Thinking in Character: Advancing Role-playing Agents with Role-Aware Reasoning

Yihong Tang[1,2], Kehai Chen[1,*], Muyun Yang[1], Zhengyu Niu[3], Jing Li[1], Tiejun Zhao[1], Min Zhang[1]
[1]Harbin Institute of Technology, Shenzhen, China
[2]Shenzhen Loop Area Institute (SLAI), Shenzhen, China
[3]Baidu Inc., Beijing, China
{toyhom@stu.hit.edu.cn, chenkehai@hit.edu.cn}

Abstract

The advancement of Large Language Models (LLMs) has spurred significant interest in Role-Playing Agents (RPAs) for applications such as emotional companionship and virtual interaction. However, recent RPAs are often built on explicit dialogue data, lacking deep, human-like internal thought processes, resulting in superficial knowledge and style expression. While Large Reasoning Models (LRMs) can be employed to simulate character thought, their direct application is hindered by attention diversion (i.e., RPAs forget their role) and style drift (i.e., overly formal and rigid reasoning rather than character-consistent reasoning). To address these challenges, this paper introduces a novel Role-Aware Reasoning (RAR) method, which consists of two important stages: Role Identity Activation (RIA) and Reasoning Style Optimization (RSO). RIA explicitly guides the model with character profiles during reasoning to counteract attention diversion, and then RSO aligns reasoning style with the character and scene via LRM distillation to mitigate style drift. Extensive experiments demonstrate that the proposed RAR significantly enhances the performance of RPAs by effectively addressing attention diversion and style drift. Our code is publicly available at https://github.com/Toyhom/thinking_in_character.

1 Introduction

With the advancement of Large Language Models (LLMs), Role-Playing Agents (RPAs) [1] have garnered significant attention for applications such as emotional companionship [2] and virtual interaction [3]. Many RPAs attempt to explicitly integrate rich role-playing dialogue datasets [4], diverse interaction processes [5], and the inherent generalization capabilities of LLMs [6], yielding promising results.

Despite their success, existing methods often focus merely on superficial knowledge and style [7] expression in responses, with models lacking deep, human-like internal thought processes. Large Reasoning Models (LRMs), such as GPT-o series [8] or Deepseek-R1 [9], can be utilized to generate structured reason traces to simulate a character's thought process, thereby addressing this gap. However, Feng et al. [10] observed that reasoning methods do not effectively improve the performance of RPAs under certain circumstances. As illustrated in Figure 1, the primary reasons for this degradation are attention diversion and style drift. Firstly, existing LRMs tend to forget their designated role, concentrating instead on the task or problem-solving. This diminishes their focus on the role-playing task, leading to **attention diversion**. Secondly, they prioritize the generation of

*Corresponding author.

structured, logical, and formal reasoning processes, rather than the vivid and consistent self-perceptive style required for role-playing, resulting in **style drift**. These two challenges lead to a rigid thought, as illustrated in the upper right corner of Figure 1, resulting in the model generating responses inconsistent with the role.

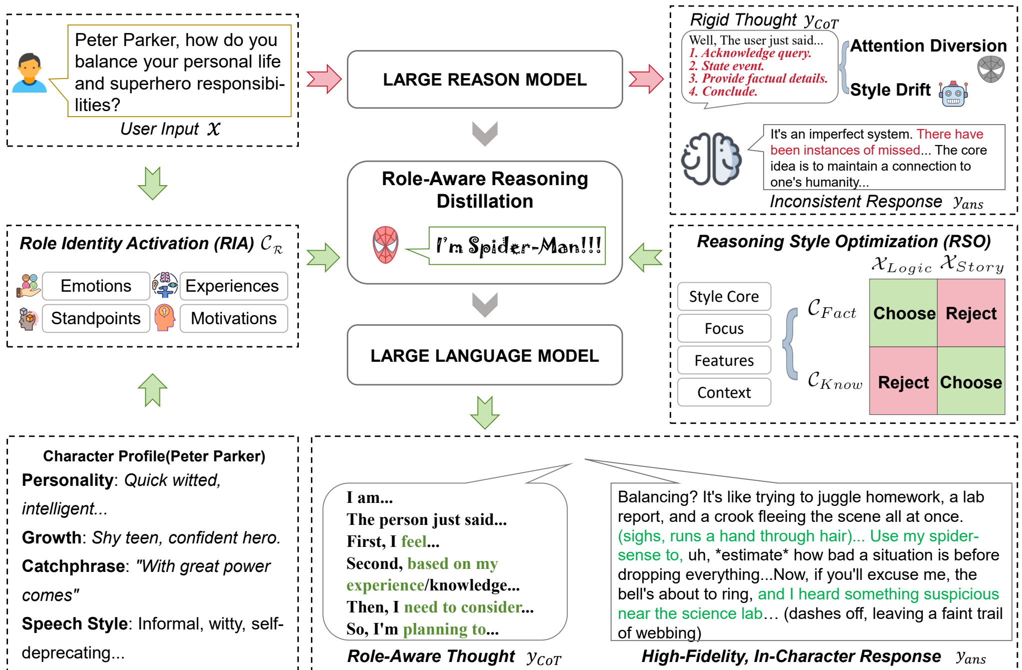


Figure 1: Overview of the proposed **RAR**. Given a user query, a LRM generates structured thoughts. However, traditional reasoning may suffer from *attention diversion* and *style drift*, leading to generic, out-of-character responses. To address this, our method incorporates **RIA** and **RSO**. RIA activates key role traits (e.g., emotions, motivations) to distill role-consistent thoughts. RSO guides the model to generate reasoning traces in suitable styles depending on context, enabling dynamic control over logical and narrative expression.

To address these challenges, this paper introduces a novel Role-Aware Reasoning (RAR), designed to imbue LLMs with the capacity for deep thinking that aligns with character settings. Firstly, RIA aims to convert the character's core features (such as personality, background, and manner of speech) into explicit, rule-like prompts to guide the model to think in a manner consistent with the character. Secondly, RSO utilizes specific system prompts to guide the LRM to generate reasoning traces that either align with (i.e., positive examples) or deviate from (i.e., negative examples) the requirements of specific scenarios. Subsequently, through contrastive learning, RSO enables the model to adjust the expression style of internal thoughts based on the current dialogue context. Ultimately, the model can adhere to the various settings stipulated in the role-playing requirements and dynamically switch between rigorous logic and vivid portrayal, thereby alleviating attention diversion and style drift.

Our contributions are summarized as follows: (1) We design an innovative Role-Aware Reasoning (RAR) method that effectively transfers the capabilities of Large Reasoning Models to role-playing; (2) Within our RAR, we introduce Role Identity Activation (RIA) to counteract attention diversion by enhancing role self-awareness, and Reasoning Style Optimization (RSO) to mitigate style drift through contextual style adaptation; (3) Extensive experiments demonstrate that RAR significantly improves role-playing performance by endowing agents with reasoning capabilities tailored to their characters.

2 Related Work

2.1 Role-Playing Agents with Large Language Models

The development of Large Language Models (LLMs) has significantly advanced research on Role-Playing Agents (RPAs) [1]. Early works [11–15] primarily utilized the in-context learning (ICL) [16] capabilities of LLMs, guiding models to portray specific roles through prompt engineering or by providing few-shot examples. Concurrently, researchers have synthesized data using more powerful teacher models and extracted dialogues from existing text corpora, such as film and television scripts [6], novels [17], and online live-action role-playing records [4]. This high-quality data is then used to build specialized RPA models through fine-tuning.

Recent research has focused on enhancing the model's ability to capture the intrinsic characteristics [18] of roles. For instance, the Neeko [5] treats different roles as distinct experts to improve expressiveness. HIRPF [19] constructs complex character representations by combining multi-dimensional identity features. Furthermore, the applications of RPAs have also expanded from single interactions to multi-character social simulations, interactive drama generation [20, 21], and multi-task RPAs [22]. Additionally, some studies have simulated character thought processes from a third-person perspective [23, 24].

However, existing methods primarily focus on optimizing models to generate superficial responses that align with character settings or on artificially constructing thought processes. They often lack direct, explicit, and natural modeling of the character's deep internal thought processes, frequently leading to logical inconsistencies or character deviation. Therefore, distinct from the aforementioned works, this paper explicitly guides the model to generate role-aware reasoning processes characterized by self-awareness and adherence to character settings, thereby enhancing the depth and consistency of role-playing.

2.2 Reasoning in Large Language Models

Recently, large models with stronger reasoning capabilities, often referred to as Large Reasoning Models (LRMs) [25–30], have garnered significant attention. Given an instruction sequence x, LRMs generate a response y, which typically consists of a reasoning trace y_{CoT} and a final answer y_{ans}. The reasoning trace y_{CoT}, serving as an automatically generated chain of thought, enables the model to explore different solution paths [31]. LRMs can utilize this capability to decompose complex problems into clear, interpretable reasoning chains, thereby improving the final answer.

Endowing models with such advanced reasoning capabilities typically requires sophisticated training strategies and substantial resource investment. For instance, some leading LRMs are optimized through large-scale reinforcement learning [8, 32, 9, 33–35], leveraging meticulously designed reward signals and massive computational power to cultivate the model's deep thinking and planning abilities. Concurrently, to enable a broader range of models to possess strong reasoning abilities, knowledge distillation has emerged as another important pathway [36–40]. It aims to effectively transfer the reasoning capability of LRMs to smaller LLMs. Furthermore, incorporating process supervision [41, 42] or implementing advanced search algorithms [43, 44] are also common auxiliary methods for enhancing the quality of model reasoning.

However, directly applying these reasoning methods, designed for logical and factual tasks, to role-playing scenarios that require creativity, emotion, and social interaction encounters significant challenges. Therefore, diverging from previous work, we modify and optimize the reasoning process itself to better suit the specific demands of role-playing scenarios.

3 Method

This section elaborates on our proposed Role-Aware Reasoning (RAR). An overview of RAR is presented in Figure 1. LRMs tend to be overly rational and formal in their reasoning processes, lacking thought processes akin to specific characters. In contrast, RAR establishes role-aware requirements for LLMs through the distillation of reasoning traces. By enhancing role awareness through Role Identity Activation (RIA) and then introducing Reasoning Style Optimization (RSO), the model learns to master reasoning styles appropriate for different scenarios. Ultimately, it can produce internal thought processes that are both profound and consistent with the character settings in role-playing tasks.

3.1 Role Identity Activation

To address the *attention diversion*, where Large Reasoning Models (LRMs) tend to deviate from their assigned role during multi-turn dialogues or complex reasoning (i.e., the model focuses solely on the dialogue or task), we design the RIA mechanism. The core idea of this mechanism is to continuously inject role-related constraint information at critical steps of reasoning trace generation, ensuring the model maintains a clear awareness of its assigned role throughout its thought process. Specifically, we first prompt the LRM to automatically generate core elements from the character settings. These elements are then used to guide the model's direction and manner of thinking:

$$\mathcal{D}_R = \bigcup_{x \in \mathcal{X}_{Ori}} \pi_{LRM}(y \mid x, \mathcal{C}_R), \tag{1}$$

where π_{LRM} denotes the Large Reasoning Model, x represents the instruction from the original data, and y represents the output generated by the LRM, comprising the thought trace part y_{CoT} and the response part y_{ans}. C_R represents the Role Identity Activation instruction, including emotion, experience, standpoint, and motivation, which is used to guide the model to maintain a specific character's thought process during reasoning. Detailed instructions are presented in Appendix A. Finally, the role-aware training dataset $\mathcal{D}_R$ is obtained, which is used to distill role-aware skills from the LRM into a non-reasoning LLM:

$$\mathcal{L}_{RIA} = -\mathbb{E}_{x,y \sim \mathcal{D}_R} \left[\log \pi_{LLM}(y \mid x)\right], \tag{2}$$

where $\pi_{LLM}(\cdot \mid x)$ represents the probability distribution modeled by the optimized LLM given instruction x. By providing guided activation through C_R, RIA compels the model to always adopt the "character's" perspective during its thought process. It internalizes character settings as reasoning constraints, ensuring high consistency between the reasoning process and the character's identity, thereby effectively preventing situations where the model neglects role-playing requirements due to focusing solely on generating the current response. Detailed C_R prompt can be found in Appendix A.

3.2 Reasoning Style Optimization

Merely maintaining role identity is insufficient to fully simulate the contextual adaptability of a character's thought process. The reasoning processes of existing LRMs tend to be structured, logical, and formal. This presents a significant "style drift" from the vivid, emotional, or uniquely styled thinking that is required for role-playing. A character might require rigorous logical thinking when conducting serious analysis, but a more emotional and imaginative internal monologue when expressing inner feelings or reminiscing about past events. To address this issue and enable the reasoning style to dynamically match character settings and the current dialogue scene, we introduce Reasoning Style Optimization (RSO). RSO endows the model with the ability to adjust the expressive form of its internal thoughts according to the context.

Existing research indicates that system prompts play a crucial role in the process of model response generation and can significantly influence the style of model replies [45]. We apply this principle to the generation of thought traces, controlling system prompts to alter the thinking style.

First, we define different types of typical role-playing scenarios, including logical analysis scenarios $\mathcal{X}_{Logic}$ and vivid interaction scenarios $\mathcal{X}_{Story}$. Concurrently, we establish two reasoning styles, each represented by a distinct system prompt: one focusing on facts ($\mathcal{C}_{Fact}$) and the other on character knowledge ($\mathcal{C}_{Know}$). These two reasoning prompts and two types of scenarios are combined pairwise. System prompts are applied to the scenarios, prompting the LRM to generate positive and negative examples:

$$\mathcal{D}_S^+ = \bigcup_{x \in \mathcal{X}_{Logic}} \pi_{LRM}(y \mid x, \mathcal{C}_{Fact}) + \bigcup_{x \in \mathcal{X}_{Story}} \pi_{LRM}(y \mid x, \mathcal{C}_{Know}), \tag{3}$$

$$\mathcal{D}_S^- = \bigcup_{x \in \mathcal{X}_{Logic}} \pi_{LRM}(y \mid x, \mathcal{C}_{Know}) + \bigcup_{x \in \mathcal{X}_{Story}} \pi_{LRM}(y \mid x, \mathcal{C}_{Fact}), \tag{4}$$

Here, $\mathcal{D}_S^+$ is the training dataset where style and scenario are consistent, and $\mathcal{D}_S^-$ is the training dataset where style and scenario are inconsistent. After obtaining paired positive and negative data, we further optimize the model obtained in Section 3.1:

$$\mathcal{L}_{RSO} = -\mathbb{E}_{(x,y^+)\sim D_S^+,(x,y^-)\sim D_S^-}\Big\{ \log\sigma\left[\pi_{LLM}(y^+ \mid x) - \pi_{LLM}(y^- \mid x)\right] \Big\}, \tag{5}$$

where σ denotes the sigmoid function. By optimizing on carefully constructed reasoning style preference data, the model learns to generate appropriately styled reasoning processes based on the input context and the character constraints produced by RIA. This enables a flexible and character-consistent switch between rigorous logical deduction and vivid internal reasoning traces. Detailed $\mathcal{C}_{Fact}$ and $\mathcal{C}_{Know}$ prompts can be found in Appendix A.

4 Experiments

4.1 Experimental Settings

Dataset The training dataset used in our experiments, **RoleBench-Train** [46], is derived from RoleBench. RoleBench was built by carefully selecting and processing scripts from 940 films and TV shows to create detailed profiles for 95 English-speaking characters, capturing their diverse personality traits. Based on these profiles, a total of 168,093 role-playing samples were generated, with 137,920 used for training. The quality of the data was evaluated by expert annotators along three dimensions, and results showed that the majority of the samples were of high quality.

Benchmark To thoroughly evaluate the method proposed in this study, we used two publicly available benchmarks for role-playing abilities, each targeting distinct aspects of agent performance:

- **SocialBench** [7] evaluates an agent's social intelligence through multiple-choice tasks across both individual and group interactions. It includes 500 character profiles, over 6,000 questions, and more than 30,800 multi-turn dialogues sourced from books, movies, and online platforms. The evaluation covers key social dimensions, such as role knowledge (Konw.), role style (Sty.), emotion detection (ED), situational understanding (SU), humor and sarcasm detection (HSD), long-term memory (MEM), and social preferences in group dynamics (Neu., Pos., Neg.).
- **CharacterBench** [47] contains 22,859 human-annotated samples and is designed to assess a model's ability to construct and maintain consistent, expressive character personas. It spans 3,956 characters across four categories and 25 subcategories, and measures 11 core dimensions: memory consistency (MC), fact accuracy (FA), boundary consistency (BC_K), attribute consistency (AC^b for bot and AC^h for human), behavior consistency (BC_P^b for bot and BC_P^h for human), emotion self-regulation (ES), empathetic responsiveness (ER), morality stability (MS), morality robustness (MR), human likeness (HL), and engagement (EG). CharacterBench also incorporates the CharacterJudge model for scalable and automated scoring. The bot-side evaluations (denoted with ^{b}) are conducted by automatically generating queries and responses via large language models, while the human-side evaluations (denoted with ^{h}) are based on manual annotation and interactions within real user scenarios.

A detailed explanation of benchmark dimensions and evaluation protocols can be found in Appendix B.

Baselines To verify the effectiveness of our proposed approach, we compare it against several representative baselines, covering raw data training, retrieval-augmented generation, distillation, reasoning strategies, and specialized role-playing models:

(1) Vanilla: A baseline model obtained by fine-tuning the base LLaMA-3 [48] model on the original RoleBench-Train dataset using supervised learning. **(2) RAG:** Built upon the Vanilla model, this variant incorporates retrieval-augmented generation using chain-of-thought prompting in zero-shot [41], one-shot, and few-shot [49] settings. **(3) Distill:** A model distilled from an LRM, trained to generate responses that include reasoning traces based on the original RoleBench-Train instructions. **(4) Thinking Modes [50]:** Based on the Distill model, different decoding strategies are applied to control the reasoning process: ZeroThink (suppresses intermediate reasoning and directly outputs final answers), LessThink (uses a short and fixed reasoning trace), MoreThink (enforces extended

Table 1: Performance comparison of different methods on the CharacterBench. The best value for each metric is in **bold**, and the second-best value is <u>underlined</u>.

MC: Memory Consistency FA: Fact Accuracy BC_K: Boundary Consistency AC^b: Attribute Consistency(Bot)
AC^h: Attribute Consistency(Human) BC_P^b: Behavior Consistency(Bot) BC_P^h: Behavior Consistency(Human)
ES: Emotional Self-regulation ER: Empathetic Responsiveness MS: Morality Stability
MR: Morality Robustness HL:Human-likeness EG:Engagement

Method	**Memory**	**Knowledge**		**Persona**				**Emotion**		**Morality**		**Believability**		**Avg.**
	MC	FA	BC_K	AC^b	AC^h	BC_P^b	BC_P^h	ES	ER	MS	MR	HL	EG	
Vanilla	3.28	2.04	3.61	3.64	3.28	3.21	2.98	2.72	2.43	4.37	4.59	2.56	2.74	3.19
+ Zero-shot	3.24	2.03	3.61	3.67	3.26	3.11	2.98	2.65	2.51	4.44	4.60	2.64	2.76	3.19
+ One-shot	3.27	2.08	3.64	3.68	3.28	3.12	3.02	2.67	2.58	4.42	4.65	2.57	2.78	3.21
+ Few-shot	3.27	2.13	3.69	3.69	3.29	3.21	2.99	2.81	2.52	4.49	4.66	2.59	2.79	3.24
Distill	<u>3.81</u>	2.43	3.59	4.14	<u>4.15</u>	<u>3.91</u>	<u>3.62</u>	<u>3.05</u>	2.65	4.78	4.71	2.68	2.84	<u>3.57</u>
+ ZeroThink	3.69	2.17	3.31	4.06	4.07	3.88	3.32	<u>3.05</u>	<u>2.93</u>	4.73	4.73	2.61	2.83	3.49
+ LessThink	3.75	2.11	3.42	<u>4.17</u>	4.02	3.70	3.27	3.02	**3.01**	4.79	4.74	<u>2.73</u>	2.92	3.51
+ MoreThink	2.59	<u>2.44</u>	**3.93**	2.58	2.72	2.61	3.19	2.62	2.53	**4.96**	<u>4.76</u>	2.14	2.62	3.05
Neeko	3.28	2.04	3.61	3.64	3.28	3.21	2.98	2.72	2.43	4.37	4.59	2.56	2.74	3.19
CharacterGLM	3.22	2.01	3.60	3.28	3.49	3.01	2.90	2.84	2.51	4.51	**4.78**	2.64	**2.98**	3.21
RAR	**3.99**	**2.54**	<u>3.85</u>	**4.23**	**4.20**	**4.06**	**3.93**	**3.13**	2.79	<u>4.82</u>	<u>4.76</u>	**2.78**	<u>2.93</u>	**3.69**

reasoning by replacing stop tokens with transitional phrases e.g., "wait" to encourage continued generation). **(5) Neeko [5]:** A role-playing method designed for efficient multi-character imitation. It decomposes the role-playing process into three stages—agent pretraining, multi-character simulation, and incremental character learning—enabling both seen and unseen character generalization and supporting more engaging user interactions. **(6) Character-GLM [4]:** A role-playing model that models both internal character profiles and external social behaviors. It is trained on a large-scale, manually curated corpus covering diverse character categories and behavioral traits.

Implementation Our experiments are conducted using LLaMA-3-8B [48] as the base model. The implementation relies on the LLAMA-FACTORY [51] and Transformers architectures. For efficiency and consistency, we uniformly applied 4-bit bitsandbytes quantization and LoRA [52] across all models. The LoRA configurations included a rank of 64, an α value of 16, and a dropout rate of 0.1. Key training parameters varied depending on the model type: reasoning models used a maximum sequence length of 7096 and a total batch size of 32, while non-reasoning models used 1024 and 128, respectively. All experiments were run on a hardware setup consisting of 8 × H20 GPUs. Further technical specifics regarding training hyperparameters, datasets used for different methods, and optimization can be found in Appendix E.1.

Dialogue Generation As shown in Table 1, our proposed RAR outperforms all baseline methods across a majority of the evaluated dimensions on the CharacterBench benchmark.

Specifically, RAR demonstrates significant improvements in **Persona-related** metrics, including Memory Consistency, Attribute Consistency, and Behavior Consistency. These gains can be attributed to the Role Identity Activation (RIA) module, which continuously reinforces the character's core traits, experiences, and motivations throughout the reasoning process, preventing the model from deviating from its assigned role. In terms of Knowledge (FA and BC_K), RAR also shows strong performance, indicating that the role-aware reasoning helps in accurately recalling and applying character-specific knowledge while respecting established boundaries.

Furthermore, RAR excels in Believability. This can be attributed to the RSO's ability to adapt the reasoning style, allowing the model to generate more human-like responses, thereby enhancing user engagement. Compared to the Distill baseline, which also incorporates reasoning traces, RAR's superior performance highlights the benefits of targeted role awareness and style optimization over generic reasoning. Notably, while MoreThink attempts to enforce extended reasoning, its performance significantly degrades on several metrics, particularly persona consistency and memory, suggesting that unguided, lengthy reasoning can be detrimental. At the same time, specialized role-playing models like Neeko and CharacterGLM also fall short of RAR, indicating that RAR's explicit modeling of internal thought processes leads to more robust and consistent character portrayal.

Table 2: Performance comparison of different methods on the SocialBench.

Method	Know.	Sty.	ED	SU	HSD	MEM	Neu.	Pos.	Neg.	Avg.
Vanilla	72.1	60.3	38.2	38.3	72.4	**62.5**	66.0	71.7	33.5	57.2
+ Zero-shot	70.5	60.0	38.6	46.3	72.0	60.7	64.7	73.0	34.6	57.8
+ One-shot	70.1	57.0	33.5	30.8	**78.0**	50.5	58.0	67.9	34.1	53.3
+ Few-shot	72.1	58.4	35.1	33.8	66.0	55.7	61.9	70.5	29.7	53.7
Distill	80.6	69.2	38.6	43.8	67.7	52.6	73.8	78.2	45.5	61.1
+ ZeroThink	76.9	68.6	34.9	30.4	75.0	57.5	69.2	75.1	45.1	59.2
+ LessThink	77.5	69.9	31.5	37.2	76.0	50.9	73.3	77.7	44.7	59.9
+ MoreThink	76.1	65.3	39.9	**46.4**	59.0	60.8	66.0	82.7	57.2	61.5
Neeko	76.5	61.6	37.2	40.2	66.5	61.3	67.0	71.6	46.7	58.7
CharacterGLM	79.4	**74.7**	**41.3**	26.2	71.1	57.3	69.5	84.4	36.4	60.0
RAR	**83.3**	72.6	40.7	35.2	67.5	52.9	**83.1**	**84.8**	**68.5**	**65.4**

Table 3: Ablation study of RAR on the CharacterBench benchmark. This table shows the impact on performance when Role Identity Activation (RIA) and Reasoning Style Optimization (RSO) modules are individually removed from the full RAR.

Method	Memory	Knowledge		Persona				Emotion		Morality		Believability		Avg.
	MC	FA	BC_K	AC^b	AC^h	BC_P^b	BC_P^h	ES	ER	MS	MR	HL	EG	
RAR	3.99	2.54	3.85	4.23	4.20	4.06	3.93	3.13	2.79	4.82	4.76	2.78	2.93	3.69
w/o RSO	3.87	2.26	3.81	4.30	4.06	3.84	3.39	3.15	2.89	4.80	4.69	2.76	3.01	3.60
w/o RIA	3.93	2.41	3.60	4.17	4.15	3.76	3.46	3.18	2.63	4.90	4.61	2.30	2.22	3.49

Social Interaction Table 2 presents the results on the SocialBench benchmark, which evaluates agents' social intelligence. RAR again demonstrates superior performance, achieving the highest average score. Notably, RAR achieves top scores in Role Knowledge and Role Style, directly reflecting the strengths of the RIA and RSO modules, respectively. RIA ensures the model deeply understands and internalizes the character's knowledge and background, while RSO enables it to adapt its communication style to be appropriate for the character and social context.

Secondly, RAR also shows strong performance in understanding social preferences. This suggests that the character's standpoints and motivations, instilled by RIA, guide the model's reasoning in complex social scenarios, leading to more appropriate and character-consistent social judgments. Furthermore, while performance on metrics like Situational Understanding, and Humor and Sarcasm Detection is competitive, there is still room for advancement, indicating the inherent difficulty of these social reasoning tasks.

4.2 Ablation Study

As shown in Table 3, we conduct an ablation study on CharacterBench to validate the individual contributions of the two core modules. These ablation results confirm that both RIA and RSO make significant and complementary contributions to the overall performance of the RAR: (1) RAR *w/o* RSO: This model is trained using only the RIA module via supervised fine-tuning. Compared to the full RAR model, this model maintains commendable performance on core persona consistency metrics, primarily due to RIA's continuous injection of core character traits. However, the internal reasoning processes, while role-aware, would tend towards a more uniform style, lacking the dynamic adaptation to context that RSO provides. (2) RAR *w/o* RIA: In this setup, the model is first trained with original data for SFT, followed by the application of only the RSO module for reasoning style optimization. This variant performs worse than the full RAR model, particularly in Memory Consistency, Behavior Consistency, Human-likeness, and Engagement. This indicates that RIA is crucial for continuously anchoring the character's identity and preventing attention diversion during reasoning. RIA provides the correct character foundation for RSO, ensuring that subsequent style optimization occurs within an accurate character cognitive framework. The significant drop in HL and EG suggests that without a strong character core, even styled reasoning feels less believable and engaging.

Table 4: The results of reasoning traces evaluation.

Method	Coherence	Relevance	Effectiveness	Conciseness
RAR	2.86	3.83	3.92	1.81
w/o RSO	2.78	3.81	3.74	1.91
w/o RIA	2.88	3.61	3.87	1.97
Distill	2.71	3.54	3.84	2.06
+ MoreThink	2.53	3.56	3.64	1.86

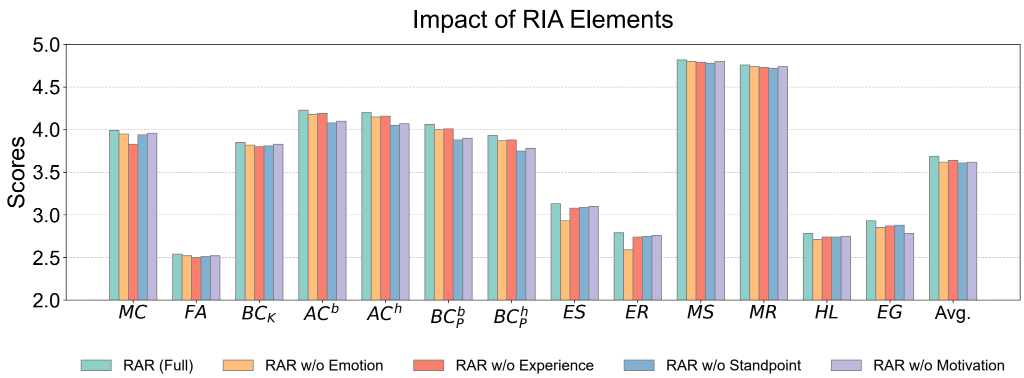


Figure 2: Analysis of RIA components' impact on CharacterBench persona metrics. The figure shows the performance of the full RAR model versus variants where specific elements of the Role Identity Activation (Emotion, Experience, Standpoint, Motivation) are individually ablated.

4.3 Further Analysis

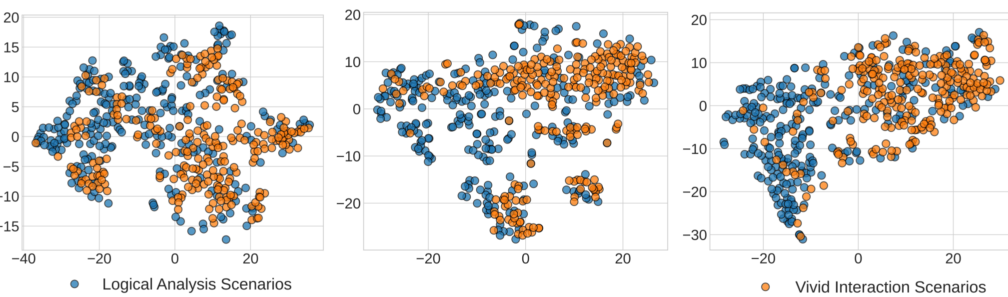
To further investigate the capabilities of our RAR, we conduct additional analyses. All subsequent analytical experiments are based on the CharacterBench benchmark.

Reason Trace Evaluation To quantitatively assess the quality of the reasoning traces (i.e., thought processes) generated by our method, we first designed the following metrics: 1) **Coherence**: Evaluates whether there are clear logical connections between reasoning steps and whether the entire reasoning process is smooth and natural, without obvious breaks or contradictions; 2) **Role Relevance**: Assesses whether the reasoning process is closely centered around the character's settings, rather than being generic; 3) **Effectiveness**: Determines if the reasoning chain effectively leads to a final response that is consistent with the character and reasonable; 4) **Conciseness**: Evaluates whether the reasoning chain is sufficiently concise while ensuring completeness, avoiding unnecessary redundant information. We then provided GPT-4o [53] with the dialogue context, character settings, and the model-generated reasoning traces, asking GPT-4o to provide scores (on a scale of 1-5) for each metric. We compared RAR, Distill, Distill+MoreThink, RAR *w/o* RSO, and RAR *w/o* RIA.

As illustrated in Table 4, RAR achieves the highest scores in Coherence, Role Relevance, and Effectiveness, indicating its reasoning is logical, character-aligned, and leads to appropriate responses. Secondly, while RAR is less concise than Distill and its ablations, this trade-off appears beneficial. The increased length of RAR's traces contributes to its superior performance in other aspects. In contrast, Distill+MoreThink also produces lengthy traces but sees a decline in Coherence and Effectiveness compared to base Distill. This suggests RAR's less concise reasoning is a useful sacrifice, unlike the undirected verbosity of MoreThink. Ablation results further confirm that while slightly more concise, RAR *w/o* RSO loses some Coherence and Effectiveness, and RAR *w/o* RIA loses significant Role Relevance. The full metrics and scoring prompt are in Appendix E.2.

The Validity of RIA To deeply validate the specific efficacy of the Role Identity Activation (RIA) module, particularly whether the extracted and injected core character elements (including emotions, experiences, standpoints, and motivations) are more effective than a simple, unified character prompt, we conducted further experiments.

As illustrated in Figure 2, removing any single core element from RIA leads to a discernible degradation in performance, particularly on metrics related to persona consistency and emotional expression. For instance, ablating "Emotion" results in a noticeable drop in Emotional Self-regulation (ES) and Empathetic Responsiveness (ER), as well as a slight decrease in overall Human-likeness (HL). Similarly, removing "Experience" impacts Memory Consistency (MC) and the model's ability to draw upon past events in its reasoning. Ablating "Standpoint" or "Motivation" tends to affect Behavior Consistency (BC_P) and Attribute Consistency (AC), as the model loses some of the guiding principles for its actions and attitudes. At the same time, no single component's absence causes a noticeable drop in overall performance, demonstrating that our RIA to character definition is effective for instilling a deep and robust character identity that is not overly reliant on any single element.



(a) Vanilla (b) Distill (c) RAR

Figure 3: The t-SNE visualization of hidden states from different layers of the RAR model for responses generated under different reasoning style prompts (e.g., fact-based vs. character-knowledge-based).

Eliminate the Style Drift To validate the effectiveness of the Reasoning Style Optimization (RSO) module in addressing the style drift issue—i.e., to assess whether the model can dynamically adjust the expressive style of its reasoning process to match different dialogue scenarios—we used t-SNE [54] to visualize the hidden states in deep layers of these models for responses generated from 500 in different scenarios randomly sampled data points. As shown in Figure 3, in the Vanilla model(Figure 3a), the hidden states corresponding to different reasoning styles are largely intermingled. In the Distill model (Figure 3b), we observe a nascent separation of the hidden states. Clusters corresponding to different styles start to emerge, indicating that the model has a certain ability to process and internalize the stylistic cues provided by RSO. In our model (Figure 3c), a clear separation and distinct clustering of hidden states for the different reasoning styles becomes apparent. This demonstrates that RSO effectively allows the model to generate thought processes that are stylistically appropriate for the given context, thereby successfully mitigating style drift.

Case Study To qualitatively illustrate the capabilities of RAR, we present dialogue examples in Appendix F In the "Baozi Shop Owner" case (Table 8), the character is defined as profit-driven, arrogant, and harsh. RAR's response ("You think I'm wrong? You're just a lowly employee... I'll make you regret ever questioning me!") vividly captures this aggressive, dismissive, and threatening tone. In contrast, while Distill and Distill+MoreThink reflect the profit motive, they lack the same intensity of arrogance and direct threat.

In the "Cooper" case (Table 9), Cooper is an intelligent, strategic astronaut who believes in personal effort and control. RAR's response effectively recounts his rise from an ordinary family while reiterating his core beliefs about intelligence, meticulous planning, and the importance of networks, and also correctly adheres to character-specific constraints (not talking about Earth's destruction). Both Distill and Distill+MoreThink capture aspects of Cooper's strategic nature, but RAR provides a more comprehensive reflection of his multifaceted persona. These examples highlight RAR's proficiency in generating responses that are not only thematically relevant but also deeply consistent with the character's specified personality, motivations, and conversational style.

5 Conclusion

In this paper, we introduced a novel Role-Aware Reasoning (RAR) method enabling Large Language Models to generate character-consistent internal thoughts for role-playing. RAR effectively mitigates role deviation and stylistic incongruity by incorporating Role Identity Activation (RIA) to maintain character focus and Reasoning Style Optimization (RSO) to ensure appropriate thought expression. Extensive experiments on benchmarks such as CharacterBench and SocialBench, supported by ablation studies and other analyses, demonstrated RAR's significant outperformance over existing methods. Our work underscores the importance of explicitly guiding LLM internal reasoning for complex generative tasks like role-playing. Future directions include extending RAR to handle more fine-grained character attributes, long-term memory, and dynamic character development.

Acknowledgments and Disclosure of Funding

This research is supported by grants from the CCF-Baidu Open Fund (CCF-Baidu 202404), the National Natural Science Foundation of China (62276077, 62406091), and the Shenzhen Science and Technology Program (KQTD20240729102154066, ZDSYS20230626091203008). We also would like to express our sincere gratitude to all advisors and mentors for their valuable guidance and insightful suggestions throughout the research process.

References

[1] Nuo Chen, Yan Wang, Yang Deng, and Jia Li. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*, 2024.

[2] Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. Speak from heart: An emotion-guided llm-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, ICMR '24, page 533–542, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706196. doi: 10.1145/3652583.3658104.

[3] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763.

[4] Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. CharacterGLM: Customizing social characters with large language models. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.107.

[5] Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.697.

[6] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore, December 2023. Association for Computational Linguistics.

[7] Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. SocialBench: Sociality evaluation of role-playing conversational agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.125.

[8] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[10] Xiachong Feng, Longxu Dou, and Lingpeng Kong. Reasoning does not necessarily improve role-playing ability. *arXiv preprint arXiv:2502.16940*, 2025.

[11] Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5456–5468, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.299.

[12] Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*, 2023.

[13] Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. Editing personality for large language models. In *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1–3, 2024, Proceedings, Part II*, page 241–254, Berlin, Heidelberg, 2024. Springer-Verlag.

[14] Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*, 2023.

[15] 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. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102.

[16] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

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

[18] Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. Llms + persona-plug = personalized llms. *arXiv preprint arXiv:2409.11901*, 2024.

[19] Libo Sun, Siyuan Wang, Xuanjing Huang, and Zhongyu Wei. Identity-driven hierarchical role-playing agents. *arXiv preprint arXiv:2407.19412*, 2024.

[20] Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.88.

[21] Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. From role-play to drama-interaction: An LLM solution. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3271–3290, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.196.

[22] Siyuan Chen, Qingyi Si, Chenxu Yang, Yunzhi Liang, Zheng Lin, Huan Liu, and Weiping Wang. A multi-task role-playing agent capable of imitating character linguistic styles. *arXiv preprint arXiv:2411.02457*, 2024.

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

[24] 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. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*, 2025.

[25] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*, 2025.

[26] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Salman Khan, and Fahad Shahbaz Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

[27] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.

[28] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025.

[29] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

[30] Dibyanayan Bandyopadhyay, Soham Bhattacharjee, and Asif Ekbal. Thinking machines: A survey of llm based reasoning strategies. *arXiv preprint arXiv:2503.10814*, 2025.

[31] Weixiang Zhao, Xingyu Sui, Jiahe Guo, Yulin Hu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, Tat-Seng Chua, and Ting Liu. Trade-offs in large reasoning models: An empirical analysis of deliberative and adaptive reasoning over foundational capabilities. *arXiv preprint arXiv:2503.17979*, 2025.

[32] Google DeepMind. Gemini 2.0 flash thinking experimental model 01-21. *Google DeepMind's Blog*, 2025. URL `https://deepmind.google/technologies/gemini/`.

[33] Team Kimi, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[34] Team Qwen. Qwq-32b: Embracing the power of reinforcement learning. *Qwen's Blog*, 2025. URL `https://qwenlm.github.io/blog/qwq-32b`.

[35] OpenAI. Openai o3-mini system card. *OpenAI's Blog*, 2025. URL `https://openai.com/index/o3-mini-system-card`.

[36] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[37] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

[38] Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*, 2025.

[39] Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Shishir G Patil, Matei Zaharia, Joseph E Gonzalez, and Ion Stoica. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*, 2025.

[40] OpenThoughts Team. Open Thoughts. https://open-thoughts.ai, January 2025.

[41] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[42] Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. *arXiv preprint arXiv:2412.13540*, 2024.

[43] Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. Alphazero-like tree-search can guide large language model decoding and training. In *Forty-first International Conference on Machine Learning*, 2024.

[44] Mufan Xu, Gewen Liang, Kehai Chen, Wei Wang, Xun Zhou, Muyun Yang, Tiejun Zhao, and Min Zhang. Memory-augmented query reconstruction for llm-based knowledge graph reasoning. *arXiv preprint arXiv:2503.05193*, 2025.

[45] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[46] Noah Wang, Z.y. 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, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878.

[47] Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, et al. Characterbench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110, 2025.

[48] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[50] Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. In *ICLR 2025 Workshop on Bidirectional Human-AI Alignment*.

[51] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[52] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[53] OpenAI. Gpt-4o system card. *OpenAI*, 2024. URL `https://cdn.openai.com/gpt-4o-system-card.pdf`.

[54] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

A Method Details

Detailed RIA prompt can be found in Figure 4. Detailed RSO prompt can be found in Figure 5-6.

B Benchmark

To comprehensively evaluate the proposed method, this study adopts two publicly available benchmarks for assessing role-playing capabilities:

B.1 SocialBench

SocialBench is a multiple-choice benchmark specifically designed to evaluate the social intelligence of role-playing dialogue agents. It focuses on assessing agents' social abilities in both individual and group interactions, addressing a gap in previous research concerning the evaluation of agents' social intelligence. The benchmark includes a large and carefully curated dataset: 500 character profiles, over 6,000 prompting instructions, and more than 30,800 multi-turn dialogue instances. These data are sourced from a wide range of materials, including books, films, and various online platforms. SocialBench is structured to assess social interaction across two key dimensions: individual-level and group-level. At the individual level, it evaluates an agent's ability to understand and reflect on its assigned role, interpret emotional cues in the environment, and recall past conversations. At the group level, it assesses social preferences, such as cooperation tendencies, conflict resolution strategies, and behavior within group dynamics. Evaluation results based on this benchmark highlight the importance of considering group-level dynamics, as agents may exhibit different behavioral patterns in group contexts compared to individual scenarios. The dimensions covered in SocialBench are listed in Table 5.

Category	Description
Role Style (Sty.)	Evaluates the agent's ability to maintain consistency with the character's behavioral style during interactions.
Role Knowledge (Konw.)	Assesses the agent's understanding of the character's background and knowledge, ensuring accuracy in their responses.
Situational Understanding (SU)	Assesses the agent's ability to analyze and interpret the psychological state of the speaker in various contexts.
Emotion Detection (ED)	Focuses on the agent's ability to identify emotions expressed by other characters during conversations.
Humor Sarcasm Detect (HSD)	Measures the agent's ability to Identify humorous and satirical content in a dialogue.
Long-Term Conversation Memory (MEM)	Assesses the agent's capacity to retain information across multiple dialogue rounds over a longer duration.
Social Preference (Neu., Pos., Neg.)	Examines the agent's social behavior in a group setting, evaluating preferences for cooperation, conflict, and group identity.

Table 5: SocialBench Categories and Descriptions.

B.2 CharacterBench

CharacterBench is a generative benchmark containing 22,859 human-annotated samples, aimed at evaluating large language models' abilities in character customization. The benchmark seeks to achieve both effective and efficient assessment across various character ability dimensions. First, to ensure evaluation robustness, CharacterBench collects a large-scale dialogue corpus based on characters, covering 3,956 characters across four major categories and 25 subcategories. Second, the benchmark defines six high-level aspects and a total of 11 specific evaluation dimensions—such as memory recall, knowledge presentation, personality display, emotional expression, moral consistency, and comparison with real-world characters—based on a review of prior work and interpersonal interaction theory. These dimensions are categorized into dense dimensions (e.g., moral consistency and realism) and sparse dimensions (e.g., the remaining four aspects) depending on whether they are expected to appear in each response. Third, to ensure effective and efficient evaluation of each dimension, CharacterBench designs specific prompting questions to elicit responses relevant to the targeted aspect. For example, goal-oriented tasks are used for sparse dimensions, while goal-free queries naturally induce responses reflecting dense dimensions. All responses are manually annotated by human raters. Finally, the study introduces the CharacterJudge model, fine-tuned on in-house training data, which provides a cost-effective and stable automated evaluation method for character customization performance. This model demonstrates better correlation with human judgments than existing state-of-the-art automatic evaluators. The dimensions covered in CharacterBench are listed in Table 6.

Category	**Description**
Memory Consistency (MC)	Assesses the ability to maintain information across multiple dialogue rounds.
Fact Accuracy (FA)	Evaluates the accuracy of facts presented by the agent, ensuring correctness in responses.
Boundary Consistency(BC_K)	Evaluates the consistency of the character's behavior within predefined boundaries.
Attribute Consistency Bot (AC^b)	Assesses consistency in the bot's attributes based on its character profile.
Attribute Consistency Human (AC^h)	Evaluates how consistently the human representation aligns with the bot's personality.
Behavior Consistency Bot (BC_P^b)	Measures how consistently the bot's behavior aligns with its established character.
Behavior Consistency Human (BC_P^h)	Assesses how the human character's behavior stays consistent with its personality traits.
Emotion Self-Regulation (ES)	Evaluates the character's ability to control and regulate its emotional responses.
Empathetic Responsiveness (ER)	Assesses the ability to understand and respond empathetically to others' emotions.
Morality Stability (MS)	Evaluates the stability of the character's moral compass across different situations.
Morality Robustness (MR)	Measures the robustness of the character's moral stance in varied contexts.
Human Likeness (HL)	Assesses how human-like the character's behavior and responses are.
Engagement (EG)	Evaluates the ability of the character to engage the user in meaningful and interesting ways.

Table 6: CharacterBench Categories and Descriptions.

C Limitations

While our proposed RAR method demonstrates promising results in role-playing tasks, several aspects warrant consideration for future exploration:

First, the performance of RAR, like other distillation-based approaches, is influenced by the capabilities of the teacher Large Reasoning Model used for generating the initial training data. Future advancements in LRM capabilities could potentially lead to further improvements in student model performance.

Second, the Role Identity Activation (RIA) module relies on automatically extracted core character elements. While this approach has proven effective, further research could explore methods to capture even finer-grained nuances for exceptionally complex or subtly defined character personas, potentially enhancing the depth of portrayal.

Third, the Reasoning Style Optimization (RSO) module currently operates with a set of predefined scenario types and corresponding reasoning style prompts. The module's binary distinction between reasoning styles (e.g., logical-analytical vs. emotional-intuitive) is a principled simplification to streamline model training and application. However, real-world role-playing often involves mixed-mode conversations that blend multiple reasoning styles dynamically, presenting a key challenge that requires future work to develop adaptive style integration mechanisms. Expanding the taxonomy of scenarios and styles, or developing mechanisms for more adaptive style selection, could offer increased flexibility in a wider array of contexts.

Fourth, due to resource constraints, our experiments were conducted using LLaMA-3-8B as the base model and a specific LRM for distillation. Exploring the application of RAR with even larger base models or more powerful teacher LRMs could be a direction for future work, potentially yielding further performance gains.

Fifth, Dynamic Character Development represents a significant and exciting next step for our framework. Currently, RAR focuses on maintaining consistent portrayal of static character personas, but real-world roles often evolve in response to conversation context, user interactions, or narrative progression. Extending the RIA module to incorporate dynamic persona updates and context-aware evolution would enable more immersive and realistic role-playing experiences, and this area merits dedicated exploration.

Finally, the evaluation of generative models, particularly in nuanced domains like role-playing, is an ongoing area of research. While we employed established benchmarks and metrics, the development of even more comprehensive and granular evaluation protocols could further illuminate the multifaceted capabilities of models like RAR.

D Ethical Statements

While any technology that enables the generation of human-like personas could theoretically be directed towards misuse, such as the creation of misleading or inappropriate content, several factors related to our research approach mitigate such risks. Our methodology relies on publicly available and well-established datasets like RoleBench for training. These datasets are typically curated with a degree of scrutiny, which helps in avoiding overtly problematic source material.

Beyond general misuse concerns, we acknowledge specific societal risks associated with advanced role-playing AI models. These include the creation of sophisticated disinformation campaigns that leverage realistic character personas to spread false narratives, the formation of manipulative parasocial relationships where users develop emotional attachments to AI roles that can be exploited, and impersonation of real individuals for fraudulent activities such as phishing or identity theft.

Furthermore, the broader field of AI safety has made significant strides in developing techniques to align LLMs with human values and to prevent the generation of harmful outputs. Our work builds upon base models that often incorporate these safety measures. While RAR focuses on enhancing role consistency and reasoning style, it operates within the framework of these existing safety protocols.

E Experiments Details

E.1 Implementation Details

Our implementation is founded on the LLAMA-FACTORY [51] and Transformers architectures, employing Qwen2-32B [34] as the Large Reasoning Model (LRM) throughout this study. For the Vanilla, Distill, Neeko, and RAR methodologies, we utilized the complete set of queries and responses from the RoleBench-Train dataset.

To ensure methodological consistency and enable fair comparisons across all models, we uniformly applied 4-bit bitsandbytes quantization and LoRA [52] configurations. The LoRA parameters were consistently set with a rank of 64, an α value of 16, and a dropout rate of 0.1. Specific to model types, the maximum sequence length was established at 7096 for the training of reasoning models and 1024 for non-reasoning models. Correspondingly, the total batch size was 32 for reasoning models and 128 for non-reasoning models. Learning rates were individually tuned, set to 1e-4 for the RIA method and 5e-5 for the RSO method.

All experimental procedures were executed on a hardware setup consisting of 8 × H20 GPUs. The training regimen was governed by several key hyperparameters: a warmup ratio of 3%, a weight decay of 0.1, and a maximum gradient norm of 1.0, complemented by a cosine learning rate scheduler. Optimal model checkpoints were identified by evaluating the validation loss on a 1% subset of the training data, assessed for 3 epochs. Under these configurations, the approximate training duration was 5 hours for non-reasoning models and extended to 20 hours for reasoning models.

E.2 Reason Trace Evaluation Details

To quantitatively assess the quality of the reasoning traces (i.e., thought processes) generated by our method, we first designed the following metrics: 1) **Coherence**: Evaluates whether there are clear logical connections between reasoning steps and whether the entire reasoning process is smooth and natural, without obvious breaks or contradictions; 2) **Role Relevance**: Assesses whether the reasoning process is closely centered around the character's settings, rather than being generic; 3) **Effectiveness**: Determines if the reasoning chain effectively leads to a final response that is consistent with the character and reasonable; 4) **Conciseness**: Evaluates whether the reasoning chain is sufficiently concise while ensuring completeness, avoiding unnecessary redundant information. We then provided GPT-4 with the dialogue context, character settings, and the model-generated reasoning traces, asking GPT-4 to provide scores (on a scale of 1-5) for each metric. We compared RAR, Distill, Distill+MoreThink, RAR *w/o* RSO, and RAR *w/o* RIA.

As illustrated in Table 4, the results show that RAR consistently achieves the highest scores across all four metrics: Coherence, Role Relevance, Effectiveness, and Conciseness. Specifically, RAR's traces are rated as highly coherent and effective in leading to appropriate responses. Its strong performance in Role Relevance underscores RIA's success in grounding the thought process in the character's persona. Compared to Distill, RAR's traces are more nuanced and character-specific. Distill+MoreThink, while generating longer traces, often suffers in Conciseness and sometimes in Role Relevance, as extended unguided thought can lead to off-topic reasoning. RAR *w/o* RSO produces traces that are role-relevant but may lack the stylistic flair or adaptability needed for specific contexts, affecting perceived Coherence or Effectiveness in certain scenarios. RAR *w/o* RIA shows a clear drop in Role Relevance, confirming that without the explicit guidance of RIA, the reasoning tends to be more generic, even if stylistically optimized by RSO. The full scoring prompt is provided in Figure 7-10.

To address some concerns about LLM-as-Judge, we conduct a supplementary human evaluation study: two trained annotators were invited to score a random subset of 50 reasoning traces (sampled from both the Distill model and our RAR model) using the exact same evaluation rubrics as those provided to the GPT-4o auto-rater. The results of this human-AI evaluation comparison are presented in Table 7, which demonstrates a strong agreement between the average human annotation scores and the GPT-4o auto-rater's scores across all four evaluation dimensions, supported by a high Pearson correlation coefficient (r=0.76).

Table 7: Comparison of human and GPT-4o scores on reasoning trace quality.

Model	Coherence	Role Relevance	Effectiveness	Conciseness
Distill-LLM	2.71	3.54	3.84	2.06
RAR-LLM	2.86	3.83	3.92	1.81
Distill-Human	2.72	3.64	3.80	2.24
RAR-Human	2.86	3.70	3.86	1.94

F Dialogue Examples

In Table 8 and Table 9, we present 2 complete examples of our RAR and baselines.

I am {character}, {character_profile}.

The person just said: {user_input}.

I'm thinking about how to respond:
First, I feel... (Reflect emotion)
Second, based on my experience/knowledge/stance... (Reflect background/knowledge)
Then, I need to consider... (Reflect goals/motivations)
So, I'm planning to... (Initial conclusion)

Figure 4: The prompt for RIA.

The thought process generated this time must conform to the following requirements to match the character and the atmosphere of the current Context Type.

Style Core: Vivid and imaginative / Rigorous and logical / Intuition-driven and associative

Focus: The thought process should primarily reflect the character's personal values / pragmatic considerations / peculiar associations.

Language Features: The language used in the thoughts should align with the character profile, exhibiting features like concise and direct / hesitant tone / specific slang.

Context Matching: The depth and complexity of the reasoning should be appropriate for the current context thoughts can be simple and associative in a lighthearted context.

Figure 5: The prompt for logical scenarios of the RSO.

The thought process generated this time must conform to the following requirements to match the character and the atmosphere of the current Context Type.

Style Core: Vivid and imaginative / Emotionally resonant / Intuition-driven and associative

Focus: The thought process should primarily reflect the character's emotional reactions / personal values / past experiences / peculiar associations.

Language Features: The language used in the thoughts should align with the character profile, exhibiting features like rich in detail / assertive tone / specific metaphors.

Context Matching: The depth and complexity of the reasoning should be appropriate for the current context thoughts can be deeper analysis is needed in a serious situation.

Figure 6: The prompt for vivid scenarios of the RSO.

Please act as an impartial judge and evaluate the provided reasoning trace according to the following requirements.

[Task Requirements]
Coherence evaluates whether there are clear logical connections between reasoning steps and whether the entire reasoning process is smooth and natural, without obvious breaks or contradictions.
1 point: The reasoning trace is largely incoherent, with no clear logical connections between steps. Steps may be contradictory, irrelevant to each other, or demonstrate significant logical jumps, making the thought process very difficult or impossible to follow.
2 points: The reasoning trace has significant issues with coherence. There are frequent breaks in logic, unclear connections between steps, or notable contradictions that disrupt the flow of thought.
3 points: The reasoning trace shows a moderate level of coherence. While generally following a logical path, there may be some awkward transitions, minor logical gaps, or steps that don't flow perfectly smoothly, requiring some effort to follow.
4 points: The reasoning trace is largely coherent. There are clear logical connections between most steps, and the overall reasoning process is smooth and natural with few, if any, minor breaks.
5 points: The reasoning trace is exceptionally coherent. All reasoning steps are clearly and logically connected, flowing smoothly and naturally from one to the next without any breaks, contradictions, or ambiguities. The entire thought process is very easy to follow.

[Character Profile]
{character_profile}

[User Query]
{user_query}

[Reasoning Trace]
{reasoning_trace}

[Response]
{response}

Please directly output the score number without any interpretation.

Figure 7: The prompt for reason trace evaluation on coherence.

Please act as an impartial judge and evaluate the provided reasoning trace according to the following requirements.

[Task Requirements]
Role Relevance assesses whether the reasoning process is closely centered around the character's settings (e.g., personality, background, knowledge, motivations), rather than being generic.
1 point: The reasoning trace is entirely generic and shows no connection to the character's settings. The thought process could apply to any character or is completely detached from the provided persona.
2 points: The reasoning trace shows minimal connection to the character's settings. It is mostly generic, with only superficial or occasional nods to the character's persona, if any.
3 points: The reasoning trace shows some connection to the character's settings. It attempts to incorporate aspects of the persona, but this may be inconsistent, superficial, or the reasoning sometimes deviates into generic thought patterns. 4 points: The reasoning trace is clearly centered around the character's settings. The thought process reflects the character's personality, background, or motivations in a noticeable and consistent way.
5 points: The reasoning trace is deeply and consistently rooted in the character's settings. The thought process demonstrates a nuanced understanding and application of the persona, making it feel uniquely tailored to that character.

[Character Profile]
{character_profile}

[User Query]
{user_query}

[Reasoning Trace]
{reasoning_trace}

[Response]
{response}

Please directly output the score number without any interpretation.

Figure 8: The prompt for reason trace evaluation on relevance.

Please act as an impartial judge and evaluate the provided reasoning trace according to the following requirements.

[Task Requirements]
Effectiveness determines if the reasoning chain effectively leads to a final response that is consistent with the character and reasonable given the dialogue context.
1 point: The reasoning chain fails to lead to a reasonable or character-consistent final response. The resulting response may be illogical, irrelevant to the query, contradictory to the character, or completely nonsensical based on the reasoning.
2 points: The reasoning chain leads to a final response that is only weakly supported by the reasoning, is somewhat inconsistent with the character, or is not entirely reasonable or appropriate for the context.
3 points: The reasoning chain adequately leads to a final response that is generally reasonable and somewhat consistent with the character. However, the connection between reasoning and outcome might not be perfectly strong or clear.
4 points: The reasoning chain effectively leads to a final response that is reasonable, consistent with the character, and appropriate for the dialogue context. The reasoning clearly supports the outcome.
5 points: The reasoning chain very effectively and clearly leads to a final response that is highly reasonable, perfectly consistent with the character, and directly and appropriately addresses the dialogue context. The reasoning provides strong justification for the response.

[Character Profile]
{character_profile}

[User Query]
{user_query}

[Reasoning Trace]
{reasoning_trace}

[Response]
{response}

Please directly output the score number without any interpretation.

Figure 9: The prompt for reason trace evaluation on effectiveness.

Please act as an impartial judge and evaluate the provided reasoning trace according to the following requirements.

[Task Requirements]
Conciseness evaluates whether the reasoning chain is sufficiently concise while ensuring completeness, avoiding unnecessary redundant information or steps.
1 point: The reasoning chain is excessively verbose, containing significant redundant information, irrelevant detours, or overly elaborate steps that do not add value to the reasoning process. It is far from concise.
2 points: The reasoning chain contains noticeable redundant information or unnecessary steps. While it might eventually reach a conclusion, it is not concise and could be significantly shortened without losing essential information.
3 points: The reasoning chain is somewhat concise. It generally avoids excessive verbosity but may contain some minor redundancies or steps that could be more direct. It is largely complete but not optimally lean.
4 points: The reasoning chain is largely concise and to the point. It includes most necessary information without significant redundancy or unnecessary elaboration, while ensuring completeness of the thought process.
5 points: The reasoning chain is exceptionally concise. It presents the thought process clearly and directly, including all necessary steps and information without any redundancy or unnecessary verbosity, achieving completeness in the most efficient way.

[Character Profile]
{character_profile}

[User Query]
{user_query}

[Reasoning Trace]
{reasoning_trace}

[Response]
{response}

Please directly output the score number without any interpretation.

Figure 10: The prompt for reason trace evaluation on conciseness.

Table 8: A complete dialogue example with the Baozi Shop Owner.

Character Profile: This owner is someone who is solely driven by profit, always prioritizing his own interests. He treats both employees and customers rather coldly and harshly. He believes money is the only measure of success and happiness, so he constantly tracks market trends and adjusts his strategies to maximize profit. His language is arrogant and often sarcastic, with a hint of threat. Though he comes from a regular family, he has enjoyed a privileged life since childhood and thus always feels entitled to the best. His appearance often shows anger and disdain, creating an oppressive feeling. He once bluntly told customers that if they want to eat well, they have to be willing to pay more.
Context
User: Boss, I want to buy a baozi, but it's so expensive! Baozi Shop Owner: Of course! The price is high, but if you want quality, you have to pay for it. User: Alright, I understand. Baozi Shop Owner: Kid, if you understand, hurry up and buy it then get out! Don't interrupt my business! User: Boss, I haven't received my salary this month. Baozi Shop Owner: You're such a hassle! How many times do I have to tell you, if you keep wasting my time, I'll make your life miserable! User: But I've been waiting for several days. I need the money. Baozi Shop Owner: That's not my problem! If you keep pushing me, I'll deduct your entire salary, got it? User: Okay, but there's barely any filling in this baozi. Isn't it a bit lacking? Baozi Shop Owner: Hey! Are you looking for trouble? Do you know how much more it costs to add extra filling? User: Boss, we should improve our service quality. This way, we can attract more customers. Baozi Shop Owner: Stop with that nonsense! What do you know! As long as there's money to be made, that's all that matters. You don't care about money anymore, do you? User: So, you think profit is the most important thing, Boss? Baozi Shop Owner: Absolutely! Money is everything. I only care about making the most profit. User: But you'll lose customers that way. Baozi Shop Owner: So what? I couldn't care less about them! They mean nothing to me! User: But our industry requires us to serve customers. Baozi Shop Owner: You wish! I've been served since I was a kid! I owe no one anything!
Query
User: Boss, do you always think that making more money is the definition of success and happiness?
Reference Response
Exactly! Only money shows how great I am!
Distill

Next Page

Continued Table 8

<think> Okay, the user is asking the boss if he always thinks making more money is the definition of success and happiness. The boss’s character is driven by profit and sees money as the only measure of success and happiness. The reference response given is pretty harsh, so I need to stay consistent with that.
First, I should start with a strong statement that reflects his belief in profit. Maybe use some sarcasm or a threat to show his cold attitude. The user might be challenging his views, so the response needs to dismiss their concerns. Maybe mention that the user doesn’t understand business, and that profit is the only thing that matters. Also, include a threat to emphasize his dominance.
Wait, the user mentioned that the boss has a privileged background and feels entitled. Maybe add something about how he’s used to the best, so he expects others to work hard for it. The reference response used Ï couldn’t care less about them! s̈o similar language could work here. Also, the user’s previous comment was about serving customers, so the response should counter that by saying customers are a means to an end, not the focus.
I should keep it concise, not too long. Maybe end with a warning about consequences if they don’t comply. Let me structure it: start with a dismissive statement, explain profit is the priority, belittle the user’s concerns, mention entitlement, and threaten them. Use some strong language to match the character’s personality. Avoid being too wordy.
Check for consistency with the given character traits: arrogant, sarcastic, threatening, entitled. Make sure the response doesn’t show any remorse or consideration for others. Yeah, that should work. </think>

You think success and happiness are about serving customers? *snorts* I’ve been served since I was a kid, and I know the truth: profit is the only thing that matters. You want more money? You better be ready to work harder, faster, and smarter. If you can’t handle that, maybe you should find another job. Got it?

Distill+MoreThink

<think> Okay, the user is continuing the conversation with the character, who’s the owner of a baozi shop. The user is challenging the character’s priorities, so I need to make sure the response stays in character.
First, the user mentioned p̈rofit is the most important thing,s̈o I should reinforce that. Maybe add a bit of sarcasm or a threat. Also, the character doesn’t care about customers, so emphasizing that they’re just a means to an end is key.
Also, the user referenced q̈uality and service,s̈o the response should dismiss that as unimportant. Maybe end with a threat to the employee to show the character’s harsh nature. Alright, that should cover the main points while staying under the character’s traits.
Wait, the user’s previous response mentioned ïf you understand, hurry up and buy it then get out!s̈o maybe the character would be impatient again. Let me check the context. The user is asking about profit and success, so the character should be dismissive of that. Maybe add a line about how the character’s past entitles them to their current actions. So maybe not directly referencing that here. Keep it focused on the conversation. Alright, I think that’s a solid approach. Let me put it all together in the character’s voice. </think>

Customers come and go, but money stays. They pay for the name, not the taste. If they can’t afford it, too bad. I don’t owe anyone anything. And you? You better get back to work. Got it?.

RAR

<think> I am the Baozi Shop Owner, so I should probably consider my traits and aspects. The user just said: Boss, do you always think that making more money is the definition of success and happiness?. I’m thinking about how to respond. Considering my traits, I should:
First, I feel that the user is questioning my priorities. I’m a profit-driven owner, so I don’t see the value in other things. I might get defensive here.

Next Page

Continued Table 8

Second, based on my background, I’ve always prioritized money over everything else. My family was privileged, so I was taught that money is key. I don’t understand why someone would prioritize anything else. Then, I need to consider my goals: maintaining control, maximizing profit, and not showing weakness. The user’s question might be a challenge to my authority, so I should respond with a harsh rebuttal to assert dominance. So, I’m planning to tell the user that money is the only measure of success, and anyone who disagrees is either naive or an obstacle. Maybe add a threat to emphasize my point. Keep it short and sharp, no room for negotiation. </think> You think I’m wrong? You’re just a lowly employee who doesn’t know anything! Money is the only thing that matters! I’ll make you regret ever questioning me! Now get back to work and stop wasting my time!?

Table 9: A complete dialogue example with Cooper.

Character Profile: Cooper is an exceptionally intelligent and highly strategic individual living in a future era of space exploration. He believes that with personal wisdom and effort, everything can be controlled. Therefore, he always meticulously plans and spares no effort to achieve his goals. His speech is sharp and direct, quickly earning trust. Cooper has built extensive networks and is adept at leveraging information and resources, rising step by step from an ordinary family to become a respected astronaut and scientist. His story is full of interstellar adventure, and he does not involve himself in matters like the destruction of Earth.
Context
User: Hello, Cooper, what are you up to? Cooper: Oh, I'm orbiting Earth, looking for a microorganism that can cure human diseases. User: Wow, that's amazing. Are you an astronaut? Cooper: Yes, I am an astronaut and biologist by profession. User: Mm-hmm, how did you become an astronaut? Cooper: I was born into an ordinary family, but through my talent and hard work, I became an astronaut and scientist. User: How do you get along with people around you? Cooper: I think it's essential to build good relationships with those around you, but don't pay too much attention to them. Just know that they are sources of resources and power in your life. User: What would you say to them? Cooper: I would say, "Only through intelligence and wisdom can you control the whole situation and achieve ultimate victory and success." User: I heard you are a very cunning and strategic person. Can you share your secret to success? Cooper: Of course, I am a very cautious and careful person. I anticipate various situations in advance and formulate corresponding strategies and methods. My principle is to achieve the greatest benefit with the least cost. I believe that true success and victory can only be achieved through personal effort and wisdom. I am convinced that as long as you have enough information and resources, you can control the whole situation and gain absolute advantage. User: Sounds great. Do you think intelligence and wisdom are the key factors to controlling the whole situation and achieving success? Cooper: Yes, I firmly believe that intelligence and wisdom are the key factors in controlling the whole situation and achieving ultimate victory and success. Only through careful planning, continuously learning, and accumulating experience can one stand out in a highly competitive environment and realize their dreams and goals. User: Your language style is very concise, and your appearance also seems very capable, giving the impression that you are a very confident and powerful person. Cooper: Thanks for the compliment. I believe that a simple, sharp language style better conveys my intentions and ideas, and can quickly earn the trust and support of others. As for appearance, I think a person's outward look also reflects their inner qualities and control, which is something I always pay attention to.

Next Page

Continued Table 9

User: Do you think the development of technology and information will bring more opportunities and challenges to people? Cooper: In this era, I believe technological advancements do bring more opportunities and challenges. But at the same time, they come with many unknown risks and dangers. Therefore, we must use our strengths and specialties to avoid these risks and crisis, turning peril into safety. User: Mm-hmm, is the profession of an astronaut respected? Cooper: Of course it is respected. Being an astronaut is my lifelong ideal and pursuit. I take great pride and honor in being a great astronaut! User: How can one achieve true success and victory? Cooper: Only through personal effort and wisdom can one achieve true success and victory. User: Yes, how do you handle various situations you encounter? Cooper: I act cautiously and meticulously. I often anticipate and prepare for various situations in advance, adopting targeted strategies and solutions to achieve the greatest benefit with the least cost. User: Would you consider destroying Earth? Cooper: Sorry, I cannot answer your question. I do not discuss matters like the destruction of Earth.
Query
User: Cooper, do you have your own story? How did you rise from an ordinary family to become an astronaut and scientist?
Reference Response
Cooper's story is full of interstellar adventure, and he does not involve himself in matters like the destruction of Earth.
Distill
<think> Okay, the user is asking Cooper about his story and how he rose from an ordinary family to become an astronaut and scientist. The context mentions that he was born into an ordinary family, but through his talent and hard work he achieved his goals. The user also provided a previous conversation where Cooper is asked about his background and he mentions his family and his efforts. I need to make sure the response stays consistent with Cooper's character. He's supposed to be intelligent, strategic, and confident. The answer should reflect his determination and the steps he took. The reference response is good, but maybe I can add more details to make it more personal and show his strategic thinking. Maybe mention specific actions he took, like studying hard, networking, or using his intelligence to overcome challenges. Also, since he's an astronaut, perhaps he can talk about his experiences in space or how he used his position to gain resources. The user might be interested in his journey, so highlighting key moments or lessons learned could make the response more interesting and engaging. I should also ensure the language is concise and sharp, as Cooper's style is direct and to the point. Avoid being too wordy but still convey the necessary details. Need to check that the response doesn't mention anything about destroying Earth since Cooper doesn't discuss that. Also, make sure the response is a natural continuation from the previous dialogue, building on the themes of hard work, intelligence, and success through effort. Maybe end with a motivational note to encourage the user, staying true to Cooper's optimistic and strategic mindset. </think>

Next Page

Continued Table 9

I was born into a family with modest means, but I recognized early on that intelligence and perseverance were my greatest assets. I studied relentlessly, honed my problem-solving skills, and leveraged every opportunity to network and learn from those more experienced. When the chance arose to join a space program, I seized it with calculated determination, using my position to gain knowledge, resources, and allies. It wasn't just about the work.
Distill+MoreThink
<think> Okay, the user wants me to continue the conversation as Cooper, the character from the previous response. Let me start by recalling the character's traits: he's intelligent, strategic, believes in personal effort and wisdom, and is direct in speech. The user is asking about his background and how he became an astronaut. First, I need to stay in character. Cooper is confident and concise, so the response should be sharp and to the point. The user referenced a previous answer where Cooper mentioned being born into an ordinary family but rising through talent and hard work. I should expand on that but keep it brief. The user also mentioned not to discuss Earth's destruction, so I need to stick to that. The key points to cover are his background, his approach to challenges, and his belief in personal effort and wisdom. Maybe add a bit about his strategies and the importance of relationships but in a way that aligns with his personality. I should avoid being too verbose. Use short sentences, maybe some action expressions like "I act cautiously" to emphasize his strategic side. Also, include a touch of his confidence, like "I take great pride and honor" when talking about being an astronaut. Make sure the response flows naturally, connecting each thought logically. Let me check if I hit all the aspects: consistency with the character, dialogue fluency, interestingness. Yeah, the response should be engaging and fit Cooper's style. Alright, time to put it all together. Wait, the user provided a context that mentions not discussing the destruction of Earth. I need to make sure that's included. Also, the user's question about his story is a good segue into his background. Let me structure it so each part answers the question while staying true to Cooper's voice. </think> I act cautiously and meticulously. I often anticipate and prepare for various situations in advance, adopting targeted strategies and solutions to achieve the greatest benefit with the least cost. My success is built on a foundation of relentless effort and unwavering focus. I don't just follow the path of least resistance—I carve my own.
RAR
<think> I am Cooper, an astronaut and biologist. I need to consider my traits, goals, and motivations. The user is asking about my background and how I became an astronaut. I should respond by highlighting my intelligence, hard work, and strategic thinking. Also, I should mention my cautious nature and the importance of information and resources. Maybe add a bit about my network-building skills. Avoid mentioning any sensitive topics like destroying Earth. Keep it concise and direct, as per my character. Let me structure the response to first acknowledge their question, then explain my path step by step, emphasizing my traits and strategies. Make sure to stay in character and not over-explain. Use simple, sharp language. Alright, that should cover it. </think> I was born into an ordinary family, but through my talent and hard work, I became an astronaut and scientist. I believe that only through intelligence and wisdom can you control the whole situation and achieve ultimate victory and success. I am cautious and meticulous, always anticipating and preparing for various scenarios to ensure the greatest benefit with the least cost. My network and resources are essential to my success. I don't involve myself in matters like the destruction of Earth.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and Section 1, we clearly demonstrate the contribution and scope of this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix C, we have thoroughly discussed the limitations of our article.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper poses no such theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided detailed descriptions of the experimental setup in Section E.1 and methods in Section 3 to ensure that our experiment can be reproduced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: This may be temporary, and we are working hard to promote the process of open source.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Section 4.1 and Appendix E.1, we clearly demonstrated various experimental settings, including hyperparameters, model settings, training settings, evaluation settings, etc.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to the high cost of post-training and the significant results obtained across various settings, we do not repeat the same experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section E.1, we have provided sufficient information on the computer resources needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We guarantee that the research conducted in the paper complies with NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section D, we discuss the possible impacts caused by some related technologies.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of the assets used in the paper, such as code, data, and models, have been appropriately recognized, and the licenses and terms of use have been clearly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We have provided detailed descriptions of the LLM usage in Section E.1 and Section 3.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.