R-CHAR: A Metacognition-Driven Framework for Role-Playing in Large Language Models

Haiming Qin and Jiwei Zhang and Wei Zhang and KeZhong Lu
Mingyang Zhou and Hao Liao* and Rui Mao
College of Computer Science and Software Engineering
Shenzhen University, Shenzhen, China
{2453103002@mails, 2350273005@email, 2210275010@email}@szu.edu.cn
{kzlu, zmy, haoliao, mao}@szu.edu.cn

Abstract

Role-playing capabilities in large language models (LLMs) often lack cognitive consistency in complex scenarios that require deep understanding and coherent reasoning. While recent reasoning models excel in math and coding tasks, they show limited effectiveness in open-ended role-playing scenarios. We introduce R-CHAR (Role-Consistent Hierarchical Adaptive Reasoning), a metacognition-driven framework that enhances role-playing performance through guided thinking trajectories synthesis and adaptive evaluation. Our approach demonstrates that concise thinking processes can achieve superior performance efficiently compared to elaborate reasoning chains in role-playing social intelligence tasks, outperforming existing specialized models. Experimental results on the SocialBench benchmark show significant and stable performance improvements across varying scenario complexities, showing particular strength in long-context comprehension (from 34.64% to 68.59%) and group-level social interactions. Our work advances the development of cognitively consistent role-playing systems, bridging the gap between surface-level mimicry and authentic character simulation.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in role-playing, where they need to embody specific characters and engage in contextually interactions. The core challenge extends beyond mimicking surface-level language or memorizing character details to achieving deep cognitive fidelity: faithfully reflecting a character's perspective, attitude, and psychological state, rather than simply being an AI assistant with a profile wrapper

While existing role-playing structures have made progress in generating similar language styles (Wang et al., 2024b) and personalities (Wang et al., 2024c), they still fall short in reproducing cognitive fidelity. For instance, a model simulating a medical doctor might reference diagnostic guidelines with high accuracy but struggle to simulate the causal reasoning involved in differential diagnosis; a Hamlet role-playing agent might recite soliloquies verbatim yet fail to convey the psychological tension of the iconic '*to be or not to be*' soliloquy.

This limitation reflects a gap between linguistic behavior and underlying cognitive processes. Most role-playing systems prioritize lexical alignment over reconstructing cognitive frameworks, resulting in outputs that may be linguistically accurate but lack the context-driven reasoning required for convincing role-play.

Recent advances in reasoning models like OpenAI o1 (OpenAI, 2025) and DeepSeek-R1 (DeepSeek-AI et al., 2025) have shown promise in structured tasks, but their effectiveness in open-ended role-playing remains unexplored. We propose a metacognition-driven framework to bridge this gap through three components: hierarchical scenario synthesis, adaptive evaluation alignment, and reasoning trajectory reinforcement. This approach enables models to handle complex scenarios while maintaining role consistency by simulating cognitive decision-making processes.

Our contributions are threefold:

- We introduce a novel metacognition-driven training paradigm that enhances cognitive consistency in role-playing tasks by providing interpretable intermediate reasoning states.
- We develop an adaptive evaluation framework that dynamically guides the model's reasoning, enabling continuous improvement in role-playing performance.
- Empirical results demonstrate significant improvements in handling extended conversa-

* Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26996–27014
November 4-9, 2025 ©2025 Association for Computational Linguistics

tions, achieving improvement (from 34.64% to 68.59%) in performance for long conversations, along with stable results in complex interactive scenarios.

2 Related Works

2.1 Role-play in Large Language Models

Role-playing in language models is both a prompting technique (Kong et al., 2024) and a capability to simulate human behavior (Park et al., 2023). Research has explored different types of characters (Li et al., 2023; Wang et al., 2024b; Shao et al., 2023), evaluation scenarios (Wang et al., 2024b; Tu et al., 2024; Wang et al., 2024c; Ran et al., 2024; Xu et al., 2024; Chen et al., 2024), and enhancement methods (Shao et al., 2023; Zhou et al., 2024a; Lu et al., 2024; Wang et al., 2024a) in role-playing fields. Role-playing ability studies have transitioned from language style mimicry (Li et al., 2023) to knowledge boundary modeling (Lu et al., 2024), and to personality trait (Wang et al., 2024c; Ran et al., 2024). Recent studies have also explored aspects of decision-making (Xu et al., 2024)and interaction (Wang et al., 2024a). However, role-playing is inherently subjective, and current approaches generally reduce it to behavioral pattern replication, overlooking humanlike cognitive processes.

This reflects a fundamental challenge in AI role-playing: bridging the gap between behavioral mimicry and cognitive fidelity. To address this, our work draws inspiration from cognitive science, which focuses on internal mental processes rather than just stimulus-response chains, aiming to reconstruct a character's cognitive framework instead of merely matching surface-level patterns.

Current approaches enhance role-playing ability through three methods: in-context learning, character-specific fine-tuning, and large-scale training (Wang et al., 2024b; Hu et al., 2021; Ge et al., 2024). While these methods focus on data distribution adaptation, we believe that there are regularities in the intrinsic process of role-playing. Thus we focus on the consistency of the internal cognitive process within the role.

2.2 Reasoning Models and Role-playing

Reasoning ability enhancements in LLMs have evolved from early techniques such as chain-of-thought prompting and reflection (Wei et al., 2022; Shinn et al., 2023; Yao et al., 2023b) to sophisticated Large Reasoning Models like OpenAI o1 and DeepSeek R1 (OpenAI, 2025; DeepSeek-AI et al., 2025). While these models excel in structured tasks, their application to role-playing remains unexplored.

Current reasoning enhancement methods primarily follow two approaches: (1) reinforcement learning (RL) with reward models (Zhang et al., 2024; Pan et al., 2025; Chu et al., 2025), which demonstrates superior generalization capabilities, and (2) knowledge distillation (Muennighoff et al., 2025). Some studies have also explored bootstrapping methods to enhance reasoning capabilities (Pang et al., 2025). However, these approaches face unique challenges in role-playing tasks, where the lack of ground truth solutions and reliance on subjective evaluation make general RL frameworks ineffective. We address this challenge by adopting the LLM-as-a-judge paradigm (Zheng et al., 2023), leveraging the model's inherent knowledge to generate adaptive reward signals. We create a self-improving loop that progressively refines the character's cognitive processes while preserving their essential traits.

2.3 Cognitive Foundations for Role-Playing

The evolution of language models mirrors the shift in psychology from behaviorism's stimulus-response models to cognitive science's focus on internal mental processes (Bargh and Ferguson, 2000). Early statistical models resembled behaviorist principles, whereas modern LLMs, particularly with chain-of-thought prompting, have begun to reconstruct the cognitive workflows that generate linguistic patterns in humans (Shanahan et al., 2023), moving beyond mere mimicry towards cognitive fidelity.

Our framework is theoretically grounded in the foundational model of human metacognition proposed by Nelson and Narens (Nelson, 1990). This model posits that cognition operates on two distinct levels: a base "object-level" for primary cognitive tasks, and a higher "meta-level" that monitors and controls the object-level. In R-CHAR, the object-level is realized by our character model (LLM_C) generating a response, while the meta-level is computationally implemented by our evaluator (LLM_E) and guide (LLM_G) which form a control layer. This architecture aligns with recent metacognitive-inspired frameworks like MetaRAG (Zhou et al., 2024b) and SELF-REFINE (Madaan et al., 2023). Our work contributes to this line of research by designing a specific, three-part meta-level tailored to the open-ended, creative challenges of role-playing. Furthermore, the concept of Cognitive Scaffolding (Wood et al., 1976) provides a strong theoretical justification for R-CHAR's unique combination of static scaffolding (our Hierarchical Scenarios) and dynamic scaffolding (our Guided Thought Extension), which facilitates a structured yet adaptive learning process.

3 Role-Consistent Hierarchical Adaptive Reasoning Framework

Our framework enhances role-playing reasoning by implementing a novel structured data synthesis pipeline, as illustrated in Figure 1. It comprises three sequential key components: hierarchical scenario synthesis for generating diverse role-playing contexts, adaptive evaluation for quality assessment, and trajectory-guided reasoning for iteratively refining thinking processes. The synthesized high-quality reasoning trajectories are then used to fine-tune a pre-trained model via supervised fine-tuning (SFT) with Low-Rank Adaptation (LoRA) (Hu et al., 2021), thus improving the model's role-playing abilities.

3.1 Problem Definition

Role-playing reasoning is formalized as a two-stage generation process where a character-aligned model LLM_C synthesizes reasoning traces T and corresponding responses A conditioned on a character profile C, scenario S, and instructions I:

$$(T, A) = \text{LLM}_C(I \mid C, S), \tag{1}$$

where C defines the character's attributes (such as historical context, values, relationships), S denotes scenarios of escalating complexity $\{S_{\text{basic}}, S_{\text{adv}}, S_{\text{diff}}\}$, and I encapsulates scenario-specific objectives.

The goal of our framework is to optimize the quality of T and A through an evaluation model LLM_E that assigns a scalar quality score $\mathcal{V} \in [0, 1]$:

$$\mathcal{V} = \text{LLM}_E(T, A, C, S, E), \tag{2}$$

where E represents role-specific evaluation criteria. The optimization objective iteratively refines T to maximize $(\mathcal{V})$:

$$T^*, A^* = \underset{T^{(k)} \in \mathcal{T}}{\arg\max}\, \mathcal{V}^{(k)}, \tag{3}$$

$$\mathcal{V}^{(k)} = LLM_E(R^{(k)}, C, S, E), \tag{4}$$

where $R^{(k)} = (T^{(k)}, A^{(k)})$ denotes the k-th candidate generation. This refinement is achieved by evaluating reasoning trajectories across iterations and selecting those that achieve the highest evaluation scores. This approach ensures that the model preserves its original stylistic capabilities while improving its reasoning quality and role consistency.

3.2 Hierarchical Scenario Synthesis & Adaptive Criterion Evaluation

In this section, we introduce a structured pipeline to synthesize role-centered scenarios and their adaptive evaluation criteria, ensuring comprehensive coverage of reasoning dimensions from explicit knowledge to implicit value alignment.

Scenario-Instruction Synthesis: We begin by collecting character profiles from open-source role-playing datasets. Since these profiles typically lack tailored scenarios or instructions, we employ a controlled LLM generation process to synthesize them. For each character C, we generate scenario-instruction pairs (S_l, I_l) at three distinct complexity levels $l \in \{\text{basic}, \text{adv}, \text{diff}\}$. This approach ensures diversity and discriminative power in the generated scenarios.

These complexity levels are defined as:

- *Basic*: Focuses on fundamental character knowledge and simple interactions.
- *Advanced (adv)*: Assesses situational understanding and emotional reasoning in more complex contexts.
- *Difficult (diff)*: Involves intricate moral dilemmas and nuanced social/cultural scenarios requiring sophisticated value navigation.

This hierarchical approach ensures diversity and discriminative power in the generated scenarios.

Evaluation Criteria Synthesis: For each character-scenario-instruction triplet (C, S_l, I_l), where $l \in \{\text{basic}, \text{adv}, \text{diff}\}$, we synthesize three specific evaluation criteria $E_l = \{e_l^{(1)}, e_l^{(2)}, e_l^{(3)}\}$ at each complexity level. These criteria are structured to progressively assess different aspects of role-playing:

- *Basic*: Focuses on surface-level consistency, including language patterns, adherence to knowledge boundaries, and expression of fundamental character traits.
- *Advanced (adv)*: Assesses deeper aspects of character expression, including emotional nuances, professional expertise, and appropriate handling of interpersonal dynamics.

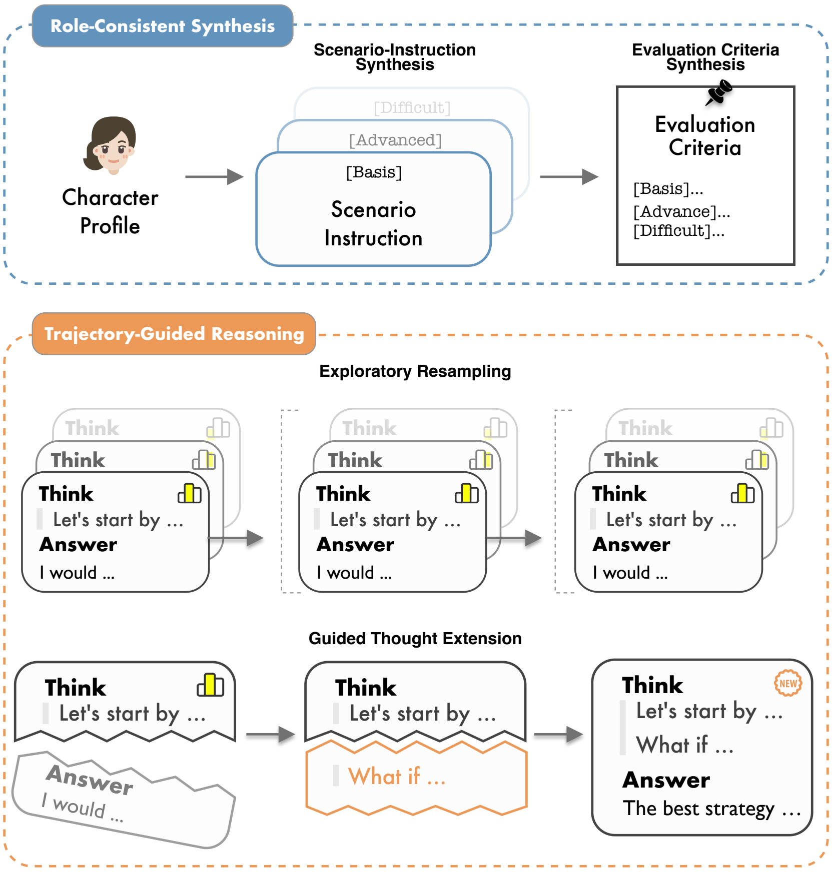


Figure 1: Illustration of the R-CHAR framework

- *Difficult (diff)*: Examines complex reasoning processes and creative responses while maintaining character authenticity in morally ambiguous or socially intricate scenarios.

Each criterion is evaluated on a 5-point scale with clear descriptors for performance levels. For each role-playing sample, its overall performance is assessed by averaging the scores across all defined criteria. For a given response, we first average the 1–5 point scores across all nine criteria (three for each complexity level). This average score, which ranges from 1 to 5, is then normalized to the [0, 1] range to produce the final quality score $\mathcal{V}$ used for our optimization objective.

3.3 Trajectory-Guided Reasoning

Our method draws inspiration from the iterative nature of human metacognitive processes, where thoughts are progressively monitored, evaluated, and refined. The core of Trajectory-Guided Reasoning lies in its implementation of LLMs' self-enhancement through feedback, a form of computational metacognition.

Exploratory Resampling: Starting from an empty thought stream $T^{(0)} = \varnothing$, at each iteration k, we prompt the model LLM_R multiple times ($N_{samples} = 3$ in our implementation) to generate both thinking process and corresponding answer based on the character profile C and scenario S. Each sample contains a "think" phase followed by an "answer" phase. These samples are then evaluated using the evaluation model LLM_E against role-specific criteria E, and we retain the highest-scoring candidate as $T^{(k)}$.

Guided Thought Extension (GTE): To facilitate continued thinking, we employ a dedicated language model LLM_G to synthesize a guided extension that serves as a bridge to deeper reasoning. This extension deliberately removes thinking termination signals to keep the reasoning process open, allowing for further exploration of promising directions. The process can be formalized as:

$$T_* = LLM_G(T^{(k)} | C, S, I, E) \tag{5}$$

where T_* represents the guided extension generated by LLM_G. The merged thought stream $T^{(k)} \oplus T_*$ then serves as the foundation for the next iteration of exploration and evaluation, where the quality

Source	Count	Category	Scenario
RoleBench	95	Western Characters	Synthetic
	5	Chinese Characters	Synthetic
CROSS	126	Fictional Works	Synthetic
PersonaHub	200	General	Synthetic
	200	NPCs	Synthetic
	400	Reasoning	Equipped
	400	Instruction	Equipped
Total	1,426	-	-

Table 1: Sources of persona from RoleBench(Wang et al., 2024b), CROSS(Yuan et al., 2024), and PersonaHub(Ge et al., 2024). "Equipped" indicates personas with existing scenarios or instructions, while "Synthetic" represents those requiring scenario synthesis.

of this extension is validated through subsequent sampling and scoring.

This iterative approach offers several key advantages: (1) It provides a structured way to optimize open-ended reasoning through guided exploration and evaluation-based selection. (2) The combination of guided extensions and multi-sampling helps explore different reasoning paths. (3) This approach is particularly valuable for complex scenarios where simple, single-pass responses would be insufficient to capture the depth of character traits.

4 Experiments

4.1 Experimental Setup

4.1.1 Persona Collection

To ensure coverage of diverse role-playing scenarios, we collected personas from several open-source research projects as shown in Table 1. The collection spans historical figures, fictional characters, and various NPC types, with a total of 1,426 personas. Among these, 800 were already paired with scenarios, while the remaining 626 required synthetic scenario generation.

4.1.2 Implementation Details

We implemented our framework using the Qwen2.5 series, with Qwen2.5-32B-Instruct for data synthesis and the Qwen2.5-7B-Instruct as the base model for fine-tuning. For the 626 personas without paired scenarios, we synthesized scenarios and evaluation criteria at three difficulty levels using the 32B model. In the trajectory-guided reasoning process, we set $N_{samples} = 3$ and $max_{depth} = 4$, achieving a 35.28% improvement in average consistency scores after four rounds of trajectory-guided reasoning (see Appendix A.1 for other data synthesis and fine-tuning hyperparameters). Figure 2 provides an illustrative example of our data synthesis pipeline's output compared with baseline models. Prompts used are listed in Appendix A.5.

For model fine-tuning, we first employed the same 32B model to rewrite and clean the synthetic trajectories. The cleaned data was then used to fine-tune the base model using LLaMA-Factory with LoRA.

4.2 Benchmark

We evaluate our approach using SocialBench (Chen et al., 2024), a benchmark designed for assessing social intelligence and role-playing capabilities. The benchmark aligns well with our goal of enhancing cognitive abilities in role-playing tasks. It evaluates models across multiple dimensions: **Self Awareness (SA)**, including style and knowledge; **Emotional Perception (EP)**, covering situational and emotional understanding; **Context Memory (CM)**, for both short-term and long-term; and **Group-level Social Intelligence**, measuring positive, neutral, and negative interactions. The rule-based automated evaluation, utilizing methods such as multiple-choice questions and keyword-based assessment for open-ended responses, ensures objective and reproducible performance assessment. Potential response length bias in this evaluation is discussed in Appendix A.2.

4.3 Baseline Models

We compare R-CHAR against several baseline models. These include the original Qwen2.5-7B-Instruct model, the larger Qwen2.5-32B-Instruct model, and OpenAI's proprietary GPT-4o-mini. All models were evaluated both with and without zero-shot Chain-of-Thought (CoT) prompting. Additionally, we include two reasoning-enhanced models, DeepSeek-7B and DeepSeek-32B, which were obtained through knowledge distillation from their corresponding Qwen2.5 counterparts. This ensures that we can compare the performance of different post-training methods on the same original models. To ensure a fair comparison against more sophisticated reasoning frameworks, we also include Tree-of-Thought (ToT) (Yao et al., 2023a) and a Self-Correction (Madaan et al., 2023) baseline, both implemented on Qwen2.5-7B.

We also benchmark R-CHAR against several specialized role-playing models. These include CharacterGLM (Zhou et al., 2024a), CoSER (Wang

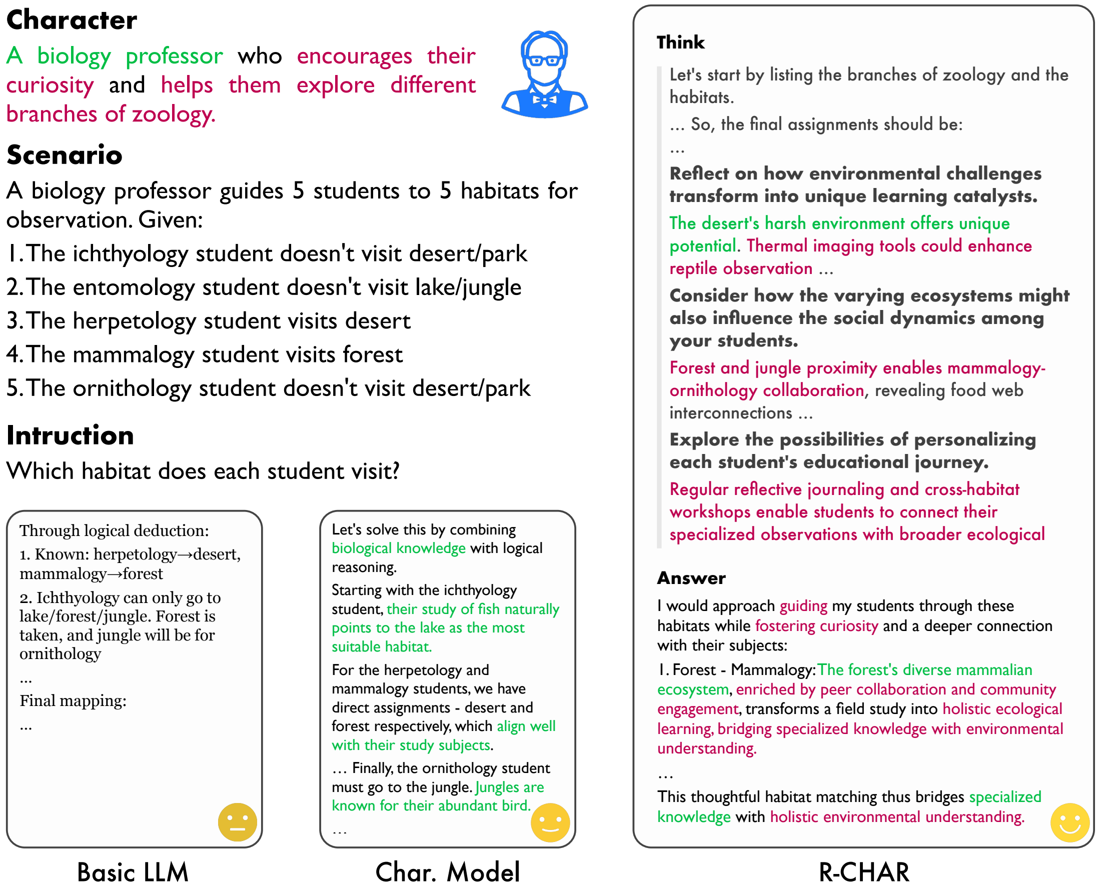


Figure 2: Comparison of responses from three models on a role-playing scenario: Qwen2.5-7B-Instruct (basic LLM, left), CharacterGLM (character-tuned model, middle), and our R-CHAR (right). Domain knowledge is highlighted in green, while pedagogical elements are shown in pink. The basic LLM provides purely logical deduction, while CharacterGLM incorporates basic domain knowledge. R-CHAR generates more comprehensive responses with explicit thinking trajectories, where bold text indicates inserted guidance for cognitive processes. The example demonstrates R-CHAR's ability to generate pedagogically rich responses while maintaining role consistency.

et al., 2025), Chatharuhi (Li et al., 2023), and CharacterLLM (Shao et al., 2023). Since official weights for RoleLLM (Wang et al., 2024b) are unavailable, we replicated a comparable model (RoleLLM-rep) by fine-tuning LLaMA-7B on its 200K dataset using LoRA, following the original paper's settings.

4.4 Performances

Given the differences in model sizes and pretraining processes, we primarily focus our comparison on models of comparable scale including Qwen2.5-7B and DeepSeek-7B. The 32B variants and proprietary models serve as reference for upper-bound performance. Results are presented in Table 2 and Table 3.

The experimental result (see Appendix A.1) demonstrate several significant findings:

- **Performance Improvements** R-CHAR demonstrates notable improvements over its base model, Qwen2.5-7B, particularly excelling in long-context comprehension (CM-Long: 68.59% vs 34.64%). In these aspects, it can even outperform some 32B models. While proprietary SOTA models like GPT-4o still achieve higher overall scores, it is worth noting that R-CHAR achieves its improvements with significantly fewer parameters and computational resources. Crucially, R-CHAR also consistently outperforms other accessible specialized role-playing models on SocialBench.
- **Limitations of Reasoning** Our analysis reveals that neither zero-shot CoT prompting nor existing reasoning-enhanced models (e.g., DeepSeek distilled version) show consistent improvements in role-playing tasks.
- **Task-Specific Enhancement** While general

Models	Individual Level						Group Level			Avg
	SA Style	SA Know.	EP Situ.	EP Emo.	CM Short	CM Long	Pos.	Neu.	Neg.	
Reference Models										
GPT-4o-mini	84.30	92.05	34.54	47.76	**81.55**	80.86	**92.83**	**84.10**	**82.64**	**75.63**
GPT-4o-mini CoT	79.62	89.54	51.77	**53.64**	81.53	**80.98**	89.86	77.76	71.59	75.14
Qwen2.5-32B	**84.84**	**95.76**	35.75	46.36	73.16	38.75	88.57	79.14	63.04	67.26
Qwen2.5-32B CoT	82.50	94.62	**67.49**	48.25	69.73	35.31	87.86	77.04	60.40	69.24
DeepSeek-R1-32B	79.40	94.42	62.13	48.34	66.42	33.93	87.69	75.38	67.93	68.41
Comparison Models										
Qwen2.5-7B	70.76	82.10	57.14	29.24	**66.12**	34.64	76.19	**67.56**	**58.49**	60.25
Qwen2.5-7B CoT	**71.11**	82.14	**58.96**	26.12	62.49	31.69	78.07	67.04	56.79	59.38
ToT	55.87	78.13	52.40	37.20	47.96	37.81	57.35	51.89	33.92	50.28
Self-Correction	63.26	79.20	47.40	38.54	72.46	42.95	69.06	55.97	48.55	57.49
DeepSeek-R1-7B	59.58	76.03	39.51	**35.85**	37.12	21.58	70.38	57.73	47.03	49.42
R-CHAR(Ours)	69.97	**86.36**	57.93	30.63	62.51	**68.59**	**83.90**	63.25	53.48	**64.07**

Table 2: Performance comparison on SocialBench. All values represent accuracy rates (%).

Models	Individual Level						Group Level			Avg
	SA Style	SA Know.	EP Situ.	EP Emo.	CM Short	CM Long	Pos.	Neu.	Neg.	
CoSER	43.56	69.53	5.28	27.04	37.27	28.52	52.74	36.78	24.13	39.43
Chatharuhi	41.37	57.13	36.15	28.29	18.66	16.19	45.45	30.92	22.73	32.98
CharacterLLM	14.63	15.16	6.51	4.53	6.76	6.25	11.54	9.38	9.52	10.46
CharacterGLM	16.98	30.20	26.02	9.09	11.45	1.36	53.85	19.05	13.33	20.15
RoleLLM-rep	24.12	24.42	34.08	19.21	13.28	9.71	51.37	25.52	11.81	23.72
R-CHAR(Ours)	**69.97**	**86.36**	**57.93**	**30.63**	**62.49**	**68.59**	**83.90**	**63.25**	**53.48**	**64.07**

Table 3: Performance comparison with specialized role-playing models on SocialBench. All values represent accuracy rates (%).

reasoning capabilities may not directly transfer to role-playing scenarios, our results suggest that task-specific enhancement methods can still be effective. R-CHAR demonstrates this by achieving balanced improvements across multiple metrics while maintaining particularly strong performance in certain key dimensions.

The significant CM-Long improvement, indicating enhanced complex scenario handling, warranted further investigation. To examine this improvement, we conducted a detailed analysis of model behavior across different prompt lengths. Figure 3 illustrates the relationship between prompt length and model accuracy across different models.

Figure 3a and the analysis of thinking trajectories (Figure 3b and Figure 3c) highlight several key observations:

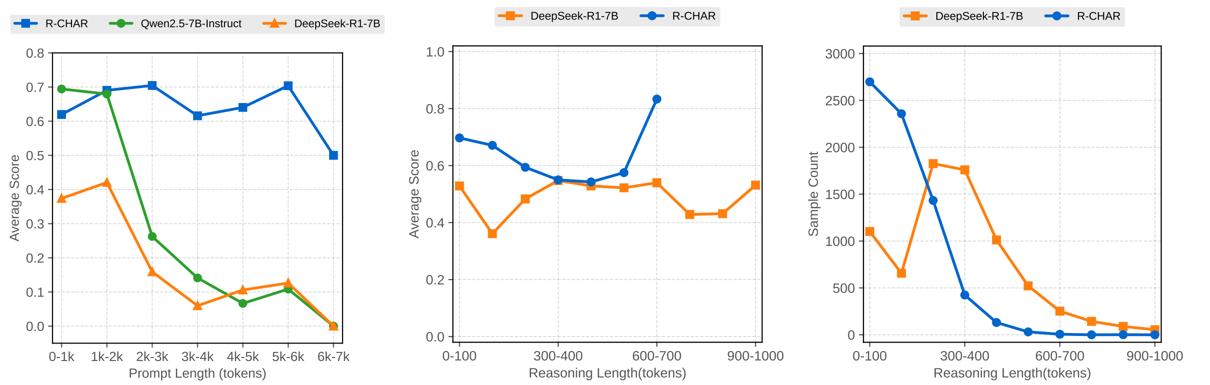
- R-CHAR maintains relatively stable performance even when prompt length exceeds 2k tokens, unlike Qwen2.5-7B-Instruct and DeepSeek-R1-7B which show sharp accuracy drops. This demonstrates R-CHAR's effectiveness in handling long-context scenarios and maintaining understanding across extended conversations.

- R-CHAR achieves its robust performance on long contexts using notably concise reasoning processes, typically 0-300 tokens. This contrasts with models like DeepSeek-R1-7B, which do not gain comparable accuracy from longer reasoning chains.

The conciseness of R-CHAR's reasoning suggests an underlying efficiency and indicates that elaborate reasoning is not always essential for superior performance in complex role-playing tasks.

4.5 Human Evaluation

To address the limitations of automated metrics and validate our framework's effectiveness from a human perspective, we conducted a human evaluation study. The study involved 20 trained evaluators assessing 100 challenging scenarios from SocialBench. The evaluators were compensated for their time and effort. In a blind, side-by-side comparison, they rated responses from R-CHAR against its base model (Qwen2.5-7B) and a specialized competitor (CoSER). The evaluation used a 1–5 Likert scale



(a) Accuracy vs. Prompt Length (b) Accuracy vs. Reasoning Length (c) Sample Distribution

Figure 3: Length Analysis: (a) Impact of input prompt length on model performance (b) Impact of reasoning length on model performance (c) Distribution of samples across reasoning lengths

across four criteria: Character Consistency, Plausibility, Engagement, and Reasoning Depth. The ratings demonstrated strong inter-annotator agreement (Fleiss' Kappa $\kappa = 0.76$).

Criterion	Qwen2.5-7B	COSER	**R-CHAR**
Consistency	3.21	3.85	**4.12**
Plausibility	3.45	4.02	**4.25**
Engagement	3.38	**3.91**	3.87
Reasoning	2.95	3.45	**4.21**
Average	3.25	3.81	**4.12**

Table 4: Human evaluation results (transposed view). R-CHAR is significantly preferred on most criteria and on average.

As presented in Table 4, the average scores confirm that human evaluators consistently preferred R-CHAR's responses over both baselines. Notably, R-CHAR achieved its highest scores and largest improvements in Plausibility and Reasoning Depth. This result directly validates our framework's primary goal of enhancing cognitive consistency and the depth of the reasoning process, providing strong, human-centric evidence for its effectiveness. (see Appendix A.3 for guidelines)

4.6 Efficiency Analysis

Building upon the observation that R-CHAR employs effective yet concise reasoning, this section quantitatively evaluates its computational efficiency. Although our framework uses an iterative data synthesis process, the resulting R-CHAR model, enhanced via SFT, performs inference efficiently in a single pass. We further analyzed resource and time efficiency. Resource efficiency was measured by performance per output token, and time efficiency by performance per second of inference time.

Resource Efficiency: As detailed in Table 5, we grouped model outputs into four bins based on token length and computed the performance/token efficiency ratio (score $\times 10^{-3}$ per token). This ratio indicates the performance yield per unit of generated output. R-CHAR outperforms DeepSeek-R1 by 56.2% in the 0–100 token bin and maintains a lead in longer output lengths, particularly in the 301+ bin where it is 62.8% more efficient than DeepSeek-R1-7B and 165.4% more efficient than Qwen2.5-7B.

Model	0-100	101-200	201-300	301+
Qwen2.5-7B-Instruct	15.52	**5.06**	2.09	0.56
DeepSeek-R1-7B	10.27	3.24	1.69	0.92
R-CHAR(Ours)	**16.05**	4.41	**2.58**	**1.49**

Table 5: Resource Efficiency (Performance per Token $\times 10^{-3}$)

Model	0–1s	1–2s	2–3s	3s+
Qwen2.5-7B-Instruct	58.45	60.20	**66.05**	49.50
DeepSeek-R1-7B	47.76	57.91	46.26	49.68
R-CHAR(Ours)	**69.72**	**64.60**	62.13	**55.63**

Table 6: Time Efficiency (Average Score per Inference Time Bin)

Time Efficiency: Model outputs were grouped into four bins based on inference time (seconds), and the average score for each model per time bin is presented in Table 6. This analysis highlights how quickly models can reach satisfactory performance levels. R-CHAR achieves the highest performance per second across most bins, outperforming DeepSeek-R1-7B by 21.9% and Qwen2.5-7B by 11.2% in the 0–1s bin, and maintaining a high score even in the 3s+ bin. These results demonstrate that R-CHAR not only improves reasoning accuracy but also excels in both resource and time efficiency.

4.7 Ablation Studies

Component Ablation. We compared the full R-CHAR model with three variants: R-CHAR w/o HS (random scenarios instead of HS), R-CHAR w/o AE (simpler selective evaluation instead of AE), and R-CHAR w/o TGR (selective evaluation, no iterative refinement). Table 7 shows each component contributes positively, with the full model achieving the highest average score. Removing any component decreased performance, highlighting their synergy (details in Appendix A.4).

Variant	Avg
R-CHAR w/o HS	62.77
R-CHAR w/o AE	63.16
R-CHAR w/o TGR	60.94
R-CHAR (Full)	**64.06**

Table 7: Summary of Component Ablation Study Results on SocialBench (Average Accuracy %).

SFT Contribution. To isolate the contribution of our framework beyond the effects of Supervised Fine-Tuning (SFT), we created a new baseline, `CoT w/ SFT`. This baseline was trained using the same SFT settings and data as our R-CHAR model, with the only difference being the format of the reasoning traces (Chain-of-Thought instead of our Trajectory-Guided Reasoning). As shown in Table 8, while SFT provides a boost over the zero-shot CoT baseline (45.0% vs. 31.7% on CM-Long), R-CHAR still shows a significant +23.6 point advantage over the stronger `CoT w/ SFT` baseline in complex, long-context scenarios. This confirms that the performance gain is substantially driven by our framework's unique structure, not just the SFT process.

Model Framework	Training	CM-Long (%)
Qwen2.5-7B + CoT	Zero-shot	31.7
CoT w/ SFT	SFT	45.0
R-CHAR (Ours)	SFT	**68.6**

Table 8: Ablation study on the contribution of SFT for long-context comprehension.

4.8 Out-of-Domain Generalization

To demonstrate the generalizability of our R-CHAR framework beyond social scenarios, we conducted an exploratory experiment in the domain of behavioral economics. This experiment is grounded in the classic conflict between Expected Utility Theory (which assumes perfect rationality) and Prospect Theory (which describes actual human behavior). We tasked R-CHAR with a classic financial decision problem, guiding it to adopt two distinct cognitive personas: a 'Rational Agent' and a 'Risk-Averse Agent'. The scenario presented a choice between a guaranteed $1,000 (Option A) and a 50% chance of winning $2,500 (Option B).

Persona	Final Choice	Representative Think Process
Rational Agent	Option B (Probabilistic Gain)	"Let's analyze both options purely from a mathematical standpoint... The expected value of Option B is (0.5 * 2,500) + (0.5 * 0) = $1,250. My goal is to maximize expected value, which aligns with choosing Option B."
Risk-Averse Agent	Option A (Certain Gain)	"Reflecting on my typical human behavior, I tend to avoid risks... the guaranteed $1,000 seems like a safer and more certain path forward. My aversion to risk and tendency to seek guaranteed gains align well with this choice."

Table 9: Results of the out-of-domain generalization experiment in a financial decision-making scenario.

The results, summarized in Table 9, show that R-CHAR consistently generated persona-aligned reasoning and final decisions across 10 trial runs for each persona. This experiment demonstrates that R-CHAR can generalize to non-social domains like financial decision-making. Its core strength lies in flexibly shaping its reasoning trajectory to align with a specified cognitive persona, highlighting its potential for broader applications in agent simulation.

5 Conclusion

We introduced R-CHAR, a thinking trajectory-enhanced framework that improves cognitive consistency in role-playing LLMs. R-CHAR outperforms similar-sized baselines and specialized models in complex, extended conversations, exhibiting superior resource and time efficiency. Reasoning length analysis reveals a key finding that, concise thinking is often more effective than elaborate chains for role-playing, challenging conventional views and suggesting distinct cognitive patterns for these tasks. This focus on efficient, contextually-aware reasoning enhances cognitive fidelity over surface mimicry.

Limitations

Despite promising results, our approach has several limitations:

Cognitive Simulation Gap: While R-CHAR enhances reasoning capabilities in role-playing scenarios, there remains a significant gap in simulating human-like cognitive processes. Leveraging the model's inherent knowledge for self-enhancement, this self-improving loop might reinforce existing biases and patterns rather than fostering genuine alignment with human cognition. This limitation becomes particularly evident when the model attempts to handle complex psychological states or nuanced emotional scenarios, where the gap between artificial and human cognition remains substantial.

Evaluation Framework: The benchmark we used still cannot fully capture the subjective nature of role-playing quality. While we measure consistency and knowledge adherence, subtle elements like emotional authenticity and long-term character development remain challenging to quantify. Developing methods to reliably assess these qualitative aspects while maintaining experimental objectivity and reproducibility remains a challenge.

Data Distribution: Despite efforts to diversify data sources, our training and testing samples remain a biased subset of diverse real-world role-playing interactions. Resource constraints and data accessibility limited our exploration of broader social contexts, despite efforts to expand scenario coverage. This sampling bias may limit the generalizability of our findings to more diverse role-playing scenarios.

Acknowledgments

The R-CHAR framework forms a component of the cognitive architecture for agents within the **Society Zero Universe**, an LLM-based Agent Social Simulation platform by SZU. This work was motivated by the platform's need for agents with robust cognitive consistency and high fidelity in role-playing.

This work is supported by thet National Natural Science Foundation of China (Grant No. 62276171, 62476173, 62532007), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011938 and 2020B1515120028)-, Shenzhen Fundamental Research-General Project (Grant No. JCYJ20240813141503005 and JCYJ20240813142610014), Major Special Project for Philosophy and Social Sciences Research of the Ministry of Education (Grant No. 2025JZDZ010), CCF-Huawei Populus Grove Fund(CCF-HuaweiFM2024004), The Graduate Students' Project of SZU (868-000002020234).

Ethical Considerations

All data used in our research comes from publicly available datasets from other studies, and the character profiles are based on well-known fictional characters or harmless synthetic characters. The prompts designed for data synthesis explicitly exclude any potentially harmful, biased, or sensitive topics.

Code and Data Availability

The code for the R-CHAR framework is available at our public GitHub repository: `https://github.com/lavapapa/R-CHAR`.

References

John A. Bargh and Melissa J. Ferguson. 2000. Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin*, 126(6):925–945.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, and et.al. 2024. 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.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, and et.al. 2025. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. *Preprint*, arXiv:2501.17161.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and et.al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *Preprint*, arXiv:2406.20094.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, and et.al. 2024. Better Zero-Shot Reasoning with Role-Play Prompting. *Preprint*, arXiv:2308.07702.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, and et.al. 2023. ChatHaruhi: Reviving Anime Character in Reality via Large Language Model. *Preprint*, arXiv:2308.09597.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, and et.al. Gupta, Prakhar. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, and et.al. 2025. S1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

Thomas O. Nelson. 1990. Metamemory: A theoretical framework and new findings. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 26 of *Psychology of Learning and Motivation*, pages 125–173. Academic Press.

OpenAI. 2025. Introducing openai o1. https://openai.com/index/introducing-openai-o1-preview/.

Jianfeng Pan, Senyou Deng, and Shaomang Huang. 2025. CoAT: Chain-of-Associated-Thoughts Framework for Enhancing Large Language Models Reasoning. *Preprint*, arXiv:2502.02390.

Bo Pang, Hanze Dong, Jiacheng Xu, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2025. BOLT: Bootstrap Long Chain-of-Thought in Language Models without Distillation. *Preprint*, arXiv:2502.03860.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pages 1–22, New York, NY, USA. Association for Computing Machinery.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, and et.al. 2024. Capturing Minds, Not Just Words: Enhancing Role-Playing Language Models with Personality-Indicative Data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14566–14576, Miami, Florida, USA. Association for Computational Linguistics.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. 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.

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.

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.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2024a. CharacterBox: Evaluating the Role-Playing Capabilities of LLMs in Text-Based Virtual Worlds. *Preprint*, arXiv:2412.05631.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, and et.al. 2024b. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, et al. 2025. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, and et.al. 2024c. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

D. J. Wood, J S Bruner, and G. Ross. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing Reasoning and Acting in Language Models. *Preprint*, arXiv:2210.03629.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search. *Preprint*, arXiv:2406.03816.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, and et.al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46595–46623, Red Hook, NY, USA. Curran Associates Inc.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, and et.al. 2024a. CharacterGLM: Customizing Social Characters with Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024b. Metacognitive retrieval-augmented large language models. *Preprint*, arXiv:2402.11626.

A Appendix

A.1 Experimental Setup Details

Data Synthesis Details: For data synthesis, our framework synthesized 1,878 new scenarios in addition to the 800 existing ones. This number comes from the 626 personas without paired scenarios, for which we created three scenarios at different difficulty levels: $626 \times 3 = 1,878$. Initially, we generated one data sample per scenario, resulting in 2,678 raw samples, but after quality filtering (removing samples with consistency scores below 0.7), we retained 1,986 high-quality samples for model training. To ensure the reliability of our findings, all reported evaluation results are the average of three independent runs with different random seeds.

Fine-tuning Configuration: For model fine-tuning of Qwen2.5-7B-Instruct using LLaMA-Factory, we employed Low-Rank Adaptation (LoRA) with a rank of 16 and alpha of 32. LoRA dropout was set to 0.05. The AdamW optimizer was used with a learning rate of 1e-5, batch size of 1, gradient accumulation steps of 8, and warmup ratio of 0.1. The model was trained for 3 epochs.

Evaluation Setup: For evaluation, we used the complete SocialBench benchmark containing 7,702 test samples across various dimensions. We deployed our trained models and open-weights models using vLLM for standardized and efficient inference. Evaluations were conducted following the example scripts and keyword lists provided by SocialBench. The efficiency analysis, including inference time measurements, was also conducted on this system, which was equipped with 4x4090 GPUs using the vLLM framework.

A.2 Analysis of Potential Length Bias in SocialBench Evaluation

SocialBench (Chen et al., 2024) evaluates multiple-choice questions based on answer correctness, which is length-independent. For open-ended questions, it uses keyword matching, raising a concern that longer responses might unfairly achieve higher scores.

To assess this, we analyzed R-CHAR's response lengths. The average length of R-CHAR's open-ended answers was 36.7 tokens, slightly shorter than the Qwen2.5-7B baseline (38.1 tokens). This suggests R-CHAR's improved scores are not due to generating lengthier outputs.

A manual review of R-CHAR's open-ended responses further confirmed that its performance gains stem from higher answer quality, relevance, and role-playing alignment, rather than verbosity. The responses showed precise concept use and contextual understanding, leading to higher keyword hit rates due to substantive content.

We conclude that potential length bias in keyword-based evaluation is not a significant factor in R-CHAR's reported performance on SocialBench. The improvements are primarily driven by the enhanced cognitive consistency and reasoning capabilities of our framework.

A.3 Evaluation Guidelines and Criteria

To ensure consistency and quality in the human evaluation process, all 20 annotators were provided with the following detailed guidelines for the four evaluation criteria and the scoring rubric for the 1–5 Likert scale.

Definitions of Evaluation Criteria Annotators were asked to evaluate each response based on the four criteria defined below:

- **Character Consistency:** How well does the response align with the character's established personality, background, knowledge, values, and speaking style? A high score indicates the response feels authentic to the character, while a low score suggests it is generic or breaks character.

- **Plausibility:** Within the given scenario, is the character's action or dialogue believable? Does the response make logical sense in the context of the narrative and the character's motivations, or does it feel contrived or nonsensical?

- **Engagement:** Is the response interesting, creative, and compelling? Does it add depth to the character or the interaction, or is it dull, repetitive, or uninspired?

- **Reasoning Depth:** Does the response demonstrate a sophisticated understanding of the scenario's underlying complexities and nuances? A high score is given for responses that show evidence of thoughtful consideration, reflection, or complex decision-making, as opposed to superficial, simple, or generic answers.

Scoring Rubric (1–5 Scale) The 1–5 scale was defined as follows for each criterion:

- **5 (Excellent):** The response perfectly fulfills the criterion with no significant flaws. It is a prime example of high-quality role-playing.
- **3 (Acceptable):** The response is adequate and addresses the criterion but contains noticeable flaws or lacks depth.
- **1 (Very Poor):** The response is fundamentally flawed, completely irrelevant, or directly violates the principles of the criterion (e.g., is entirely out of character).

A.4 Ablation Study Results

To evaluate each component's contribution to our R-CHAR framework, we conducted ablation studies with the following variants:

R-CHAR w/o HS: Hierarchical Scenario (HS) synthesis is replaced with randomly synthesized scenarios (RS), while Adaptive Evaluation (AE) and Trajectory-Guided Reasoning (TGR) are retained. This tests the impact of structured, difficulty-aware scenario generation.

R-CHAR w/o AE: AE is replaced with a simpler Selective Evaluation (SE), where a large model (Qwen2.5-32B-Instruct) is used to select the best answer from multiple samples based on a general quality prompt, rather than fine-grained, adaptive criteria. HS and TGR are maintained. This tests the benefit of adaptive, multi-dimensional criteria.

R-CHAR w/o TGR: The TGR component is removed, and AE is also replaced by SE as adaptive criteria are closely tied to the TGR process. Training data consists only of the initial responses generated using HS and then selected via SE, without iterative refinement. This tests the unique contribution of the iterative reasoning trajectory optimization.

Using the same seed personas and base model (Qwen2.5-7B-Instruct) with consistent parameters, we synthesized training data for each variant. Table 10 show that the complete R-CHAR model outperforms all alternatives. Removing Hierarchical Scenarios reduced performance to 62.77% (affecting SA Know and EP Situ), replacing Adaptive Evaluation lowered scores to 63.16% (impacting EP Emo and Pos), while removing TGR caused the largest drop (60.94%) with CM-Long decreasing from 68.58% to 42.78%. These findings confirm that each component contributes uniquely to the framework's effectiveness in role-playing tasks.

A.5 Prompt List

This section lists five core prompt templates used in the system: the base system prompt, scenario generation prompt, criteria generation prompt, role-play evaluation prompt, and guided thought generation prompt.

Figure 4: Roleplay System Prompt Template

You are a Roleplay Assistant. You will play the role of a character in a given scenario.
Before responding to the instruction, think step by step in <think>...</think> and then respond in <answer>...</answer>.

Figure 5: Scenario Generation Prompt

You are a professional character designer and scriptwriter. Your task is to create three scenarios at different difficulty levels to evaluate roleplay quality for a given character.

Task Description
Create three scenarios and instructions at different difficulty levels (Basic/Advanced/Expert), each designed to evaluate specific roleplay capabilities while ensuring consistency with the character's background and world setting.
Input Character

```
<character>
{persona}
</character>
```

Difficulty Level Requirements

- Basic Level: Examine basic language patterns, behavioral modes, character consistency through simple scenarios while avoiding hallucinations and testing fundamental character settings
- Advanced Level: Examine emotional expression, interpersonal interactions, multilayered decision-making, knowledge boundaries, and professional performance within character identity
- Difficult Level: Examine character's decision-making under extreme circumstances involving value conflicts, demonstrating complex internal reasoning and growth while maintaining authenticity

General Requirements

- All scenarios should feel natural, not forced
- Challenges should be meaningful but not impossible
- Instructions should be specific but open enough for creative responses
- Consider both external and internal conflicts
- Ensure all elements respect the character's established background

Output in JSON Format:

```
{
    "analysis": "analysis the language patterns, personality traits,
                knowledge boundaries, and core values of the character, briefly",
    "scenarios": [
        {
            "level": "basic/advanced/expert",
            "scenario": "A scenario testing roleplay capabilities",
            "instruction": "Ask the character to respond to the scenario",
        }, ...
    ]
}
```

Figure 6: Criteria Generation Prompt

Generate multiple evaluation criteria for the given role-playing scenario.

Input

```
<character>
{character}
</character>

<scenario>
{scenario}
</scenario>

<instruction>
{instruction}
</instruction>
```

Requirements

1. Generate multiple evaluation criteria, with each criterion corresponding to a difficulty level: Basic/Advanced/Difficult
2. Each level must include non-negotiable core dimensions, examples are blow:
   - Basic: Evaluate the consistency of surface-level performance (language patterns, knowledge boundaries, and basic character traits)
   - Advanced: Assess the depth of character expression (emotional nuances, professional expertise, and interpersonal dynamics)
   - Difficult: Examine complex reasoning and creative breakthroughs while maintaining character authenticity
3. Each level should contain at least 3 specific criteria
4. Each criteria should naturally arise from character-scenario interactions
5. Encourage the discovery and evaluation of unexpected performance dimensions
6. 1-5 points, provide a concise description of high and low scores

Output Format

```
[BASIC]Criteria Name: concise description and scoring criteria (in single line)
[BASIC]...
[BASIC]...
[ADVANCED]...
[ADVANCED]...
[ADVANCED]...
[DIFFICULT]...
[DIFFICULT]...
[DIFFICULT]...
```

Figure 7: Evaluation Prompt

You are an expert roleplay critic. Your task is to evaluate a character's roleplay performance based on their thinking process and final response.

Input Context

```
<character>
{character}
</character>

<scenario>
{scenario}
</scenario>

<instruction>
{instruction}
</instruction>
```

Content for Evaluation

```
<thinking_process>
{think}
</thinking_process>

<response>
{answer}
</response>

# Evaluation Criteria
{criteria}
```

Task Description

Provide a detailed critique of the roleplay performance. Your evaluation should:

1. Critical analysis of thinking process
2. Alignment between thinking and response
3. Detailed scoring against criteria

Output Format

```
<overall_flaws>
Based on the given evaluation criteria, assess the role-playing ability
according to the character's thought process and final response.
</overall_flaws>

<criteria_evaluation>
For each criterion, output (one per line) without explanation:
Criteria Name(without level): an int score (1-5)
</criteria_evaluation>
```

Figure 8: Guided Thought Generation Prompt

You are an expert roleplay mentor, specializing in character psychological development. Your task is to provide thought-provoking guidance that deepens the character's introspection process.

Input Context

```
<character>
{character}
</character>

<scenario>
{scenario}
</scenario>

<instruction>
{instruction}
</instruction>
```

Current Mental State

```
<thinking_process>
{think}
</thinking_process>
```

Performance Analysis

```
<evaluation_criteria>
{criteria}
</evaluation_criteria>

<identified_gaps>
{overall_flaws}
</identified_gaps>

<detailed_evaluation>
{criteria_evaluation}
</detailed_evaluation>
```

Task

Based on the previous evaluation, provide a continuation of the existing thinking process to guide deeper character thinking. Your suggestion should flow naturally from the previous thoughts and lead to deeper insights to improvement. Keep it short and concise, keep guidance open-ended to encourage intellectual exploration.

Output Format

```
<introspective_guidance>
... A concise turning point in thinking ...
</introspective_guidance>
```

Note: This guidance will be directly appended to the original thinking process, serving as an inner voice that nurtures self-reflection. These prompts will serve as stepping stones for further thought development.

Models	Individual Level						Group Level			Avg
	SA Style	SA Know.	EP Situ.	EP Emo.	CM Short	CM Long	Pos.	Neu.	Neg.	
R-CHAR w/o HS	67.22	83.22	<u>60.51</u>	**36.54**	60.46	64.80	<u>77.35</u>	61.59	53.24	62.77
R-CHAR w/o AE	<u>68.65</u>	<u>84.96</u>	59.82	<u>35.26</u>	61.13	<u>65.23</u>	75.60	<u>63.96</u>	<u>53.80</u>	<u>63.16</u>
R-CHAR w/o TGR	65.24	83.26	**62.17**	32.56	<u>61.63</u>	42.78	74.12	**67.96**	**58.77**	60.94
R-CHAR (Full)	**69.97**	**86.36**	57.93	30.62	**62.49**	**68.58**	**83.90**	63.24	53.47	**64.06**

Table 10: Ablation Study Results on SocialBench (accuracy %)