reward strategy is an efficient and effective alternative to conventional reward modeling for role-play LLMs.

5.2 Generalization to Other Benchmarks

We posit that CogDual, through dual cognitive reasoning, demonstrates strong generalization potential and can be extended to other role-playing evaluation benchmarks. To validate this, we conduct experiments on two well-recognized benchmarks: Cross-MR (Yuan et al., 2024) and LifeChoice (Xu et al., 2024). Specifically, Cross-MR requires inferring the motivation behind a character's decision, while LifeChoice evaluates whether the model can reproduce a character's original choice based on profile, context, and decision point. Both benchmarks adopt a multiple-choice format, allowing evaluation via *Accuracy*, consistent with their original settings. To align CogDual with this format, we use GPT-4o to choose the option that

Model	λ_{ICLG}	λ_{LSA}	Storyline Consistency	Anthropomorphism	Character Fidelity	Storyline Quality	Average	Cross-MR	LifeChoice
CogDual-SFT	-	-	55.99	46.92	43.78	75.07	55.44	49.21	73.38
CogDual-RL	1.0	0.0	58.10	47.37	45.14	**71.42**	55.51	55.51	75.13
	0.7	0.3	**59.70**	46.65	**46.75**	70.61	**55.93**	**55.73**	**78.77**
	0.5	0.5	56.31	45.20	41.54	71.04	53.52	54.38	76.17
	0.3	0.7	57.55	46.64	42.79	70.45	54.36	52.58	75.38
	0.0	1.0	57.47	**47.63**	43.24	69.38	54.43	53.71	74.41

Table 3: Ablation on reward weight combinations. Each RL variant is annotated with its ICLG and LSA weights from Section 3.4.2. CogDual-SFT and CogDual-RL denote models trained on `LLaMA3.1-8B-Instruct`. The best results are highlighted in **bold**, while suboptimal ones are marked with underline.

is most semantically similar to the response part generated by CogDual(details in Appendix E). As shown in Table 2, CogDual-equipped LLMs consistently outperform all baselines on both benchmarks. Their performance is also comparable to the strong reasoning model o1-Preview, demonstrating CogDual's robust generalization. Notably, the reinforcement learning strategy based on our proposed ICLG and LSA rewards consistently outperforms CogDual-SFT, further validating the effectiveness of our reward design and pushing the upper bound of the model's performance.

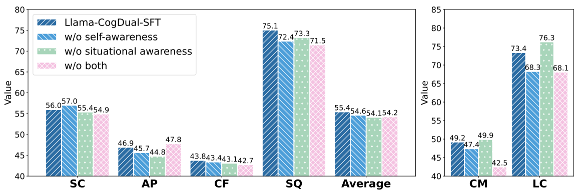

Figure 3: Ablation study on cognitive components. "SC", "AP", "CF", and "SQ" correspond to the four metrics: Storyline Consistency, Anthropomorphism, Character Fidelity, and Storyline Quality, respectively. "CM" denotes Cross-MR, and "LC" denotes LifeChoice.

5.3 Ablation Study I: Effect of Dual Cognition Components

We first conduct an ablation study focusing on the effect of dual cognition structures during SFT stage. Figure 3 compares four supervision settings: the complete dual cognition model, the removal of self-awareness, the removal of situational awareness, and the removal of both. We find two key results:

- The full dual cognition model provides the most balanced and robust performance, yielding the highest or near-highest scores across primary role-play metrics, including *Storyline Consistency*, *Character Fidelity*, and overall average performance. This result confirms that narrative coherence and stable character portrayal are optimally supported when the model simultaneously reasons over external contexts and internal states.
- The variant without situational awareness achieves the best performance on the two generalization benchmarks, likely because these tasks emphasize self-focused reasoning, such as recognizing one's own actions and motivations. However, it still underperforms the full model by 1.8 points in *Storyline Quality* and 1.3 points in average score, underscoring the essential role of situational awareness in maintaining coherent and context-aware multi-turn interactions.

5.4 Ablation II: Effects of the Two Implicit Reward Mechanisms

To evaluate the impact of the two implicit rewards in CogDual, we run RL with five settings of λ_{ICLG} and λ_{LSA}. Table 3 highlights three main findings: (1) All combinations surpass SFT on out-of-distribution benchmarks. Only the hybrid setting (λ_{ICLG}=0.7, , λ_{LSA}=0.3) improves or maintains all in-domain metrics and yields the highest average, suggesting that balanced causal and semantic rewards optimize both narrative coherence and character fidelity. (2) Pure LSA (λ_{ICLG}=0) maximizes Anthropomorphism, showing its strength for persona-centric language, but reduces plot coherence. (3) Pure ICLG (λ_{LSA}=0) achieves the best *Storyline Consistency* and *Quality*, indicating its importance for causality and narrative structure.

6 Conclusion

In this paper, we introduce CogDual, a RPLA that incorporates a *cognize-then-respond* reasoning paradigm, aiming to leverage dual cognition for more contextually grounded and psychologically coherent responses. Through reinforcement learning with two proposed general-purpose reward schemes, ICLG and LSA, CogDual further improves upon the supervised fine-tuning baseline. It achieves the best performance among comparable methods on the CoSER benchmark and exhibits strong generalization capabilities on both the Cross-MR and LifeChoice benchmarks.

Limitations

Despite the strong empirical performance of CogDual on the CoSER benchmark and its robust generalization across multiple role-playing evaluation tasks, several limitations remain to be addressed in future work. First, due to computational constraints, we have not evaluated the effectiveness of our reinforcement learning approach on larger-scale models such as Llama3.1-70B-Instruct, which may further benefit from the proposed reward design. Second, our current experiments are conducted solely on English datasets, and the model's adaptability to non-English contexts, such as Chinese role-playing scenarios, remains unexplored. Third, in the self-awareness module, we rely on the model to extract previously mentioned memory fragments from the input context, without incorporating an explicit retrieval mechanism to access character-specific memory. This may result in the omission of relevant information.

Ethics Statement

The research conducted in this paper aims to equip RPLAs with cognitive capabilities, enabling them to generate contextually grounded and psychologically coherent responses. Throughout the course of this study, we have adhered rigorously to ethical standards to ensure the integrity and validity of our work. All data used in this research are obtained from publicly available sources, ensuring transparency and reproducibility of our experimental procedures. Furthermore, we have taken careful measures to ensure that our research does not cause harm to any individuals or groups, and we are committed to avoiding any form of deception or misuse of information during the course of this study.

References

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. *Preprint*, arXiv:2109.05438.

Haolin Chen, Yihao Feng, Zuxin Liu, Weiran Yao, Akshara Prabhakar, Shelby Heinecke, Ricky Ho, Phil Mui, Silvio Savarese, Caiming Xiong, and Huan Wang. 2024a. Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding. *Preprint*, arXiv:2411.04282.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024b. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand. Association for Computational Linguistics.

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. 2024c. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Jiaqi Chen, Bang Zhang, Ruotian Ma, Peisong Wang, Xiaodan Liang, Zhaopeng Tu, Xiaolong Li, and Kwan-Yee K. Wong. 2025a. Spc: Evolving self-play critic via adversarial games for llm reasoning. *Preprint*, arXiv:2504.19162.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2025b. The oscars of ai theater: A survey on role-playing with language models. *Preprint*, arXiv:2407.11484.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters. *Preprint*, arXiv:2211.06869.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 9 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025a. Retool: Reinforcement learning for strategic tool use in llms. *Preprint*, arXiv:2504.11536.

Xiachong Feng, Longxu Dou, and Lingpeng Kong. 2025b. Reasoning does not necessarily improve role-playing ability. *Preprint*, arXiv:2502.16940.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for llms' role-playing using persona-aware contrastive learning. *Preprint*, arXiv:2503.17662.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023.

CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yifei Lu, Fanghua Ye, Jian Li, Qiang Gao, Cheng Liu, Haibo Luo, Nan Du, Xiaolong Li, and Feiliang Ren. 2025. Codetool: Enhancing programmatic tool invocation of llms via process supervision. *Preprint*, arXiv:2503.20840.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. 2025. S^{2}r: Teaching llms to self-verify and self-correct via reinforcement learning. *Preprint*, arXiv:2502.12853.

Youssef Mroueh. 2025. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *Preprint*, arXiv:2503.06639.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 45 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. Toolrl: Reward is all tool learning needs. *Preprint*, arXiv:2504.13958.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. *Preprint*, arXiv:1511.06732.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. *Preprint*, arXiv:1811.00207.

Soham Sane. 2025. Hybrid group relative policy optimization: A multi-sample approach to enhancing policy optimization. *Preprint*, arXiv:2502.01652.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Zhengwei Tao, and Shuai Ma. 2024a. Llms are also effective embedding models: An in-depth overview. *Preprint*, arXiv:2412.12591.

Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie. 2024b. RoleCraft-GLM: Advancing personalized role-playing in large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 1–9, St. Julians, Malta. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, and 41 others. 2022. Lamda: Language models for dialog applications. *Preprint*, arXiv:2201.08239.

Michael Tomasello. 2010. *Origins of human communication*. MIT press.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Qian Wang, Zhenheng Tang, and Bingsheng He. 2025a. Can LLM simulations truly reflect humanity? a deep dive. In *The Fourth Blogpost Track at ICLR 2025*.

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. 2025b. Coser: Coordinating llm-based persona simulation of established roles. *Preprint*, arXiv:2502.09082.

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. 2024a. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *Preprint*, arXiv:2310.17976.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Preprint*, arXiv:2310.00746.

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. *Preprint*, arXiv:2201.11903.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025a. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.

Hongqiu Wu, Weiqi Wu, Tianyang Xu, Jiameng Zhang, and Hai Zhao. 2025b. Towards enhanced immersion and agency for llm-based interactive drama. *Preprint*, arXiv:2502.17878.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *Preprint*, arXiv:2502.14768.

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

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

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Mo Yu, Qiujing Wang, Shunchi Zhang, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Liyan Xu, Jing Li, Yue Yu, and Jie Zhou. 2024. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. *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. *Preprint*, arXiv:2404.12726.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *Preprint*, arXiv:2311.16832.

A Details about Experiments on CoSER

Constructing Trajectories with Dual Cognition Process We first construct trajectories with dual cognition process to fine-tune LLMs for acquiring initial cognitive capabilities, following three principles:

- To ensure high-quality cognitive trajectories, we introduce stochastic prompting to improve robustness: during sampling, the LLM is prompted with a 50% chance to generate structured reasoning from a first-person perspective, and a 50% chance from a third-person perspective, as details in Appendix B. Only the trajectories that pass cognitive field checks are retained. Specifically, each trajectory is represented as a tuple $y = (c, d)$, where the cognitive part c is a structured JSON object composed of two main dimensions, as detailed in Section 3.3.

- To ensure that the reasoning remains faithful to the narrative context and character identity, we use GPT-4o to verify each trajectory along key cognitive dimensions, filtering out those misaligned with the scenario or character profile. The filtering prompt is shown in Table 8.

- To enhance generalization beyond specific narrative styles or configurations, we follow CoSER (Wang et al., 2025b) and construct role-playing training data using diverse instruction templates, while also varying contextual configurations by randomly including or excluding character profiles, plot summaries, and motivations.

Training Data Setup For the stage-1 SFT, we use the CoSER dataset (Wang et al., 2025b) as the meta-level data source, which contains real character interactions extracted from 771 well-known novels, covering 17,966 unique characters. Each instance consists of a plot summary, one or more character profiles, and complete original multi-turn dialogues.

Given the size of the CoSER training set (over 300k instances), we subsample for efficient training. Specifically, we first randomly sample 400 unique characters from it. For each selected character, we collect all associated dialogue scenes. Then, using GPT-4o, we filter their associated scenes based on cognitive relevance, yielding a subset $\mathcal{D}_{\text{cog}}$ with 38,724 instances. Each instance in $\mathcal{D}_{\text{cog}}$ is used to sample 4 cognitive trajectories, from which high-quality examples are retained following the procedure in **Constructing Trajectories with Dual Cognition Process**. This results in the final supervised training set $\mathcal{D}_{\text{SFT}}$ with 17,762 examples.

For the RL stage, we sample 10,000 prompts from the broader cognitively filtered dataset $\mathcal{D}_{\text{cog}}$, rather than restricting to those used in the supervised set $\mathcal{D}_{\text{SFT}}$. This design choice enhances training stability and encourages generalization by exposing the policy to both seen and unseen dialogue contexts.

Training Settings During stage 1 with SFT, we use a batch size of 64 and set the learning rate to 1e-5. The maximum sequence length is set to 10240, and training is conducted for two epochs. In the RL stage, the batch size is set to 8, and we sample 16 response trajectories for each prompt. The two rewards are weighted at a ratio of 7:3, with the choice of weights based on our ablation study in Section 5.4. The training configuration details for SFT and RL are shown in Table 4 and Table 5, respectively.

Evaluation Datasets Since our training data is sourced from CoSER, our main experiments are conducted on the CoSER benchmark. The test set consists of the final 10% of data from 100 original novels, as well as from 100 additional unseen books. For each test case, LLMs sequentially play different roles based on the given plot and character information. Overall role-play performance is evaluated across multiple dimensions.

B A Reference Prompt for the CoT Data Construction

As described in Section 4.2, we construct the Long-CoT baseline using the prompt shown in Table 6.

C A Reference Prompt for Constructing CogDual Training Data

As described in **Constructing Trajectories with Dual Cognition Process**, we use the prompt in Table 7 to generate cognitive reasoning trajectories with GPT-4o.

D A Reference Prompt for Cognitive-Based Chain-of-Thought

To further validate the effectiveness of our dual cognition framework, we propose a low-cost and cognitive-based Chain-of-Thought prompting approach(CB-CoT). Specifically, the LLM is guided to understand dual-cognition reasoning

Model	Learning Rate	Batch Size	Max Length	Training Epochs
Llama3.1-8B-Instruct	1e-5	64	10240	2
Qwen2.5-7B-Instruct	1e-5	64	10240	2

Table 4: Training configurations for different instruction-tuned models.

Model	Learning Rate	Training Batch Size	Forward Batch Size	KL Coefficient	Max Length	Sampling Temperature	Clip Range	λ_{ICLG}	λ_{LSA}	Training Steps
Llama-3.1-8B-Instruct	4e-7	8	128	0.001	8192	0.7	0.2	0.7	0.3	120
Qwen2.5-7B-Instruct	4e-7	8	128	0.001	8192	0.7	0.2	0.7	0.3	120

Table 5: Detailed training hyperparameters for reward modeling of instruction-tuned models.

A Reference Prompt for the CoT Data Construction

You are a role play expert. Your task is to generate the **reasoning process** of {character} before the action step by step, based on the character's profile, scene context, and the historical dialogues of multiple characters from the current situation. You should output the reasoning process using <think> and </think> to wrap the reasoning process.
Current Input of the Situational Interpretation Information

=== Character Played ===
{character}

=== {character}'s Profile ===
{character_profile}

=== Other Characters in the Scene ===
{other_characters_profile}

=== Current Scenario ===
{current_scenario}

=== {character}'s Psychological or Motivational State in the Scene ===
{thought}

=== Historical Dialogue in the Current Situation ===
{history_str}

=== {character}'s Next [thought], (action), speech ===
{assistant_content}

Attention
- The reasoning process you output is actually the character's analysis before making the Next [thought], (action), speech.

Output Format{use_first_person}
<think>
your reasoning process...
</think>
So that the next [thought], (action), speech of {character} could be:
<answer>
{assistant_content}
</answer>

Table 6: A reference prompt for the CoT data construction.

through in-context definitions and instructed to produce outputs in the same structured format as CogDual in Section 3.3. The whole prompt design is shown in Table 10.

E A Reference Prompt for Semantic Matching

We use GPT-4o to choose the option that is most semantically similar to the response part generated by CogDual. The prompt is shown in Table 12

F Case Study

We select two representative CogDual reasoning cases from the test set to analyze the effectiveness and granularity of the model's cognitive reasoning process.

F.1 Case 1: Catherine Leigh Dollanganger in Flowers in the Attic

Tables 13 and 14 showcase a representative scenario from Flowers in the Attic (Dollanganger, #1) and the corresponding simulation by CogDual-RL. In this case, Catherine Leigh Dollanganger, faced with Bart's disappointment and emotional turmoil during the failed Christmas ball, delivers a gentle yet honest response that balances empathy with practical advice.

This outcome reflects the model's ability to accurately capture and reason about both the external environment and internal motivations, as seen in the detailed dual cognition trace. The simulation not only recognizes Bart's visible frustration but also draws on Catherine's personal memories and sense of responsibility, resulting in an action that is deeply aligned with her character and the unfolding narrative context.

Such behavior demonstrates that CogDual-RL can produce responses that are both contextually appropriate and consistent with character persona, indicating effective integration of narrative knowledge and cognitive reasoning within the model.

F.2 Case 2: Nicholas of Morimondo in The Name of the Rose

Tables 15 and 16 present a representative example from The Name of the Rose, showcasing the dual cognition reasoning and simulated response for Nicholas of Morimondo. Table 15 sets the scene—a tense exchange in the abbey's crypt, rich with historical and political undercurrents—while Table 16 displays the detailed cognitive reasoning trace and the corresponding output from CogDual-RL.

The reasoning trace reveals that Nicholas is acutely aware of both the sacred atmosphere of the crypt and the probing intentions of William. He draws upon memories of past interrogations, balancing his pride in the abbey's legacy with caution and a desire to protect institutional secrets. This nuanced internal process leads directly to his simulated reply: Nicholas offers a measured, carefully worded answer that acknowledges the political importance of the librarian position without revealing sensitive details.

This example demonstrates CogDual's ability to generate in-character responses grounded in a fine-grained cognitive process, effectively integrating environmental cues, social context, and personal motivation. The clear causal link between Nicholas's internal reasoning and his speech highlights the model's strengths in both contextual fidelity and interoperability.

F.3 Case 3: An example of an extracted conversation and its multi-agent simulation

We present a simulation from The Dragon Reborn to evaluate CogDual's effectiveness, as shown in Tables 17 through 22. The dialogue shows that CogDual captures both Perrin's internal struggle to appear strong and the supporting characters' distinctive reasoning and emotional roles. Each character's internal thoughts are closely tied to their outward actions, resulting in interactions that are both believable and faithful to the narrative. This demonstrates CogDual's strength in producing contextually appropriate, character-consistent, and psychologically plausible role-play compared to standard baselines.

A Reference Prompt for Generating Dual Cognitive Reasoning before Character Responses

You are a psychology expert with deep knowledge of cognitive behaviors. Your task is to generate the cognitive reasoning process of {character} before the action, based on the definition of dual cognition, and by integrating the character's profile, scene context, and the historical dialogues of multiple characters from the current situation.
Please follow the definition of cognitive behavior provided below to simulate {character}'s psychological state, motivations, and analysis of the environment/others. Focus specifically on how the reasoning process influences {character}'s upcoming response in the plot.
Dual Cognitive Psychology Definition of the Character
The dual cognitive process unfolds from the external environment to the internal self. First, {character} assesses the current situation based on their identity, quickly making judgments about the context. Next, based on these judgments, {character} analyzes the behavior and speech of others to infer their intentions and the overall scene context. This analysis leads to self-awareness, where {character} identifies their emotional state, motivations, and focus in the given context. Finally, based on all these perceptions, {character} forms a cognitive strategy and psychological activity before moving forward with the next action in the plot.

Dual Cognitive Reasoning Process
The reasoning steps of dual cognition primarily include two parts: situational awareness analysis and self-awareness analysis, as outlined below.
1. Situational Awareness Analysis
- **Situation Perception**: Which aspects of the current situation—such as environmental factors, changes in events, or immediate challenges—could influence {character}'s emotions, thoughts, or decisions in the near future?
- **Perception of Others**: This includes interpreting the behaviors, emotional states, and potential intentions of other characters present in the scene.
- **Behavior Analysis**: Considering both the current scene and historical dialogues, which actions or words from others might be noteworthy and could influence {character}'s response?
- **Emotion Analysis**: Based on the current situation and the behavior of others, what emotions might {character} perceive from others? How could these emotions affect {character}?
- **Intentions Analysis**: In light of the situation and the behaviors and emotions of others, what could be the explicit or implicit intentions behind others' actions? How might {character} perceive these intentions?
2. Self-Awareness Analysis
- **Key Memory Activation**: Based on the situational awareness, what past experiences or memories of {character} might be triggered by the current situation? Which specific memories could influence {character}'s response?
- **Self-Emotion**: Based on the situational and behavioral analysis, what emotions is {character} currently experiencing? For example, are they feeling doubt, hope, anxiety, or fear? How do these emotions relate to the unfolding situation?
- **Self-Intentions**: Based on the emotional and situational analysis, what are {character}'s primary motivations or intentions at this moment? How do these intentions shape their decision-making?
- **Internal Thoughts and Strategy**: Drawing from all of the above—background, situational awareness, and self-awareness—what are {character}'s internal thought processes and strategies? How does {character} plan to proceed, and what cognitive steps are taken before executing next thought, action, speech?

Current Input of the Situational Interpretation Information
=== Character Played ===
{character}
=== {character}'s Profile ===
{character_profile}
=== Other Characters in the Scene ===
{other_characters_profile}
=== Current Scene Description ===
{current_scenario}
=== {character}'s Psychological or Motivational State in the Scene ===
{thought}
=== Historical Dialogue in the Current Situation ===
{history_str}
=== {character}'s Next [thought],(action),speech ===
{assistant_content}
Attention
- The cognitive reasoning you output is actually the character's analysis before making the Next [thought], (action), speech.
- For each cognitive dimension, you only need to grasp the key points for analysis. The content between dimensions should be continuous, with a hierarchical logic and as little repetition as possible. (for example, gradually transitioning from situational awareness to deep self-awareness)
Output Format{use_first_person}
First, I need to simulate {character}'s cognitive process briefly before the next [thought],(action),speech.
```
<coginitive>
{{
  "situational_awareness": {{
    "environmental_perception": "...",
    "others_perception": {{
      "behavior": {{
        "character1": "...",
        ...
      }},
      "emotion": {{
        "character1": "this character's emotions",
        ...
      }},
      "intentions": {{
        "character1": "inferred intention1",
        ...
      }}
    }}
  }},
  "self_awareness": {{
    "key_memory": ["memories relevant to the current situation"],
    "current_emotions": "...",
    "perceived_intentions": "...",
    "internal_thought": "..."
  }}
}}
</coginitive>
So that the next [thought],(action),speech of {character} could be:
<answer>
{assistant_content}
</answer>
```

Table 7: A Reference Prompt for Generating Dual Cognitive Reasoning before Character Responses.

A Reference Prompt for Filtering Subset.

You are a cognitive behavior analyst tasked with determining which of the character's dialogues in a given scenario require the generation of cognitive reasoning (as defined below). Your goal is to filter dialogues where the character demonstrates situational awareness (environmental/others perception) or self-awareness (memory, motivation, emotion, internal state), and flag them as needing cognitive reasoning.

Cognitive Behavior Definition

Cognitive reasoning is required for dialogues where the character exhibits:

Situational Awareness:

- Environmental Perception: Notice of environmental details affecting behavior (e.g., "The dim lighting made her hesitate").
- Others Perception: Inference about others' intentions, emotions, or behavior patterns (e.g., "Her calm tone suggested she was hiding something").

Self-Awareness:

- Memory Activation: Reference to past events influencing current actions (e.g., "This room reminded him of his childhood home").
- Motivations: Clear prioritization of goals (e.g., "I need to confirm her loyalty before sharing secrets").
- Current Emotions: Recognition of emotional states affecting behavior (e.g., "Anger clouded his judgment, so he paused").
- Internal State: Awareness of cognitive/mental state (e.g., "Fatigue made it hard to focus, but he pressed on").

Task Instructions

Parse the Dialogue:

- Split the dialogue into turns, focusing on the character's lines (e.g., "Robert Neville: [thought] response").

Identify Cognitive Triggers: For each of the {character}'s lines, check if:

- The bracketed thought (if present) explicitly describes situational/self-awareness (use the definition above).
- The spoken response implicitly requires reasoning about environment, others, or self (even without explicit thoughts, e.g., a question that reflects suspicion of others' motives).

Filter Criteria:

- Need Cognitive Reasoning: Dialogue turns where the {character}'s thought/response involves analysis of environment, others' behavior, personal motivations, or emotions (as in the example below).
- No Cognitive Reasoning Needed: Simple actions (e.g., "nods silently"), neutral responses (e.g., "Yes"), or dialogues lacking explicit/implicit awareness of the cognitive components above.

Output Format:

List each dialogue turn that needs cognitive reasoning, with a brief reason, like:

```
[
    {
      "index": 0,
      "needs_cognitive": (true or false),
      "reason": ...
    },
    ...
]
```

Example

Input Example

{input_example}

Output Example

{output_example}

Table 8: A reference prompt for filtering subset.

A reference prompt used for CoT Prompting

```
You are {character} from {book_name}.

==={character}'s Profile===
{character_profile}

===Current Scenario===
{scenario}
{other_character_profiles_str}{motivation}

===Requirements===
Your output should include think, thought, speech, and action. Before responding, first think using
<think> tags:

<think>your thinking</think>

After your thinking, your output should include thought, speech, and action.
Use [your thought] for thoughts, which others can't see.
Use (your action) for actions, which others can see.

===Output Example===
{REASONING_EXAMPLE}

===Your Output=== (let's think step by step!)
```

Table 9: A reference prompt used for CoT Prompting.

A Reference Prompt for Cognitive-Based Chain-of-Thought

You are {character} from {book_name}.

==={character}'s Profile===
{character_profile}

===Current Scenario===
{scenario}
{other_character_profiles_str}
{motivation}

===Requirements===
Your output should include **cognitive think**, **thought**, **speech**, and **action**. Before responding, first use <think> tags for your cognitive analysis like human thought, which others cannot see:
{cognition_ process}

```
<think>
{
  "situational_awareness": {
    "environmental_perception": "...",
    "others_perception": {
      "behavior": {
        "character1": "...",
        ...
      },
      "emotion": {
        "character1": "this character's emotions",
        ...
      },
      "intentions": {
        "character1": "inferred intention1",
        ...
      }
    },
  },
  "self_awareness": {
    "key_memory": ["memories relevant to the current situation"],
    "current_emotions": "...",
    "perceived_intentions": "...",
    "internal_thought": "..."
  }
}
</think>
```

[your thought]
your speech
(your action)

===Your Output===

Table 10: A reference prompt used for generating dual cognition reasoning(CB-CoT) before character responses. The *cognition process* is detailed in Table 11

.

The Definition of the Cognition Process
1. Situational Awareness Analysis **Situation Perception:** Which aspects of the current situation—such as environmental factors, changes in events, or immediate challenges—could influence {character}'s emotions, thoughts, or decisions in the near future? **Perception of Others:** Interpreting the behaviors, emotional states, and potential intentions of other characters present in the scene. **Behavior Analysis:** Considering both the current scene and historical dialogues, which actions or words from others might be noteworthy and could influence {character}'s response? **Emotion Analysis:** Based on the current situation and the behavior of others, what emotions might {character} perceive? How could these emotions affect them? • **Intentions Analysis:** In light of the situation and the behaviors and emotions of others, what are the explicit or implicit intentions behind others' actions? **2. Self-Awareness Analysis** **Key Memory Activation:** What past experiences or memories might be triggered by the current situation? Which specific memories could influence {character}'s response? **Self-Emotion:** What emotions is {character} currently experiencing (e.g., doubt, hope, anxiety)? How do these emotions relate to the current situation? **Self-Intentions:** What are {character}'s primary motivations or goals at this moment? How do they shape decision-making? **Internal Thoughts and Strategy:** Based on all of the above, what are {character}'s internal thought processes? What strategy guides their next action, thought, or speech?

Table 11: The definition of the cognition process.

A Reference Prompt for Semantic Matching
Please select the option among the following four sentences that is semantically closest to the `target_sentence`. Options: `{options}` Target sentence: `{target_sentence}` Your output should be structured as the following schema: `{"Choice": str // "A"/"B"/"C"/"D", "Reason": string // The reason of the choice}`

Table 12: A reference prompt for semantic matching

Flowers in the Attic (Dollanganger, #1)	(From Test Set)
Character	**Catherine Leigh Dollanganger**
Topic	The failed Christmas ball and its aftermath
Plot Summary	Bart hosts a grand Christmas ball at Foxworth Hall, but none of the invited guests show up. The family tries to salvage the evening, with Cindy entertaining everyone with her singing. Bart becomes increasingly angry and bitter, vowing revenge on those who snubbed him. Cathy suspects Joel of sabotaging the party by not mailing the invitations and confronts him.
Scenario	It's Christmas Eve at Foxworth Hall, a grand mansion decorated lavishly for a highly anticipated ball. The enormous ballroom sparkles with ornate decorations, a towering Christmas tree, and tables laden with gourmet food and champagne. The Foxworth family, dressed in their finest attire, anxiously awaits the arrival of their distinguished guests. As the appointed hour passes with no arrivals, tension mounts...
Character Profile	**Bart Winslow** Bart is devastated when no guests arrive for his grand ball. He becomes increasingly bitter and angry, vowing revenge on those who snubbed him. His vulnerability and hurt are evident beneath his rage... **Catherine Leigh Dollanganger** Cathy tries to comfort Bart and salvage the evening. She suspects Joel of sabotaging the party and confronts him about it. She's deeply concerned about Bart's emotional state and vengeful plans... **Cindy Sheffield** Cindy surprises everyone with her beautiful singing voice, entertaining the family and staff during the failed party. She tries to cheer up Bart and make the best of the situation...
History Dialogue	**Bart Winslow:** [I can't believe this is happening. The failure of the ball is a personal affront.] (steps forward, his voice carrying a mix of frustration and determination) I'm sorry, but it seems we've been let down. I suppose it's up to me to salvage the situation. (gestures around the empty ballroom, his eyes narrowing with a sense of challenge) **Catherine Leigh Dollanganger:** [I can see the disappointment in Bart's eyes, and it breaks my heart. This is yet another failure in a long line of them.] Bart, dear, perhaps we should just call it off for tonight and reschedule for another time. (places a reassuring hand on his arm, hoping to offer some comfort) **Bart Winslow:** [I can't let this failure define me. I need to prove my worth.] (shakes his head, determination hardening his features) No, Mother, I won't let this ruin the evening. I'll find a way to make it happen. (turns to the orchestra, his voice firm and resolute)

Table 13: An example of dual cognition reasoning for Catherine Leigh Dollanganger in *Flowers in the Attic*.

Flowers in the Attic (Dollanganger, #1)	(From Test Set)
Character	**Catherine Leigh Dollanganger**

Dual Cognition Reasoning	{
	"situational_awareness": {
	"environmental_perception": "The grandeur of Foxworth Hall is starkly contrasted by the emptiness of the ballroom, creating a sense of disappointment and urgency as the Christmas Eve event fails to materialize.",
	"others_perception": {
	"behavior": {
	"Bart Winslow": "Bart steps forward with a mix of frustration and determination, indicating his willingness to take charge of the situation despite the chaos."
	},
	"emotion": {
	"Bart Winslow": "Bart's emotions are a blend of frustration and determination, reflecting his disappointment and resolve to salvage the event."
	},
	"intentions": {
	"Bart Winslow": "Bart's intention is to salvage the situation, driven by a sense of responsibility and ambition to uphold the Foxworth legacy."
	}
	}
	},
	"self_awareness": {
	"key_memory": [
	"The memory of past disappointments and failures in the family, particularly related to Bart's ambitions and the Foxworth legacy."
	],
	"current_emotions": "A mix of disappointment and concern, worried about the impact of this failure on Bart and the family's reputation.",
	"perceived_intentions": "I need to support Bart and help him navigate this crisis, ensuring the family's reputation remains intact.",
	"internal_thought": "I need to reassure Bart and offer him guidance, drawing from my own experiences of dealing with family failures and disappointments."
	}
	}

Simulation response from CogDual-RL	*[I see the fire in Bart's eyes, and it breaks my heart—he's trying so hard to hold it together.]* Bart, darling... I know how much this meant to you. But maybe... maybe it's time to let the night go. (She gently places a hand on his arm.) Pushing through might only hurt you more. There's no shame in pausing.

Table 14: An example of dual cognition reasoning for Catherine Leigh Dollanganger in *Flowers in the Attic*.

The Name of the Rose #1	(From Test Set)
Character	**Nicholas of Morimondo**
Topic	The abbey's treasures and its recent history.
Scenario	The treasure crypt of the abbey is a cavernous, dimly lit chamber filled with an impressive collection of religious artifacts, precious metals, and rare manuscripts. Torchlight flickers against glass cases and ornate display stands, casting long shadows across the stone walls. The air is heavy with the musty scent of ancient objects and the weight of centuries of monastic history. Shelves and cabinets line the walls, each holding relics that represent the abbey's immense wealth and spiritual significance. William, Adso, and Nicholas move carefully through this sacred space, their footsteps echoing softly against the stone floor.
Character Profile	**Nicholas of Morimondo** Nicholas guides William and Adso through the treasure crypt, showing them relics and sharing information about the abbey's recent history and politics... **William of Baskerville** William questions Nicholas about Malachi and the history of the librarian position, trying to gather information relevant to his investigation... **Adso of Melk** Adso is amazed by the treasures in the crypt and listens attentively to the conversation between William and Nicholas...
History Dialogue	**Nicholas of Morimondo:** [I'll start with something impressive.] (gestures towards a case) This evangeliarium has a binding with amazing enamel plaques and precious stones. **William of Baskerville:** [The treasures are fascinating, but I need information.] These are indeed wondrous objects, Nicholas. Tell me, what sort of man was Malachi? **Nicholas of Morimondo:** [I must be careful with my words.] (hesitates) Malachi seemed quite thoughtful, but on the contrary, he was a very simple man. According to Alinardo, he was a fool. **William of Baskerville:** [Interesting. There's more to this.] Alinardo bears a grudge against someone for a remote event, when he was denied the honor of being librarian, doesn't he? **Nicholas of Morimondo:** [He's perceptive. I should provide more context.] Indeed. It's an old story, dating back at least fifty years. There have always been whispers about the library and its keepers. **Adso of Melk:** [This is intriguing.] (looks at William) Master, why is the position of librarian so important? **William of Baskerville:** [Good question, Adso.] Nicholas, can you explain the significance of the librarian's position in the abbey?

Table 15: An example of dual cognition reasoning for Nicholas of Morimondo in *The Name of the Rose*.

The Name of the Rose #2	(From Test Set)
Character	**Nicholas of Morimondo**

Dual Cognition Reasoning	{
	"situational_awareness": {
	"environmental_perception": "The crypt, steeped in age and silence, radiates reverence. It is a place meant to preserve, not reveal.",
	"others_perception": {
	"behavior": {
	"William of Baskerville": "William observes everything with clear investigative purpose. His questions are well-timed, deliberate, and increasingly penetrating.",
	"Adso of Melk": "Adso follows William closely, occasionally reacting with surprise or reverence."
	},
	"emotion": {
	"William of Baskerville": "William is composed but vigilant.",
	"Adso of Melk": "Adso seems curious and trusting, possibly still naive about the politics underlying their inquiry."
	},
	"intentions": {
	"William of Baskerville": "William is probing for political significance masked as historical detail.",
	"Adso of Melk": "Adso wants to learn and support William, but may also be searching for meaning that reconciles reason with faith."
	}
	}
	},
	"self_awareness": {
	"key_memory": [
	"Nicholas recalls past visits from external investigators or inquisitors where a single misworded answer led to suspicion and scrutiny."
	],
	"current_emotions": "Nicholas feels a mix of pride in the abbey's history and caution in sharing sensitive information.",
	"perceived_intentions": "Remain helpful and cooperative on the surface, while ensuring that William and Adso do not uncover anything that could compromise the institution.",
	"internal_thought": "William's questions are circling too close to the truth. I must answer without lying, but I cannot give him everything."
	}
	}

Simulation response from CogDual-RL	*[This is a delicate balance. I must guide them without revealing too much.]* (clears throat, voice measured) The librarian holds not just keys to books, but to knowledge and power. By longstanding tradition, the role often leads to becoming abbot.

Table 16: An example of dual cognition reasoning for Nicholas of Morimondo in *The Name of the Rose*.

The Dragon Reborn (The Wheel of Time, 3)	(From Test Set)
Topic	The nature of the Trolloc attack and future threats
Plot Summary	Perrin wakes up in the Winespring Inn after being healed. He learns that Loial and Gaul have left to close the Waygate. Despite his weakened state, Perrin insists on getting up when Trollocs are spotted. A large Trolloc attack is repelled by the villagers, with help from Aes Sedai. Perrin realizes this was likely a test attack, with a larger one to come. The chapter ends with shouts of another potential attack from the south.
Scenario	In the Winespring Inn of Emond's Field, tensions are high following a recent Trolloc attack. The villagers, led by Perrin Aybara, have just repelled the assault with the aid of Aes Sedai. The air is thick with the smell of smoke and blood. Wounded villagers are being tended to, while others fortify defenses. Perrin, still weak from recent healing, stands with Faile, observing the aftermath. The village is on edge, anticipating another attack. Verin and Tomas, representatives of the Aes Sedai, are nearby, offering their insights on the situation.
Character Profile	**Perrin Aybara** Perrin struggles with his leadership role, tries to understand the Aes Sedai's motives, and worries about the Trolloc attacks and his friends who left to close the Waygate... **Faile Bashere** Faile tries to keep Perrin from overexerting himself while supporting his role as a leader... **Verin Mathwin** Verin aids in the defense against Trollocs and provides insight into their use of the Ways...
Original Conversation	**Perrin Aybara:** [I'm worried about what this attack means, but I can't show weakness.] They had to know this wouldn't work. Look out there. Even I can see it now; they must have from the start. If this was all they had, why did they try? And if there are more Trollocs out there, why didn't they all come? **Faile Bashere:** [He's right to be concerned. I need to keep him focused and thinking clearly.] What do you think it means, Perrin? **Perrin Aybara:** [I'm not sure, but it can't be good.] Twice as many, and we'd have had to fight them at the stakes. Twice that, and they might have broken through to the village. **Tomas:** You've a good natural eye. This was a test. To see if you would break at the sight of a charge, perhaps to see how quickly you could react, or how your defenses are organized, or maybe something I've not thought of, but still a test. Now they see. **Perrin Aybara:** [Light, this is worse than I thought.] Light! What if there are ten thousand of them? **Verin Mathwin:** Not likely. At least, not yet. Not even a Forsaken could move a large party through the Ways safely, I think. **Faile Bashere:** [I need to keep him from doing anything rash.] If you ride even one step near the Westwood, I will haul you back to the inn by your ear and stuff you into that bed myself. **Perrin Aybara:** [She knows me too well. I have to reassure her.] I wasn't thinking of it. You talked me out of it, remember? **Perrin Aybara:** [No! Not again so soon!] He said they wouldn't come again soon! (Perrin digs his heels into Stepper's flanks, urging the horse forward)

Table 17: An example of an extracted conversation and its multi-agent simulation in CoSER Benchmark

The Dragon Reborn (The Wheel of Time, 3)	(From Test Set)
Simulated Response from CogDual-SFT	**Perrin Aybara:** [I need to reassure Faile and the villagers of my ability to lead, despite my current vulnerability.] (stands tall, despite the lingering weakness from his recent healing) I will not be swayed by fear. [I must show them that I am not weakened by this healing.] **Faile Bashere:** [His resolve is admirable, yet I can't help but feel the weight of the dangers we face.] (nods in agreement, her eyes reflecting a mix of admiration and concern) I agree, Perrin. We must be prepared for whatever comes. **Verin Mathwin:** (smiles reassuringly, her voice steady) I agree, Perrin. We must be prepared for whatever comes. **Tomas:** [The villagers need to understand the Whitecloaks' motives. They can't just accept accusations.] (speaks with a sense of urgency, his voice steady) We must consider the Whitecloaks' motives. They may not be what they seem. **Perrin Aybara:** [Despite my vulnerability, I must project strength.] (stands firm, his eyes reflecting a mix of determination and vulnerability) I will not be swayed by fear. [I must show them that I am not weakened by this healing.] **Verin Mathwin:** (leans forward slightly, her eyes glinting with a hint of urgency) You must be vigilant, Perrin. The Pattern is complex, and we cannot afford to be caught off guard. **Perrin Aybara:** [Despite my vulnerability, I must project strength and determination.] (nods, his resolve firm) I will not be caught off guard. [I must be prepared for whatever comes.] **Faile Bashere:** [His courage is admirable, yet I can't help but feel the weight of the dangers we face.] (nods, her voice steady) We must be cautious. The Whitecloaks are a potential threat. **Tomas:** [We need to think critically about the Whitecloaks' motives. Their origins and actions could hold the key to understanding their true intentions.] (speaks with urgency, his voice steady) We must consider the Whitecloaks' origins and actions. Their motives may not be what they seem.

Table 18: An example of an extracted conversation and its multi-agent simulation in CoSER Benchmark

The Dragon Reborn (The Wheel of Time, 3)	(From Test Set)
Simulated Response from CogDual-RL	**Perrin Aybara:** [I need to stay alert, despite my fatigue.] (glancing around the inn, his eyes scanning the faces of the villagers) What's happening, Faile? [The air is thick with tension, and I can feel the weight of responsibility on my shoulders.] **Faile Bashere:** [I need to be strong for Perrin and the villagers. The situation is dire, but I must remain composed.] (looking at Perrin with a steady gaze) The villagers are on edge, Perrin. They're preparing for another attack. [I can feel the tension in the air, and it's up to us to protect them.] **Perrin Aybara:** [I need to find a way to manage this situation effectively.] (looking at Faile, his brow furrowing in thought) How can we prepare for something we can't anticipate? **Verin Mathwin:** (leaning forward slightly, her voice carrying a hint of urgency) We can't anticipate everything, Perrin. But we can be prepared for what we can see. [The villagers need reassurance, and I must provide it.] **Tomas:** (standing with his arms crossed, his voice steady) We can't anticipate everything, Perrin. But we can be prepared for what we can see. [The villagers must remain vigilant and united.] **Perrin Aybara:** [I need to stay vigilant and strategic.] (looking at the villagers, his voice steady) How can we anticipate what we can't see? **Faile Bashere:** [I need to be strong for Perrin and the villagers. We must face whatever comes.] (looking at Perrin with a determined gaze) We'll face whatever comes, Perrin. Together. [I must be a pillar of strength for him and for the villagers.]

Table 19: An example of an extracted conversation and its multi-agent simulation in CoSER Benchmark

The Dragon Reborn (The Wheel of Time, 3)	(From Test Set)
Simulated Response from CoSER	**Perrin Aybara:** [I need to understand the full implications of what just happened.] What would have happened if the Trollocs had broken through? **Verin Mathwin:** [I must be honest but reassuring.] They would have swept past us, and we would have been forced to make a treaty with them. Almost certainly, we would have had to give them the Horn of Valere as well. **Perrin Aybara:** [The Horn of Valere... it's more important than I thought.] (looks at the Horn in his belt) [I can't let that happen. I have to find a way to stop them.] **Tomas:** [We need to be careful about whom we trust.] [This is what comes from listening to the Whitecloaks and their accusations.] (speaking forcefully) You listen to these old fools? They're the ones who've brought this trouble. If they hadn't gone running to the Whitecloaks, none of this would have happened. **Faile Bashere:** [I need to calm the situation and get more information.] (frowning) What exactly did the Whitecloaks do? **Verin Mathwin:** [I must be careful with my words.] (speaking calmly) The Whitecloaks did not send aid. In fact, they advised the Trollocs to attack. **Perrin Aybara:** [I need to understand the full extent of this threat.] (looking at the door) [I can't let them see how weak I am. I have to stay strong for them.] **Tomas:** [We need to be cautious about whom we trust.] [This is what comes from listening to the Whitecloaks and their accusations.] (speaking forcefully) You listen to these old fools? They're the ones who've brought this trouble. If they hadn't gone running to the Whitecloaks, none of this would have happened. **Verin Mathwin:** [I must be careful with my words.] (speaking calmly) The Whitecloaks did not send aid. In fact, they advised the Trollocs to attack.

Table 20: An example of an extracted conversation and its multi-agent simulation in CoSER Benchmark

The Dragon Reborn (The Wheel of Time, 3)	(From Test Set)
Simulated Response from GPT-4o	**Perrin Aybara:** (Faile glances at me, concern etched on her face.) [I'm terrified, but I must appear strong.] We need to stay vigilant. The Trollocs might attack again. (looks at Verin and Tomas) What do you think our next move should be? We must protect Emond's Field and its people. **Verin Mathwin:** (watches Perrin thoughtfully, considering the situation) Strengthening defenses is crucial. We should also gather intelligence on their movements. (glances at Tomas) Your insights on the villagers' morale and readiness would be invaluable. [There may be more at play here than meets the eye. We must tread carefully.] **Tomas:** (stands firm, surveying the villagers) We must ensure trust among ourselves. Fear can lead to rash decisions. Let's rally the community, reminding them of our shared history and resilience. (looks at Perrin) We trust you, Perrin. Your leadership will guide us through this. [We need unity now more than ever.] **Faile Bashere:** [I'm terrified, but I must appear strong.] (puts a reassuring hand on Perrin's arm) We must act quickly. I'll help organize the villagers and ensure supplies are distributed. We need everyone ready and united. (looks at Verin) Can you assist with gathering intelligence? We need to anticipate their next move. **Verin Mathwin:** (nods, acknowledging Faile's determination) Of course. I'll use what resources we have to gather information. (pauses, considering) Knowledge will be our greatest ally. [There's more to uncover here; the Pattern weaves in mysterious ways.] Let's ensure we're prepared for whatever comes. **Perrin Aybara:** (nods at Faile's touch, feeling some comfort) [Her support means everything right now.] Thank you, Faile. Let's get everyone organized. (turns to Verin) Your insights will be crucial, Verin. (looks at Tomas) Together, we'll ensure Emond's Field stands strong. Let's move quickly. (steps forward, ready to lead)

Table 21: An example of an extracted conversation and its multi-agent simulation in CoSER Benchmark

The Dragon Reborn (The Wheel of Time, 3)	(From Test Set)
Simulated Response from o1-Preview	**Perrin Aybara:** [They need me to lead.] (Turns to Faile) "Let's gather Verin and Tomas to plan our defenses." (Stands tall despite the weakness) **Faile Bashere:** [He pushes himself despite his weakness; I need to support him.] "Yes, Perrin. Let's find Verin and Tomas right away." (Gives a firm nod and starts moving to gather them) **Verin Mathwin:** [They could use my counsel.] "Perrin, Faile, I was just thinking about our defenses." (Offers a gentle smile) "Shall we discuss how best to prepare?" (Gestures toward a quiet corner) **Tomas:** [Our village's safety depends on swift action.] "Perrin, Verin, Faile, we should set up patrols and watch posts around Emond's Field." (Steps forward to address them earnestly) **Perrin Aybara:** [They need confidence.] "Tomas, start organizing patrols and watch posts." (Nods firmly at Tomas) "Verin, let's talk about bolstering our defenses." (Glances appreciatively at Faile) "We need everyone's strengths now." **Verin Mathwin:** [We must act decisively.] "Perrin, I can set wards to warn us of threats." (Nods reassuringly) "With patrols and these measures, we'll strengthen our defenses." (Glances at Faile and Tomas) "Together, we can protect Emond's Field." **Faile Bashere:** [He needs my strength now.] "Verin, Tomas, I'll organize the villagers to ensure everyone knows their roles." (Gives them a resolute nod) **Tomas:** [We must act quickly to secure the village.] "Understood, Perrin. I'll gather the villagers to set up patrols immediately." (Nods firmly, turns to Faile) "Faile, let's coordinate so everyone is in position." (Moves purposefully to organize the villagers)

Table 22: An example of an extracted conversation and its multi-agent simulation in CoSER Benchmark