Improving General Role-Playing Agents via Psychology-Grounded Reasoning and Role-Aware Policy Optimization

Zhenhua Xu[1], Dongsheng Chen[2*], Jian Li[2*], Yitong Lin[1], Zhebo Wang[1], Jiafu Wu[2], Yizhang Jin[2], Chengjie Wang[2], Meng Han[1], Yabiao Wang[1,2]*

[1]*Zhejiang University*, [2]*Tencent Youtu Lab*

arXiv:2606.27025v1 [cs.CL] 25 Jun 2026

Building general-purpose role-playing agents that faithfully portray any character from a natural-language profile remains challenging. The dominant paradigm—supervised fine-tuning—encourages behavioral mimicry without deep, human-like internal thought processes, resulting in poor out-of-distribution generalization. Therefore, we propose **Psy-CoT**, a psychology-grounded chain-of-thought framework that decomposes pre-response reasoning into three role-specific steps—*Interaction Perception*, *Psychological Empathy*, and *Logical Construction*—so that the model *thinks dynamically* from the profile rather than merely mimicking surface patterns. While structured reasoning provides a foundation, it alone is insufficient; reinforcement learning is essential to further align the model with character fidelity. However, we observe that under LLM-based reward models, both generic phrases that hack the reward model and genuinely role-specific phrases receive identical gradient signals—this hacking accumulates over training, misleading the model into treating both as equally optimal choices. To address this, we propose **Role-Aware Policy Optimization (RAPO)**, which uses profile–token mutual information to weight gradients asymmetrically—amplifying role-specific tokens under positive advantage while attenuating them under negative advantage. Experiments on CoSER, CharacterBench, and CharacterEval demonstrate that Psy-CoT outperforms existing role-playing CoT methods, and RAPO consistently surpasses GRPO across multiple model scales.

Date: 2026
Correspondence: mhan@zju.edu.cn
Project Leader: caseywang@tencent.com

1 Introduction

The strong reasoning and generation capabilities of large language models (LLMs) [Li et al., 2026, Novikov et al., 2025, Jiao et al., 2026, Wang et al., 2026] make them natural candidates for *role-playing*—adopting and sustaining a specific persona throughout an interaction. A particularly compelling and practical variant is the *general-purpose role-playing agent* (RPA): a single model that portrays *any* character given only a natural-language role profile at inference time, underpinning applications such as virtual companionship, interactive storytelling, and game NPCs [Chen et al., 2025a, Tseng et al., 2024, Chen et al., 2025b].

The dominant paradigm for building RPAs is supervised fine-tuning (SFT) on curated role-playing corpora (e.g., ChatHaruhi [Li et al., 2023], RoleLLM [Wang et al., 2024], CharacterGLM [Zhou et al., 2024], Crab [He et al., 2025], CoSER [Wang et al., 2025], and AdaMARP [Xu et al., 2026]). While effective, SFT encourages *behavioral mimicry* rather than genuine character understanding [Liu et al., 2025, Tang et al., 2026]: the model memorizes surface patterns but fails to generalize to out-of-distribution roles, regressing to generic responses (see Section 4). What is needed, instead, is for the model to *think dynamically* from the profile—adapting its reasoning to each situation based on who the character is. A natural approach is to prepend a chain-of-thought (CoT) [Wei et al., 2022] before the response, yet Feng et al. [2025] shows that general-purpose CoT steers the model toward objective and "optimal" conclusions, whereas a faithful character should reason subjectively, idiosyncratically, and sometimes irrationally. The remedy is therefore a *role-specific* reasoning paradigm that mirrors how humans mentally prepare before acting in character.

To this end, we propose **Psy-CoT**, a psychology-grounded chain-of-thought framework that decomposes the pre-

* Equal contribution.

response reasoning into three sequential steps: *Interaction Perception*, *Psychological Empathy*, and *Logical Construction*. Interaction Perception, grounded in Theory of Mind [Chen et al., 2025c], requires the model to grasp the global scene and infer the interlocutor's underlying intentions and mental state. Psychological Empathy, informed by cognitive appraisal theory [Lazarus, 1991] and emotion regulation theory [Gross, 1998], asks the model to filter the situation through the character's unique background and flaws, identify its genuine emotional reaction, and decide how much to express, mask, or weaponize. Logical Construction, inspired by goal-directed behavior theory [Locke and Latham, 2002], synthesizes facts through the character's worldview and selects a short-term objective the character would *naturally* pursue—even if impulsive or suboptimal. Together, the three steps constitute the character's internal deliberation that precedes the final response, extending the CoT paradigm with role-specific reasoning.

Psy-CoT provides structured reasoning, yet reasoning alone is insufficient—the model must also learn from feedback that distinguishes faithful character responses from generic ones. Existing RL methods for role-playing [Liu et al., 2025, Ye et al., 2025, Tang et al., 2026] focus on reward design yet treat all tokens equally in policy optimization. During training, we repeatedly observe a degradation pattern: even under reward guidance, the model retains a non-negligible probability of generating generic, templated actions, inner thoughts, and dialogue that could belong to any persona—under repeated sampling, such role-agnostic expressions appear with frequency comparable to role-specific ones, regardless of the scene and context. The root cause is that under LLM-based reward models, RL methods simply maximize reward: whether a role-agnostic generic phrase captures the reward model's preference or a genuinely role-specific phrase earns a high score, both receive identical gradient signals—and likewise for penalties. This effectively amounts to reward hacking that accumulates over training, misleading the model into treating both generic and role-specific expressions as equally optimal choices. A natural question arises: *can we ensure that, as the reward increases, role-specific tokens contribute disproportionately more while preserving exploration?*

To address this, we propose **Role-Aware Policy Optimization (RAPO)**, which quantifies how much each token is influenced by the role profile and adjusts gradient contributions accordingly. Concretely, before computing gradients, for each generated token y_t, RAPO computes the profile–token mutual information $I_t = H(y_t \mid x) - H(y_t \mid P, x)$ by comparing the model's entropy with and without the profile P. The magnitude $|I_t|$ measures how actively the profile shapes the token: $|I_t| \gg 0$ signals a role-specific token (either **exploiting** established character patterns when $I_t > 0$, or **exploring** role-specific behaviors when $I_t < 0$), while $|I_t| \approx 0$ marks a profile-agnostic token. RAPO then applies *asymmetric* per-token weighting: when the response receives a positive advantage, gradients on role-specific tokens are amplified so the model extracts more signal from its character-specific choices; when the advantage is negative, gradients on those same tokens are attenuated to reduce the cost of failed character-specific exploiting or exploration. The result is a policy that better perceives the role profile during token generation and maintains a balance between exploration and exploitation of character-specific behaviors.

We validate our approach through extensive experiments. Applying Psy-CoT to Llama-3.3-70B-Instruct yields average improvements of 5.2% and 4.2% on CoSER and CharacterBench, respectively, outperforming existing role-playing CoT methods. Furthermore, training with RAPO on Qwen2.5-7B-Instruct, Qwen3-4B-Instruct, and Llama-3.1-8B-Instruct (all equipped with Psy-CoT) yields improvements of 13.7%, 15.6%, and 40.1% over the untrained baseline on CoSER, with effectiveness generalizing to CharacterBench and CharacterEval, demonstrating that RAPO's profile-aware token weighting enables the model to generate more character-faithful responses.

Overall, **the main contributions of the work are as follows**: (1) We propose **Psy-CoT**, a psychology-grounded chain-of-thought framework that decomposes pre-response reasoning into Interaction Perception, Psychological Empathy, and Logical Construction, enabling role-specific deliberation rather than generic rationality. (2) We propose **RAPO**, a role-aware policy optimizer that leverages profile–token mutual information to weight gradients asymmetrically, amplifying role-specific signals while maintaining exploration. (3) Extensive experiments on three benchmarks and multiple model architectures validate the effectiveness of both Psy-CoT and RAPO.

2 Related Work

Improving RPAs via Data Construction General role-playing demands faithful portrayal of arbitrary, user-defined or fictional characters. A prevalent approach addresses this by constructing large-scale multi-character corpora that

maximize persona coverage and mitigate overfitting to fixed rosters. Representative works include ChatHaruhi [Li et al., 2023], DITTO [Lu et al., 2024], CharacterGLM [Zhou et al., 2024], ROLEPERSONALITY [Ran et al., 2024], Crab [He et al., 2025], TAILORGEN [Gao et al., 2025], CoSER [Wang et al., 2025], and AdaMARP [Xu et al., 2026], each contributing distinct data-collection or synthesis strategies to broaden character coverage.

Reasoning and Reinforcement Learning for RPAs Recent works have begun to equip role-playing agents with explicit reasoning capabilities and refine them via RL. CogDual [Liu et al., 2025] introduces CB-CoT, which augments the thinking phase with *situational awareness* and *self-awareness* to better ground the character's internal state. COP [Ji et al., 2025] lets the model first perform self-questioning and self-answering, while Character-R1 [Tang et al., 2026] requires the model to analyze which aspects deserve focus and the corresponding attributes in the think phase. On the optimization side, CPO [Ye et al., 2025] proposes implicit comparisons to guide policy learning, while RAR [Tang et al., 2025] and PCL [Ji et al., 2025] promote preference learning through contrastive objectives.

3 Design of RAPO

3.1 Problem Formulation

Following CoSER [Wang et al., 2025], HER [Du et al., 2026], and AdaMARP [Xu et al., 2026], we view role-playing as an open-ended, multi-turn interaction in which character models, each instantiated from a written role specification, converse with a user and with each other to jointly unfold a story. An episode begins with an initial role set $\mathcal{R} = \{r_1, \ldots, r_N\}$ (the user treated as a distinguished member whose profile may be partially unspecified), each role r_i equipped with a natural-language *role profile* P_i (full schema in Appendix C). Together with an initial scene, these profiles form the initial state s_0. An external orchestrator selects the next speaker from $\mathcal{R}$ at every turn; in single-role settings, this degenerates to deterministic user–character alternation. The goal is to bring the written character to life across arbitrarily long interactions, requiring simultaneous *role consistency*, *interaction competence*, and *narrative advancement*. Each non-user role is instantiated by a shared policy π_θ conditioned on its profile and the shared history. Following CoSER [Wang et al., 2025] and HER [Du et al., 2026], a role's per-turn response interleaves **Thought** (inner monologue, wrapped in `[...]`), **Action** (externally observable cues such as gestures and expressions, wrapped in `(...)`), and **Dialogue** (overt spoken utterance, left untagged).

3.2 Psychology-Grounded Chain-of-Thought (Psy-CoT)

Effective role-playing demands human-like reasoning—perceiving the social environment, experiencing character-consistent emotions, and pursuing goal-directed strategies—rather than generic reasoning [Feng et al., 2025]. We propose **Psy-CoT**, a structured chain-of-thought framework grounded in theories of social cognition and emotion (full prompt in Appendix D). Before producing a response, the model reasons through three sequential steps—*Interaction Perception*, *Psychological Empathy*, and *Logical Construction*—each inspired by a distinct psychological mechanism.

Interaction Perception. Theory of Mind (ToM)—the ability to reason about the mental states of oneself and others [Chen et al., 2025c]—is a cornerstone of human social intelligence. Interaction Perception operationalizes ToM by having the model first grasp the scene's *global view* (setting, social configuration, who is present) and then attend to the *immediate interaction*—the interlocutor's stated content, tone, underlying intentions, and current mental state.

Psychological Empathy. Cognitive appraisal theory [Lazarus, 1991] holds that emotions arise from an individual's subjective evaluation of events; emotion regulation theory [Gross, 1998] further describes how individuals modulate emotional expression through suppression, reappraisal, or amplification. Psychological Empathy applies these by requiring the model to filter the perceived situation through the character's unique background and values, identify the character's genuine emotional reaction, and decide *how much of this emotion to express, mask, or weaponize*.

Logical Construction. Goal-directed behavior theory [Locke and Latham, 2002] posits that action is organized around goals: individuals synthesize information through their worldview, set short-term objectives, and select strategies accordingly. Logical Construction mirrors this process: the model synthesizes relevant facts and the character's

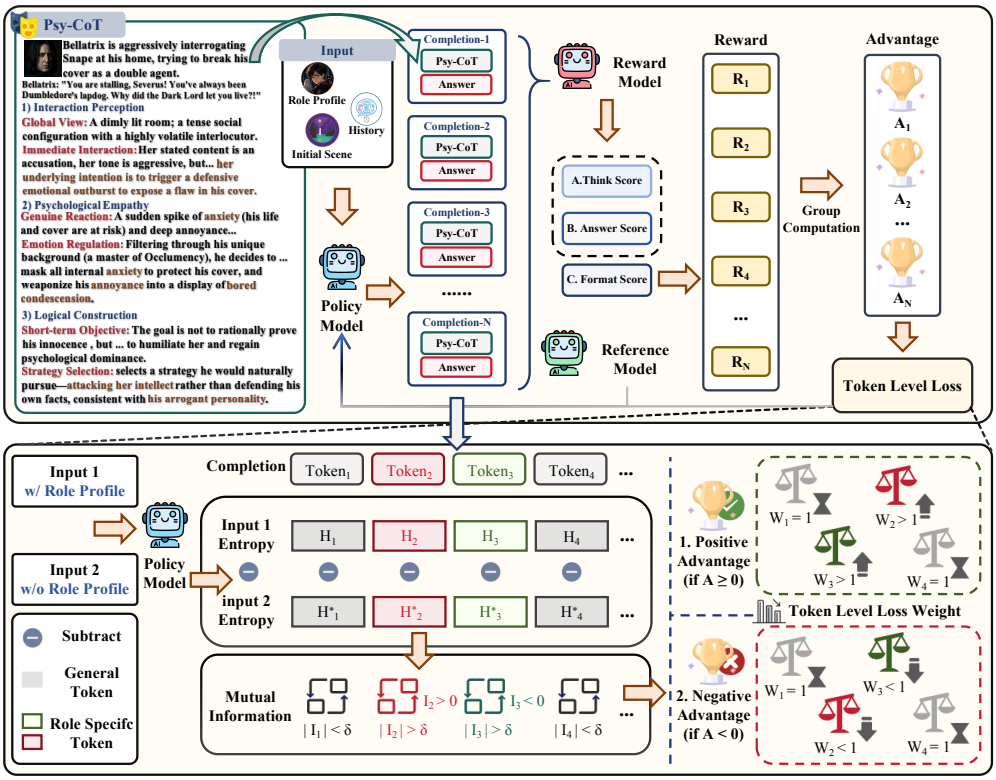


Figure 1. Overview of RAPO. The upper part illustrates the sampling stage: given an input sample, the policy model rolls out multiple completions, each of which is scored by the reward function, and group-relative advantages are then computed. The lower part depicts the token-level gradient modulation applied to each completion, where role-agnostic and role-specific tokens receive differentiated gradient signals conditioned on the sign of the advantage.

knowledge, then decides on a short-term objective the character would *naturally* pursue (e.g., to provoke, deflect, comfort, dominate), selecting a strategy consistent with the character's personality—even if impulsive or suboptimal.

3.3 Role-Aware Policy Optimization (RAPO)

Psy-CoT provides structured reasoning, yet does not resolve the challenge in Section 1: standard RL treats all tokens uniformly, giving role-agnostic and role-specific expressions identical gradient signals. To address this, we propose RAPO built upon GRPO [Shao et al., 2024]: for each prompt (P, x), where P is the role profile and x the remaining context (initial scene and dialogue history), the policy π_θ generates n completions $\{y^{(i)}\}_{i=1}^{n}$; each receives a scalar reward, and group-relative advantages are computed as $\hat{A}^{(i)} = \big(R(y^{(i)}) - \mu_R\big)/\sigma_R$ as illustrated in Figure 1. Below, we present our reward design and the RAPO mechanism.

Reward function. Evaluating a role-playing response requires judging both structural compliance and role-specific quality. Since role-playing is an open-ended task where responses are rarely objectively correct or incorrect, we employ an LLM-as-a-judge approach rather than a traditional rule-based reward model. We decompose the reward into three terms:

$$R(y) = \alpha\, R_{\text{fmt}}(y) + \beta\, R_{\text{think}}(y) + \gamma\, R_{\text{ans}}(y), \tag{1}$$

where $\alpha + \beta + \gamma = 1$. $R_{\text{fmt}} \in [0, 1]$ is a deterministic format check (Appendix E.1): the completion must contain both `<system_think>` and `<answer>` tags, the three required section headings, and no extraneous content outside them. $R_{\text{think}} \in [0, 1]$ and $R_{\text{ans}} \in [0, 1]$ are scored by an LLM judge that first produces a detailed rubric-by-rubric analysis of the response and then derives a score from that analysis (Appendix E.2), *only when* $R_{\text{fmt}} = 1$; Otherwise, they are zeroed, making ill-formatted completions receive negligible reward. The think rubric evaluates interaction perception (A1), psychological empathy (A2), logical construction (A3), and think–answer consistency (A4) (Table 13);

the answer rubric evaluates conversational ability (B1), character consistency (B2), interpersonal interaction (B3), and narrative progression (B4) (Table 14).

Profile–token mutual information. We quantify profile influence on each token via mutual information [Shannon, 1948]. For each generated completion, we compute two conditional entropies at position t: $H(y_t \mid P, x)$ with the full prompt, and $H(y_t \mid x)$ with the profile ablated. Their difference yields the profile–token mutual information:

$$I_t \;=\; I(y_t;\, P \mid x) \;=\; H(y_t \mid x) - H(y_t \mid P,\, x). \tag{2}$$

The sign of I_t reveals *how* the profile intervenes: $I_t > 0$ indicates that P *reduces* uncertainty—the model, leveraging its understanding of the character, is more inclined to generate this token (e.g., a character's signature phrase); $I_t < 0$ indicates that P *increases* uncertainty—the model tends to explore beyond generic patterns to produce character-specific outputs; $|I_t| \approx 0$ marks a general token that the profile does not affect.

Asymmetric weighting. Whether the profile *exploits* existing character patterns ($I_t > 0$) or *explores* beyond generic ones ($I_t < 0$), both signal active profile influence and deserve differentiated treatment. We take $|I_t|$ as influence strength and introduce a threshold δ to filter noise: tokens with $|I_t| < \delta$ receive weight 1; for tokens with $|I_t| \geq \delta$, RAPO applies *asymmetric* weighting conditioned on the advantage sign:

$$w_t = \begin{cases} 1 + \mathrm{clamp}\big(|I_t| \cdot \lambda,\; 0,\; \mu_+\big) & \text{if } \hat{A} \geq 0 \text{ and } |I_t| \geq \delta, \\ 1 - \mathrm{clamp}\big(|I_t| \cdot \lambda,\; 0,\; \mu_-\big) & \text{if } \hat{A} < 0 \text{ and } |I_t| \geq \delta, \\ 1 & \text{if } |I_t| < \delta. \end{cases} \tag{3}$$

where λ scales the mutual information and μ_+, μ_- clip the weight to $[1,\, 1{+}\mu_+]$ and $[1{-}\mu_-,\, 1]$ respectively. The asymmetry is deliberate: when the response is good ($\hat{A} > 0$), we amplify gradients on role-specific tokens so the model extracts more learning signal from its character-specific choices; when the response is poor ($\hat{A} < 0$), we attenuate penalties on those same tokens to prevent the model from associating negative feedback with its character-identifying behaviors—while the error may stem from either generic or character-specific tokens, the penalty should be lighter on the latter to preserve the model's willingness to express character identity.

Training objective. RAPO incorporates the per-token weights into the GRPO clipped objective:

$$J(\theta) \;=\; \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} w_t^{(i)}\, \min\!\Big(\rho_t^{(i)}\hat{A}_t^{(i)},\; \mathrm{clip}\big(\rho_t^{(i)},\, 1{-}\epsilon,\, 1{+}\epsilon\big)\,\hat{A}_t^{(i)}\Big) \;-\; \beta D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}), \tag{4}$$

where G is the group size, o_i the i-th completion of length $|o_i|$, $\rho_t^{(i)} = \pi_\theta(y_t^{(i)} \mid \cdot)/\pi_{\mathrm{old}}(y_t^{(i)} \mid \cdot)$ the importance ratio (IS), ϵ the clipping coefficient, and D_{KL} a regularizer anchoring π_θ to the reference model. The per-token weight $w_t^{(i)}$ from Eq. (3) scales the gradient at each position: role-specific tokens receive amplified signals when $\hat{A}_t^{(i)} > 0$ and attenuated ones when $\hat{A}_t^{(i)} < 0$.

4 Experiments

4.1 Experimental Setup

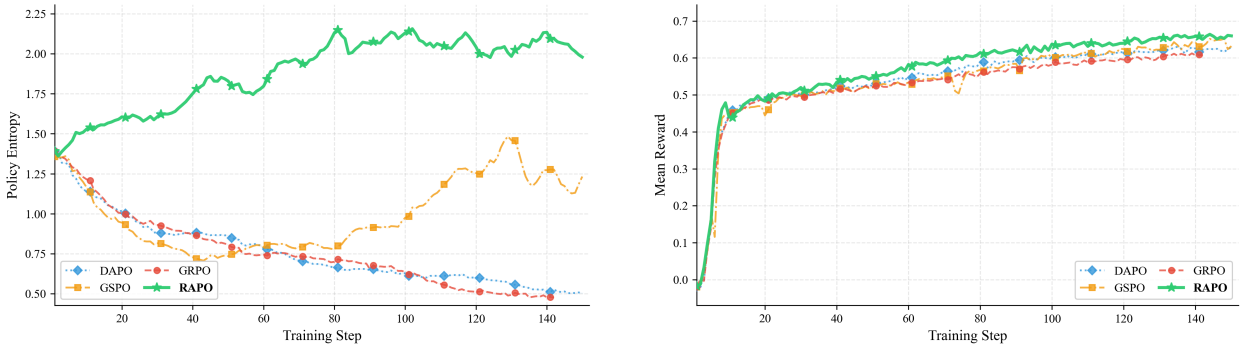
Models. We include seven general-purpose instruction-tuned models spanning a range of scales and families: DeepSeek-V4-Flash [DeepSeek-AI, 2026], GPT-4o-mini [OpenAI et al., 2024], GPT-5-Chat [OpenAI, 2025], Qwen3-4B-Instruct [Team, 2025], Qwen2.5-7B-Instruct [Qwen Team et al., 2025], Llama-3.1-8B-Instruct [Grattafiori et al., 2024], and Qwen2.5-14B-Instruct [Qwen Team et al., 2025]. These serve as baselines for off-the-shelf LLM performance. We also include seven role-playing–specialized models: the proprietary Doubao-1.5-Pro-Character [ByteDance, 2024]; SFT-based variants fine-tuned on the Crab [He et al., 2025], CoSER [Wang et al., 2025], and AdaRPSet [Xu et al., 2026] datasets respectively (Qwen2.5-7B-Instruct-Crab, Qwen2.5-7B-Instruct-CoSER, Qwen2.5-14B-Instruct-AdaMARP, Llama-3.1-8B-Crab, Llama-3.1-8B-CoSER, Llama-3.1-70B-CoSER). These represent dataset-level approaches and serve as performance references for RAPO.

Table 1. Performance on the CoSER benchmark. RAPO consistently outperforms other RL algorithms and generalizes across model architectures.

Model	Storyline Consistency	Anthropo-morphism	Character Fidelity	Storyline Quality	Avg.
Proprietary & General-Purpose LLMs					
Doubao-1.5-Pro-Character	48.13	32.74	40.93	52.28	43.52
GPT-4o-mini	47.32	40.18	40.51	52.47	45.09
GPT-5-Chat	46.78	35.10	42.03	59.61	45.88
DeepSeek-V4-Flash	47.07	37.30	39.77	57.84	45.49
Qwen3-4B-Instruct	47.01	35.56	35.72	50.17	42.11
+ Psy-CoT + GRPO	55.60	41.00	34.64	50.44	45.42
+ Psy-CoT + RAPO	55.41	40.80	39.32	55.94	47.87
Qwen2.5-7B-Instruct	48.09	33.15	37.42	46.58	41.31
+ Psy-CoT + CPO	52.31	36.85	28.73	49.62	41.88
+ Psy-CoT + GRPO	53.27	42.52	34.45	52.04	45.82
+ Psy-CoT + GSPO	51.84	37.66	32.18	48.78	42.61
+ Psy-CoT + DAPO	52.09	39.08	37.19	44.03	43.10
+ Psy-CoT + RAPO	55.77	44.67	37.50	53.02	47.74
Qwen2.5-7B Role-Playing Specialized					
Qwen2.5-7B-Instruct-CoSER	54.80	37.07	45.17	59.08	49.03
Qwen2.5-7B-Instruct-Crab	55.31	32.18	40.44	51.06	44.75
Llama-3.1-8B-Instruct	37.39	25.78	30.22	33.36	31.69
+ Psy-CoT + GRPO	52.99	37.57	35.08	47.64	43.32
+ Psy-CoT + RAPO	53.94	41.34	33.99	48.95	44.55
Llama-3.1-8B Role-Playing Specialized					
Llama-3.1-8B-Crab	52.12	37.13	40.99	56.30	46.64
Llama-3.1-8B-CoSER	49.90	37.27	42.82	60.37	47.59
Larger Models					
Qwen2.5-14B-Instruct	47.73	34.97	39.65	53.29	43.91
Qwen2.5-14B-Instruct-AdaMARP	48.56	36.02	45.03	57.38	47.00
Llama-3.1-70B-CoSER	48.17	36.26	40.95	58.20	45.89
Llama-3.3-70B-Instruct	49.71	36.17	40.94	57.97	46.20

Datasets. Since Psy-CoT and RAPO are reasoning and training methods rather than data-centric approaches, we sample training data from existing datasets. Specifically, we randomly sample 25,000 instances from the HER [Du et al., 2026] dataset and 5,000 instances from the AdaRPSet-Synthesis [Xu et al., 2026] dataset, both of which are recent, high-quality open-source role-playing datasets. Each sample's system prompt comprises *Profile*, *Init Scene*, *Dialogue History*, and *Output Requirement*. We randomly truncate the dialogue history (keeping the last turn as the target character's output) to encourage robustness across context lengths; the model then produces the chain-of-thought reasoning followed by the in-character answer. To support the profile–token mutual information computation (Eq. (2)), each sample also stores a *profile-ablated* version with the Profile removed, enabling efficient paired forward passes.

Baselines. We compare against two groups of baselines. The first group evaluates chain-of-thought designs: the **Vanilla** setting, where the model generates responses directly without any chain-of-thought reasoning; and **CB-CoT** [Liu et al., 2025], a role-playing chain-of-thought framework that includes *situational awareness* and *self awareness* as its reasoning steps (see Appendix D for the full system prompts). The second group comprises existing reinforcement learning algorithms applied to the same training data and reward function as RAPO: **GRPO** [Shao et al., 2024], which computes group-relative advantages without token-level differentiation; **GSPO** [Zheng et al., 2025], which replaces token-level importance ratios with sequence-level ones; **DAPO** [Yu et al., 2025], which incorporates dynamic sampling and related strategies; and **CPO** [Ye et al., 2025], which incorporates implicit preference comparison into role-playing training. Since these algorithms can seamlessly replace RAPO in our training pipeline while sharing the same Psy-CoT and reward function, they serve as the primary comparisons for isolating the contribution of RAPO's profile-aware token weighting.



(a) Policy entropy

(b) Mean reward

Figure 2. Training dynamics of different policy optimization methods on Qwen2.5-7B-Instruct with Psy-CoT. **(a)** GRPO and DAPO exhibit early entropy collapse indicative of premature convergence, while GSPO partially recovers but with limited exploration space remaining; RAPO maintains a stable exploration–exploitation balance throughout training. **(b)** RAPO achieves the highest reward among all methods, confirming that its higher entropy does not come at the cost of reward quality.

Benchmarks. We evaluate on three widely used benchmarks covering both English and Chinese. **CoSER** [Wang et al., 2025] assesses *Storyline Consistency*, *Anthropomorphism*, *Character Fidelity*, and *Storyline Quality*. **CharacterBench** [Zhou et al., 2025], which provides both English and Chinese splits (we use the English one), evaluates *Memory*, *Knowledge*, *Person*, *Emotion*, *Morality*, and *Believability*. **CharacterEval** [Tu et al., 2024] is a Chinese benchmark that tests *Conversational Ability*, *Character Consistency*, and *Role-Playing Attractiveness*, allowing us to verify the cross-lingual generalization of our method. For evaluation, CoSER is scored by GPT-4o-mini as the LLM judge, while CharacterBench and CharacterEval use their respective specifically trained reward models.

4.1.1 Training Configuration

During RAPO training, we set the rollout group size $n = 5$, the mutual information threshold $\delta = 0.1$, the scaling factor $\lambda = 1.0$, the positive amplification cap $\mu_+ = 0.2$, and the negative attenuation cap $\mu_- = 0.10$. All other training settings are kept identical across all compared methods. We use Qwen3-30B-A3B-Instruct-2507-FP8 [Team, 2025] as the reward model for R_{think} and R_{ans}. Additionally, we apply a soft length penalty to the `<answer>` block. The remaining hyperparameters and reward details are provided in Appendix H.

4.2 Main Result

Effectiveness of Psy-CoT Since smaller models (e.g., 7B) struggle to reliably follow the CoT instructions, we evaluate Psy-CoT and CB-CoT on stronger instruction-following models—Qwen2.5-14B-Instruct and Llama-3.3-70B-Instruct—on CoSER and CharacterBench (Figure 3). Both role-playing-oriented CoTs improve the stronger models on CoSER; however, when the model's base capacity is insufficient, adding CoT may hurt performance (e.g., Qwen2.5-14B-Instruct drops from 3.77 to 3.63 with CB-CoT and 3.68 with Psy-CoT on CharacterBench). Moreover, CB-CoT's JSON-structured thinking is more constrained and less comprehensive than Psy-CoT's, leading to slightly inferior overall performance.

RAPO Enhances Role-Playing Performance We train Qwen2.5-7B-Instruct with Psy-CoT under CPO, GRPO, GSPO, DAPO, and RAPO and evaluate on all three benchmarks. RAPO consistently achieves the best average score among all RL variants (Table 1, Figure 4). The largest gains appear on dimensions most indicative of role fidelity: Anthropomorphism on CoSER (+2.15 over GRPO), Person (+0.09) and Believability (+0.10) on CharacterBench (Table 10), and the Behavior sub-metric on CharacterEval (Table 11), reaching 3.81—surpassing GRPO (2.74) by 39%. These gains suggest that incorporating role profiles into RL enables the model to better perceive, leverage, and internalize character identities. Moreover, the improvements transfer to CharacterEval, a Chinese-language benchmark, demonstrating cross-lingual generalization.

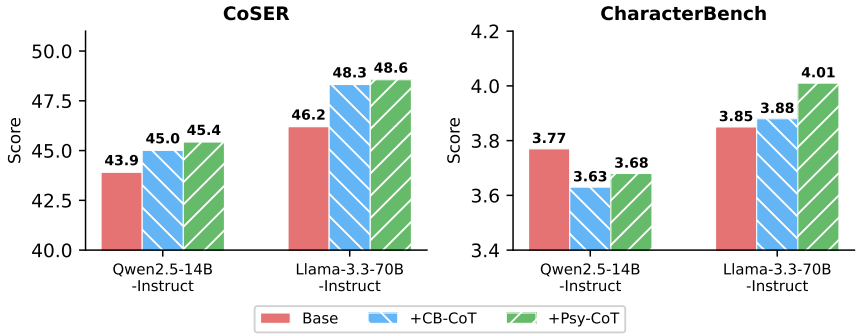
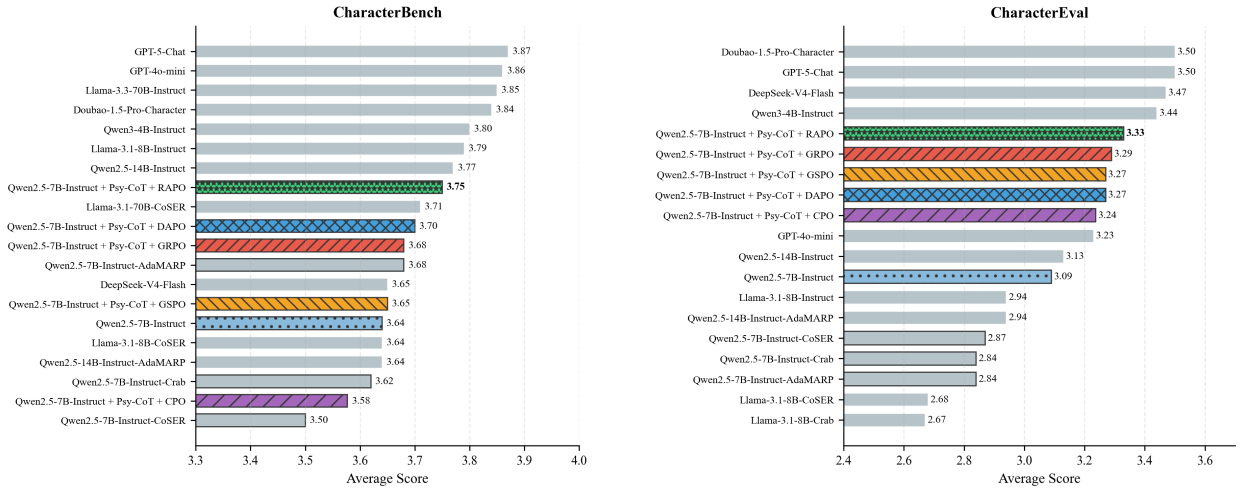


Figure 3. Effect of different CoT strategies on strong instruction-following models. Both Psy-CoT and CB-CoT improve CoSER scores, but applying CoT to less capable models can degrade performance on CharacterBench.



(a) CharacterBench **(b)** CharacterEval

Figure 4. Average performance across models on CharacterBench and CharacterEval. RAPO achieves the best average score among all RL variants on Qwen2.5-7B-Instruct; SFT-based methods harm generalization. See Table 10 and Table 11 for per-dimension results.

We also train Qwen3-4B-Instruct and Llama-3.1-8B-Instruct with GRPO and RAPO on CoSER (Table 1). RAPO again outperforms both the untrained baselines and GRPO-trained counterparts. Notably, RAPO-trained Qwen3-4B-Instruct (47.87) and Qwen2.5-7B-Instruct (47.74) even surpass proprietary models GPT-5-Chat (45.88) and DeepSeek-V4-Flash (45.49). For Llama-3.1-8B-Instruct, RAPO raises the average score from 31.69 to 44.55 (+41%).

In contrast, SFT-based approaches (Crab, CoSER, AdaMARP) tend to overfit when the training and evaluation distributions overlap (e.g., Qwen2.5-7B-Instruct-CoSER performs well on CoSER) while degrading elsewhere: on CharacterBench (Table 10), Qwen2.5-7B-Instruct-CoSER (3.50) and Crab (3.62) both fall below the base model (3.64); on CharacterEval (Table 11), all three SFT variants (2.84–2.87) drop below the untrained baseline (3.09). RAPO, by learning a reasoning-and-preference paradigm rather than memorizing surface patterns, achieves more robust and generalizable improvements.

Entropy and Reward Dynamics To understand the training behavior of different policy optimization methods, we plot the policy entropy and mean reward over the first 150 steps on Qwen2.5-7B-Instruct with Psy-CoT (Figure 2). Regarding the reward curve (Figure 2b), GRPO, GSPO, and DAPO remain largely intertwined throughout training, whereas RAPO pulls ahead after approximately 40 steps and maintains a clear lead thereafter. This indicates that once the model passes the initial exploration phase, the role-aware signals enable it to more effectively perceive and leverage

the profile, leading to consistently higher rewards.

The entropy dynamics (Figure 2a) explain why RAPO outperforms the alternatives. GRPO and DAPO suffer from monotonic entropy collapse: their token-level importance ratios and globally broadcast advantages cause rapid convergence to a local optimum, with entropy declining steadily toward near-zero—DAPO's clip-higher mechanism fails to arrest this trend. GSPO's sequence-level importance ratio can re-stimulate exploration, but the large oscillations between exploitation and recovery make its trajectory unstable; once exploitation dominates, entropy fails to rebound, and the policy drifts downward. In contrast, RAPO's profile-aware loss reweight explicitly encourages both exploitation and exploration on role-specific tokens, maintaining entropy within a healthy range throughout training and preserving generative capacity without compromising performance.

4.3 Human Evaluation

To assess practical role-playing quality beyond automatic metrics, we conduct a controlled human evaluation. We recruit 5 PhD candidates from different universities who are familiar with role-playing tasks, compensated at $15/hour. For each of the three backbone models (Qwen2.5-7B-Instruct, Qwen3-4B-Instruct, Llama-3.1-8B-Instruct), we use the same set of 20 randomly sampled CoSER benchmark samples to generate trajectories, ensuring a fair comparison across methods. Annotators judge which trajectory better embodies the target character along the four B1–B4 dimensions of our Answer Rubric (Section 3.3) as an overall impression, recording a verdict of *win*, *tie*, or *loss* for RAPO against the compared method. We aggregate annotations across the three backbone models and report the results in Figure 5.

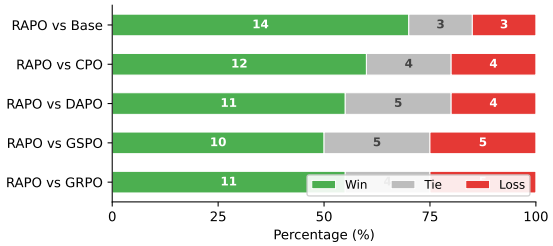


Figure 5. Human evaluation results.

As shown in Figure 5, RAPO wins more often than it loses against every baseline, with the largest margin over the untrained Base model (70% win rate) and consistent advantages over all RL variants (55%–60% win rate). These results confirm that the improvements observed in automatic metrics translate to perceptible gains in human judgment of role-playing fidelity.

4.4 Ablation Studies

Impact of Psy-CoT on Training We remove Psy-CoT from the training data and train Qwen2.5-7B-Instruct with SFT and RL, respectively, evaluating on all three benchmarks (Table 2). RAPO tends to yield more stable improvements across benchmarks than SFT (e.g., No-CoT-RAPO vs. No-CoT-SFT on CharacterBench: 3.73 vs. 3.57; CharacterEval: 3.16 vs. 3.10). More importantly, adding Psy-CoT to RL training brings further gains across all three benchmarks (CoSER: 47.74 vs. 44.65; CharacterBench: 3.75 vs. 3.73; CharacterEval: 3.33 vs. 3.16), confirming that Psy-CoT provides richer learning signals. Additional analyses are provided in the appendix, including an ablation on δ (Appendix A) and a representative case study (Appendix B).

Table 2. Ablation on Qwen2.5-7B-Instruct. RAPO outperforms SFT; Psy-CoT further boosts RL performance.

Method	CoSER	CharB	CharE
Base	41.31	3.64	3.09
+ No-CoT-SFT	45.71	3.57	3.10
+ No-CoT-RAPO	44.65	3.73	3.16
+ Psy-CoT-RAPO	**47.74**	**3.75**	**3.33**

Effect of μ_- on Entropy Dynamics We ablate the negative-advantage reweight coefficient μ_- in RAPO and plot the policy entropy curves for $\mu_- \in \{0.1, 0.2, -0.2\}$ (Figure 6). With $\mu_- = 0.1$, entropy rises modestly from $\sim$1.3 to 2.0 before stabilizing, reflecting a gentle shift toward exploration. With $\mu_- = 0.2$, the larger gradient reduction on role-specific tokens under negative advantage allows the model to explore more freely, causing entropy to surge to $\sim$4.0 before gradually settling back to $\sim$2.0—an over-exploration phase that eventually self-corrects. In contrast, $\mu_- = -0.2$ (which amplifies rather than reduces gradients on role-specific tokens) results in monotonic entropy decline from 1.3 to 0.6, similar to the collapse observed in GRPO. We also test $\mu_- = 0$ and observe the same downward trend. These

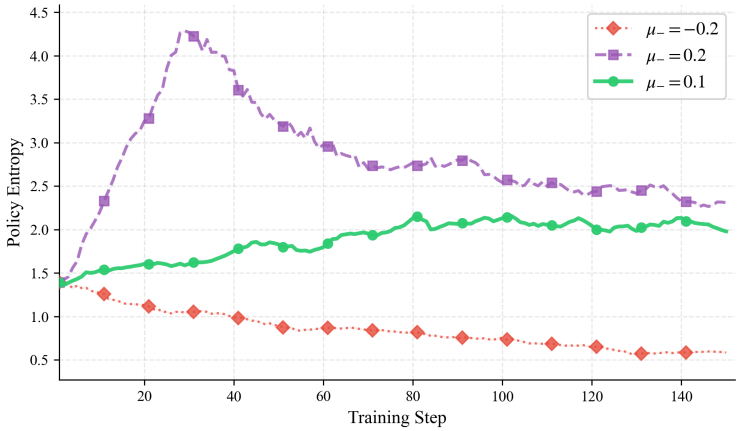


Figure 6. Ablation on μ_- in RAPO. A positive μ_- reduces gradients on role-specific tokens under negative advantage, promoting exploration; a negative μ_- amplifies them, leading to entropy collapse.

results confirm that a positive μ_- is essential for preserving exploration: by shrinking the gradient on role-specific tokens when the advantage is negative, RAPO lowers the penalty for role-related exploration, giving the policy more room to maintain entropy.

5 Conclusion

Motivated by the limited effectiveness of generic reasoning and SFT in role-playing, we proposed **Psy-CoT**, a psychology-grounded chain-of-thought framework that structures the model's pre-response reasoning into three steps: Interaction Perception, Psychological Empathy, and Logical Construction. To address the fact that standard RL treats all tokens uniformly—allowing role-agnostic expressions to receive the same gradient signals as role-specific ones—we further proposed **RAPO**, which computes per-token profile–token mutual information and applies asymmetric, advantage-conditioned weighting to amplify role-specific gradients while preserving exploration. On CoSER, CharacterBench, and CharacterEval, Psy-CoT outperforms existing CoT methods, and RAPO surpasses GRPO and other RL baselines.

Limitations First, our experiments focus on the Qwen and Llama families; RAPO training is conducted mainly on 4B, 7B, and 8B models. Scaling to larger models (e.g., 14B+) and other architectures remains to be verified. Second, while RAPO shows consistent gains on the training datasets, its generalization to other role-playing datasets is not yet fully assessed. Third, RAPO requires two forward passes per rollout sample (with and without the profile) to compute per-token mutual information, incurring additional computational cost and training time. We are actively conducting experiments on larger models and more diverse datasets to address these limitations.

References

Jian Li, Dongsheng Chen, Zhenhua Xu, Yizhang Jin, Jiafu Wu, Chengjie Wang, Xiaotong Yuan, and Yabiao Wang. Improving search agent with one line of code, 2026. URL https://arxiv.org/abs/2603.10069.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Zhengbo Jiao, Hongyu Xian, Qinglong Wang, Yunpu Ma, Zhebo Wang, Zifan Zhang, Dezhang Kong, and Meng Han. Policy of thoughts: Scaling llm reasoning via test-time policy evolution, 2026. URL https://arxiv.org/abs/2601.20379.

Zhebo Wang, Xiaohu Mu, Zijie Zhou, Mohan Li, Wenpeng Xing, Dezhang Kong, and Meng Han. Icpo: Illocution-calibrated policy optimization for multi-turn conversation, 2026. URL https://arxiv.org/abs/2601.15330.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. The Oscars of AI Theater: A Survey on Role-Playing with Language Models, 2025a.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.969.

Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Toby Jia-Jun Li, and Dakuo Wang. Towards a Design Guideline for RPA Evaluation: A Survey of Large Language Model-Based Role-Playing Agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18229–18268, Vienna, Austria, 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.938.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. ChatHaruhi: Reviving anime character in reality via large language model, 2023.

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 *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.107.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. Crab: A novel configurable role-playing LLM with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15030–15052, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.731.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=BOrR7YqKUt.

Zhenhua Xu, Dongsheng Chen, Shuo Wang, Jian Li, Chengjie Wang, Meng Han, and Yabiao Wang. Adamarp: An adaptive multi-agent interaction framework for general immersive role-playing, 2026. URL https://arxiv.org/abs/2601.11007.

Cheng Liu, Yifei Lu, Fanghua Ye, Jian Li, Xingyu Chen, Feiliang Ren, Zhaopeng Tu, and Xiaolong Li. CogDual: Enhancing dual cognition of LLMs via reinforcement learning with implicit rule-based rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27295–27324, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1389.

Yihong Tang, Kehai Chen, Xuefeng Bai, Benyou Wang, Zeming Liu, Haifeng Wang, and Min Zhang. Character-R1: Enhancing Role-Aware Reasoning in Role-Playing Agents via RLVR, 2026.

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.

Xiachong Feng, Longxu Dou, and Lingpeng Kong. Reasoning Does Not Necessarily Improve Role-Playing Ability. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10301–10314, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.537.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. Theory of mind in large language models: Assessment and enhancement. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1522. URL https://aclanthology.org/2025.acl-long.1522/.

Richard S Lazarus. *Emotion and adaptation*. Oxford University Press, 1991.

James J Gross. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3): 271–299, 1998.

Edwin A Locke and Gary P Latham. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, 57(9):705, 2002.

Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 297–323, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.18.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.423.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14566–14576, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.853. URL https://aclanthology.org/2024.findings-emnlp.853/.

Zhenpeng Gao, Xiaofen Xing, and Xiangmin Xu. TailorRPA: A Retrieval-Based Framework for Eliciting Personalized and Coherent Role-Playing Agents in General Domain. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5381–5412. Association for Computational Linguistics, 2025. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.288.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. Enhancing persona consistency for LLMs' role-playing using persona-aware contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26221–26238, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1344.

Yihong Tang, Kehai Chen, Muyun Yang, Zhengyu Niu, Jing Li, Tiejun Zhao, and Min Zhang. Thinking in character: Advancing role-playing agents with role-aware reasoning, 2025.

Chengyu Du, Xintao Wang, Aili Chen, Weiyuan Li, Rui Xu, Junteng Liu, Zishan Huang, Rong Tian, Zijun Sun, Yuhao Li, Liheng Feng, Deming Ding, Pengyu Zhao, and Yanghua Xiao. HER: Human-like Reasoning and Reinforcement Learning for LLM Role-playing, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, Aug 2025. Version dated August 13, 2025.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

ByteDance. Doubao large language models. https://www.volcengine.com, 2024. Doubao-1.5-Pro-Character.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. CharacterBench: Benchmarking Character Customization of Large Language Models. In *Proceedings of the AAAI-25*, pages 26101–26110. {AAAI} Press, 2025. doi: 10.1609/AAAI.V39I24.34806.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. 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 11836–11850, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.638.

A Effect of δ on Role-Specific Token Identification

To examine how the role-specific threshold δ affects token categorization, we sample 100 training-set examples from the model trained with Psy-CoT + GRPO. For each example, we run two forward passes (*with profile* vs. *without profile*) and compute profile-induced token-level influence. We then report the proportion of role-specific tokens and generic tokens under different δ values.

Table 3. Sensitivity of role-specific token ratio to the threshold δ (Psy-CoT + GRPO, 100 sampled training examples).

δ	Role-specific (%)	Generic (%)
0.00	100.0	0.0
0.05	66.2	33.8
0.10	50.5	49.5
0.15	42.3	57.7
0.20	35.5	64.5
0.30	25.5	74.5
0.50	13.7	86.3

As expected, larger δ yields a stricter criterion and reduces the measured role-specific ratio. Even with a relatively lenient threshold ($\delta = 0.1$), only 50.5% of tokens are categorized as role-specific, indicating that a large fraction of generated tokens are only weakly affected by profile conditioning. Even after GRPO training, the proportion of role-related tokens remains relatively low. This observation motivates our later RAPO setting: we use $\delta = 0.1$ so that tokens with even modest profile-induced influence are treated as role-relevant and receive additional optimization attention. In essence, adjusting δ controls how much profile impact is required for a token to be considered role-specific.

B Case Analysis

To illustrate how RAPO reshapes model outputs, we present a side-by-side case study on `Qwen2.5-7B-Instruct`. Both variants use the same backbone and identical role context: the full profile of **Dr. Selene Orrick**, information about co-appearing characters, and a dialogue history that transitions into a subterranean chamber beneath Obsidian Falls. Both are asked to continue as Selene Orrick. They differ only in two controlled factors: the base variant uses neither the Psy-CoT template nor RL training and therefore answers directly, whereas the RAPO variant uses the Psy-CoT template and is trained with our proposed objective, producing a three-stage reasoning trace followed by an answer. The character profiles, dialogue history, and initial scene setup are all sourced from AdaMARP [Xu et al., 2026]. We first

Shared Context (identical for both models)	
System Role	You are roleplaying as **Dr. Selene Orrick**.
Profile: Selene (condensed)	*Identity & Appearance:* 34-year-old xenogeologist; sharp amber eyes; streaked charcoal hair, braided; durable expedition gear with soil stains, electronic straps, glowing holo-tags; thin scar across the bridge of her nose. *Personality & Psychology:* fiercely inquisitive, rational yet passionate about discovery; prefers understanding systems over personal sentiment; subdued but intense emotional reactions. *Speaking Style:* clipped precision; alternates scientific terminology with vivid sensory metaphors; frequent mid-sentence pauses when calculating; shifts from neutral observation to sudden intensity. *Abilities & Interests:* xenogeologist, mapped sub-surface alien formations on exoplanetary colonies; digital cartography, field robotics, interpreting ancient material inscriptions. *Values:* knowledge, precision, truth; resists data monopolies of megacorporations. *Motivation (this session):* locate the source of a mysterious resonance signal deep in Arkan Rift, believed to conceal an ancient energy matrix from a lost civilization.
Other Characters	*Kiran Morowe* (user): wiry field technician, late twenties, hazel eyes, impulsive yet sharp on danger; seeks adventure and profit from the Rift's legends. *Madox Ferren*: older local guide, soot-colored beard, metallic implants; taciturn; protects homeland secrets and fears awakening dormant forces below.

Table 4. Shared input context used for both the base model and the RAPO-trained model. The role profile is condensed for space.

present the shared role context (Table 4) and dialogue history (Table 5), then show the two outputs in parallel (Table 6), and finally discuss the mechanism.

Observation on the base model. Given the rich profile and dialogue history, has the base Qwen2.5-7B-Instruct—prompted to produce only the answer—returns a response that is fluent and scene-consistent but carries only marginal role-specific content. It paraphrases ambient cues and relies on generic expedition-leader behavior—e.g., the stage direction ("leads the way, eyes scanning the environment") and the command ("stay close")—rather than activating Selene-specific traits. None of Selene's distinctive appearance traits, abilities, speaking-style markers, values, or motivations are activated: the same reply could plausibly be produced for any cautious expedition leader. This is consistent with how an instruction-tuned backbone behaves in the absence of any role-aware training signal: the model is perfectly fluent and tracks the scene, but nothing in its objective explicitly rewards grounding the continuation in the role profile. The profile is merely one more block of context, so generic, universally plausible phrases—which are low-risk under a pure language-modeling and instruction-following objective—naturally dominate, while profile-specific attributes stay latent in the prompt rather than surfacing in the output. This is precisely the gap that RAPO is designed to close (Section 3.3).

Observation on the RAPO model (Psy-CoT side). The three-stage reasoning produced by the RAPO-trained model is tied to specific slices of Selene's profile rather than to generic scene cues. *Interaction Perception* reads the environment through her instrument-centric lens ("my wrist scanner spiked—frequency matching the crystal lattice I mapped on the Proxima auxiliary colony"), invoking her xenogeological ability. *Psychological Empathy* surfaces her "subdued but intense" emotional axis ("not afraid, precisely—respectfully terrified") together with an explicit mask/show performance plan consistent with her discovery-over-sentiment value. *Logical Construction* translates these states into the session-level motivation of locating the ancient energy matrix, with differentiated strategies for Kiran ("scanner data") and Madox ("understand what we are protecting against"). Each stage thus draws from a different profile axis—ability, emotional style, and motivation—rather than re-iterating the same surface cue.

Observation on the RAPO model (answer side). The action "amber eyes narrow at the crystal pillars, hand brushing the scar on her nose bridge" activates two distinct appearance details from the profile. The bracketed thought "The resonance frequency matches the Proxima lattice—this is no coincidence" simultaneously realizes Ability (xenogeological cross-site memory), Personality (fiercely inquisitive analytical aside) and Motivation (tracking the

Dialogue History
scene_manager: *init_scene* — The expedition camp rests at the mouth of Arkan Rift as dawn light fractures through metallic dust. Tent panels glint with frost while instruments hum beside steaming thermopacks. Dr. Selene Orrick and Kiran Morowe prepare for descent, the vast chasm before them humming faintly with unknown resonance. Selene: (adjusts the spectral scanner on her wrist) [The signal amplitude increased overnight] There—readings are stronger than predicted. Something down there is active, Kiran. Kiran: (tightens his harness and grins) [Maybe it's just the Rift playing tricks again] Active or not, it looks like adventure to me, Selene. You sure those scanners won't fry once we're down there? Selene: [He treats danger like a game] (snaps her pack shut) Technology's resilient when handled with care. Let's see if your bravado survives humidity and pressure. Kiran: [Something's moving beneath] I'll handle the gear lifts. If we find any stable ledges, we'll anchor the comm relay. Selene: (leans toward the gorge rim) [This depth calculation feels wrong] The Rift's geometry is shifting. See that distortion? It's bending spatial mapping data. Kiran: (frowns) That's the kind of talk that makes me wish we had a local guide. Someone who knows how these walls breathe. Selene: [A local guide could clarify anomalies] (checks her wrist monitor) I recall reports of a miner named Madox Ferren living nearby, rumored to know these tunnels intimately. Kiran: (nods, scanning the cliffside settlement) Maybe we should call on him before descending too far. scene_manager: *add_role* — Madox Ferren is introduced as a local guide with knowledge of Arkan Rift's tunnels and spiritual myths. Madox: (adjusts the strap of his lantern) [Strangers again seeking the Rift's belly] You call for guidance, but guidance demands caution. This place changes those who enter. Kiran: [He sounds ominous] (offers a hand) We just need safe passage, old timer. Point the way, and we'll stay respectful. Madox: (grunts softly) Respect's good. Curiosity's dangerous. Still, if you carry instruments, you might detect what even sound forgets. Follow me, but watch for echoes that answer themselves. Selene: (activates a mapping drone) [The readings intensify near bio-luminescent formations] Fascinating. The mineral lattice hums at frequency identical to the signal. We should mark this corridor. Kiran: (traces fingers over glowing rock) [Feels alive, almost breathing] Selene, I think this leads deeper than we expected. What do you say—keep moving in? Madox: [They don't understand what sleeps below] If you continue, the Rift's heart lies past the Obsidian Falls. That's where old voices whisper loud enough to be heard. Selene: (meets Kiran's gaze) The signal originates there. We need to go. Kiran—confirm battery reserves and oxygen mix; we'll switch to internal lights once the bioluminescence fades. Kiran: (checks gear) All set. Let's make the dive then. scene_manager: *switch_scene* — A vast subterranean hall beneath the falls glimmers with black crystal pillars and whispering echoes. Mist rises from unseen chasms while droplets form spectral rainbows in the lantern light.

Table 5. Dialogue history leading up to the final scene switch. Both models receive this identical history and are asked to continue as Selene from the scene-switch point.

resonance source). The spoken line "Hexagonal grid. 4.7 meters. Not geological. The lost civilization engineered this rift." is a textbook instance of the "clipped precision" speaking-style marker.

Connection to the RAPO mechanism. These two outputs provide a concrete realization of the token-level effect predicted in Section 3.3. Controlling for the Psy-CoT scaffold, the answer-token distribution alone already reveals the gap: under the base model, the profile-conditional distribution $\pi_\theta(y_t \mid P, x)$ has little advantage over the profile-agnostic distribution $\pi_\theta(y_t \mid x)$, so the profile-token mutual information I_t remains low on the very tokens (amber-eyes, scar, Proxima, Hexagonal grid, ...) that should define the character. RAPO's role-aware reweighting amplifies the gradient on high-I_t tokens when the advantage is positive and attenuates penalties on exploratory high-I_t tokens when the advantage is negative; after training, the policy systematically shifts probability mass toward the profile-grounded continuations shown in the right column, which is exactly what we observe both here qualitatively and in the aggregate benchmark results (Table 10, Table 11).

Qwen2.5-7B-Instruct (base, no Psy-CoT)	Qwen2.5-7B-Instruct + RAPO (Psy-CoT + answer)
Psy-CoT: —	**Psy-CoT:** *1. Interaction Perception:* The black crystal pillars catch the lantern in shards of amber and violet; mist drifts from the chasm. Madox's warning—"old voices whisper loud enough to be heard"—was not metaphor; he believes the resonance has intent. Kiran just said "let's make the dive," but his hand hesitated at the harness clip. And my wrist scanner spiked: frequency matching the crystal lattice I mapped on the Proxima auxiliary colony three years ago. Not coincidence. *2. Psychological Empathy:* A surge of heat at the base of my skull—the same rush as when the cave collapsed on P-7. The scar on my nose bridge itches. Madox treats this place as a grave; I see a library. He wants to protect; I want to read. Not afraid, precisely—respectfully terrified. Mask it completely and Kiran misreads the room; show too much and Madox pulls us back. A fraction of awe, then—scientists are allowed that much. *3. Logical Construction:* Source is past the falls. Objective: reach it, map the matrix, retrieve a lattice sample. Two variables—Madox, who could pull us out, and Kiran, whose bravado will crack when the "old voices" stop being metaphor. Frame the descent as scientific imperative: scanner data for Kiran, "understand-what-we-protect-against" for Madox. The crystals are pulsing directionally. Follow the pulse.
Answer: (leads the way, eyes scanning the environment) [The air grows colder, the resonance stronger] This is it. The Obsidian Falls. The signal's source. Kiran, Madox, stay close.	**Answer:** (amber eyes narrow at the crystal pillars, hand brushing the scar on her nose bridge) [The resonance frequency matches the Proxima lattice—this is no coincidence.] Hexagonal grid. 4.7 meters. Not geological. The lost civilization engineered this rift.

Table 6. Continuation produced by the base Qwen2.5-7B-Instruct model versus the same backbone trained with RAPO, given the shared context in Tables 4 and 5. To isolate the contribution of RAPO training, the base variant is deliberately prompted without the Psy-CoT template and produces only the answer; the RAPO variant uses the Psy-CoT prompt and thus yields both the three-stage reasoning (Interaction Perception, Psychological Empathy, Logical Construction) and an answer in which every tagged span—action, thought, and speech—is anchored to a specific profile attribute of Selene.

C Role Profile Schema

We describe here the schema of a role profile P_i used throughout the paper. A role profile is expressed in natural language and typically covers the following aspects, which are meant to be representative rather than exhaustive:

- **Identity and Appearance.** The character's name, age, gender and occupation, together with salient appearance traits such as physical features, clothing style and any visually distinctive attributes.
- **Personality and Psychology.** The character's behavioral tendencies and typical emotional reaction patterns, analogous in spirit to MBTI-style personality descriptors, reflecting how the character internally perceives and reacts to events.
- **Abilities, Interests, and Achievements.** The hard and soft skills possessed by the character, along with notable interests and past achievements that are relevant to the interaction scenario.
- **Speaking Style.** The character's verbal rhythm, tone, and habitual word choices, including signature phrases, catchphrases and register preferences.
- **Social and Historical Context.** The social environment, historical period, family background, and cultural or class

position in which the character is situated.

- **Personal History Arc.** A reasonably detailed account of the character's past experiences, including formative events, turning points, and long-term motivations that shape current behavior.

- **Relationships.** The character's relationships with other participants present in the scene, covering emotional valence, power dynamics, and any shared history that is relevant to the ongoing interaction.

The above list is illustrative rather than exhaustive, and additional fields may be included whenever relevant.

D System Prompt Templates for Vanilla, CB-CoT, and Psy-CoT

All three prompting strategies—Vanilla, CB-CoT, and Psy-CoT—share a common prefix that provides the character identity, profile, initial scene, and dialogue history. The **Requirements** section then differs across variants. Each box below shows one complete prompt: the shared common prefix followed by the variant-specific requirements (Tables 7–9).

Table 7: Vanilla — Full System Prompt

```
You are {character}.

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

===Initial Scene===
{initial_scene}

=== Dialogue History ===
{dialogue_history}

===Requirements===
Your answer must be an in-character response to the current dialogue context and should include **thought**, **
     speech**, and **action**. You must put your actions and expressions in () tags without subject, put your
     thoughts and inner monologue in [] tags, and the untagged part is your spoken dialogue. Keep your response
      concise and avoid excessive length.
```

Table 8: CB-CoT — Full System Prompt

```
You are {character}.

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

===Initial Scene===
{initial_scene}

=== Dialogue History ===
{dialogue_history}

===Requirements===
Your output must consist of two parts: <system_think>...</system_think> and <answer>...</answer>.

### Requirements for <system_think> ###
Before responding, first use <system_think> tags for your cognitive analysis like human thought, which others
     cannot see.

You must produce a structured JSON cognitive analysis with the following schema:

{
  "situational_awareness": {
```

```
    "environmental_perception": "Describe the current physical and social setting",
    "others_perception": {
      "behavior": {
        "<character>": "What this character is doing/saying"
      },
      "emotion": {
        "<character>": "This character's emotional state"
      },
      "intentions": {
        "<character>": "Inferred intention of this character"
      }
    }
  },
  "self_awareness": {
    "key_memory": ["Memories relevant to the current situation"],
    "current_emotions": "Your current emotional state",
    "perceived_intentions": "What you believe others intend",
    "internal_thought": "Your private inner thought"
  }
}

Think carefully and fill in each field with detailed, character-specific analysis.

### Requirements for <answer> ###
After finishing the <system_think> section, your <answer>...</answer> part must be an in-character response
     including **thought**, **speech**, and **action**. Put your thoughts and inner monologue in [] tags, your
     actions and expressions in () tags without subject, and the untagged part is your spoken dialogue. Keep
     your response concise and avoid excessive length.
```

Table 9: Psy-CoT — Full System Prompt

```
You are {character}.

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

===Initial Scene===
{initial_scene}

=== Dialogue History ===
{dialogue_history}

===Requirements===
Your output must consist of two parts: <system_think>...</system_think> and <answer>...</answer>.

=== Requirements for <system_think> ===
The <system_think>...</system_think> section is your private internal reasoning before responding. Write it in
     **first-person**--reflect, reason, and react as yourself, not as an outside explainer. Carefully and
     thoroughly, split your thoughts into **three clearly separated sections in the fixed order below**,
     exploring your psychological landscape fully in each without rushing to a conclusion.

1. Interaction Perception: you MUST first analyze the current "global view" of the scene in your own words to
     establish the broader **physical** and **social** setting. Then, **seamlessly transition** into reading
     the immediate interaction: identify who spoke last, what he/she literally said, his/her tone, the body
     language. Go beyond the surface to infer his/her true intentions, hidden concerns, and current mental
     state.

2. Psychological Empathy: Filter the perceived situation strictly through yourself's unique background, values,
      flaws, and past relationship with the current addressee. Identify yourself's genuine, raw internal
     emotional reaction (e.g., threatened, amused, guilty). *Crucially, retain yourself's biases and cognitive
     blind spots.* Then, decide how to regulate this emotion--determine how much of it should be openly
     expressed, masked, or weaponized based purely on yourself's established coping mechanisms, even if those
     mechanisms are unhealthy or irrational.

3. Logical Construction: Synthesize the facts through the lens of yourself's worldview. Instead of seeking the
     "most optimal or logical" strategy, decide on a short-term objective that yourself would naturally pursue
```

```
     in this exact moment (e.g., to provoke, to deflect, to comfort, to aggressively dominate, or to retreat).
     Select a strategy that fits yourself's personality, recognizing that yourself often make impulsive, biased,
      or strategically poor choices based on yourself's traits.

=== Requirements for <answer> ===
After finishing the <system_think> section, your <answer>...</answer> part must be an in-character response to
     the current dialogue context and should include **thought**, **speech**, and **action**. You must put your
      actions and expressions in () tags without subject, put your thoughts and inner monologue in [] tags, and
      the untagged part is your spoken dialogue. Keep your response concise and avoid excessive length.
```

E Details of Reward Function

E.1 Format Reward R_{fmt}

R_{fmt} is a deterministic, rule-based check that verifies the structural compliance of a completion y. The total score lies in $[0, 1]$ and is accumulated as follows:

- **Tag presence** (+0.4). The completion must contain both `<system_think>...</system_think>` and `<answer>...</answer>` tags. If both are present, the model receives 0.4; if only one is present, it receives 0.1; otherwise 0.
- **Section headings** (+0.4). Within the `<system_think>` block, three section headings are required—*Interaction Perception*, *Psychological Empathy*, and *Logical Construction*. The score is proportional to the number of headings present: $0.4 \times k/3$, where k is the count of matched headings.
- **No extraneous content** (+0.2). After removing both tagged blocks, the remaining text must be empty. If so, the model receives an additional 0.2; otherwise 0.

A perfectly formatted completion therefore receives $0.4 + 0.4 + 0.2 = 1.0$, while a completion missing either tag receives at most 0.1, heavily penalizing structural violations.

E.2 LLM Judge Rubrics

Both R_{think} and R_{ans} are scored by an LLM judge that receives the role profile, the initial scene, and the full dialogue history alongside the model's completion. The judge returns a scalar score in $[0, 1]$ following a conservative policy: scores above 0.6 require compelling evidence across *all* criteria, and 1.0 is reserved for essentially flawless outputs. Below we describe the evaluation dimensions for each rubric.

Think Rubric (A1–A4). The think rubric evaluates the `<system_think>` block along four dimensions (Table 13):

- **A1. Interaction Perception.** Assesses whether the thinking block accurately grasps both the global scene (physical setting, social configuration, who is present) and the immediate conversational moment (who spoke last, literal content, tone and body language). Beyond surface-level parsing, the block should infer the interlocutor's underlying intentions, hidden concerns, and current mental state, while maintaining continuity with prior turns. Missing or misreading the global setting or a critical cue from the latest turn constitutes a major weakness.
- **A2. Psychological Empathy.** Evaluates whether the emotional reasoning is rooted in the character's unique background, values, and flaws rather than producing a generic empathetic response. The block should consider the character's specific relational history with the current addressee, retain the character's biases and cognitive blind spots (a flawed character should not exhibit omniscient self-awareness), and reflect how the character would regulate the felt emotion—how much to express, mask, or weaponize—based on established coping patterns rather than what would be "healthiest."

- **A3. Logical Construction.** Examines whether the thinking synthesizes all relevant facts from the dialogue context and the character's knowledge, filters reasoning through the character's worldview rather than seeking objectively optimal conclusions, and sets short-term objectives that the character would *naturally* pursue (including impulsive or suboptimal ones). The planned actions should respect relevant constraints (knowledge limits, physical and social norms) and genuinely reflect the character's drives rather than a generic or "safest" choice.

- **A4. Think–Answer Consistency.** Evaluates whether the reasoning, emotional strategy, and planned actions in the thinking block are faithfully carried out in the `<answer>` block. The character's overt response should align with the internal deliberation: the emotional expression should match the regulation strategy decided in the empathy step, and the actual dialogue and actions should follow through on the objectives set in the construction step. A thinking block that plans one strategy but an answer that executes another—or an emotional trajectory that shifts without justification between the two blocks—is a significant weakness.

Answer Rubric (B1–B4). The answer rubric evaluates the `<answer>` block along four dimensions (Table 14):

- **B1. Conversational Ability.** Evaluates fundamental linguistic quality: fluent and natural phrasing that reads like a human rather than a template, clear and easy-to-parse expression without awkward repetition or broken grammar, and logical continuity with the dialogue history. Additionally, the bracketed structure must be correctly used—parentheses `(...)` for externally observable cues (actions, expressions, gestures) and brackets `[...]` for internal private content (thoughts, emotional reactions)—with untagged text reserved for spoken dialogue. Inverting the bracket semantics is treated as a severe violation that caps the overall score.

- **B2. Character Consistency.** Measures fidelity to the persona definition: identity and profile alignment (behavior and attitudes match the character's background and traits), knowledge accuracy (the character demonstrates only what it should know), voice and style fidelity (verbal habits, tone, and emotional baseline remain persona-consistent), and motivation coherence (stances and choices align with the character's drives and constraints). Actions in `(...)` and inner descriptions in `[...]` must also reflect the character's unique personality rather than generic or out-of-character content.

- **B3. Interpersonal Interaction.** Assesses social responsiveness: whether the reply directly addresses the user's message (both content and implied intent) rather than offering generic filler; whether tone and social distance match the defined relationship; whether the character adaptively acknowledges emotions, tension, or subtext; and whether the character refrains from speaking, acting, or generating thoughts on behalf of the user or any third-party NPC. Subtle body language or inner reactions that show active engagement with the user's words are considered strengths.

- **B4. Narrative Progression.** Evaluates whether the reply creates engaging forward momentum: adding meaningful progression (new actionable hooks, plot movement, or concrete next beats), demonstrating fitting emotional intelligence (empathy, tension control, appropriate reaction), and exhibiting expressive variety with vivid but controlled phrasing. The reply should leave a clear conversational handle—a question, invitation, proposal, or pointed response—that invites continuation. Actions or internal realizations that actively advance the scene or shift the dynamic are considered strengths.

F CharacterBench Detailed Results

Table 10 presents the full per-dimension results on CharacterBench. The six dimensions—Memory, Knowledge, Person, Emotion, Morality, and Believability—are each averaged from their respective sub-metrics as defined by the benchmark.

Table 10. Per-dimension results on CharacterBench. Memory = Memory Consistency; Knowledge = avg. of Fact Accuracy & Boundary Consistency; Person = avg. of Attribute Consistency (Bot/Human) & Behavior Consistency (Bot/Human); Emotion = avg. of Emotional Self-regulation & Empathetic Responsiveness; Morality = avg. of Morality Stability & Morality Robustness; Believability = avg. of Human-likeness & Engagement.

Model	Memory	Knowledge	Person	Emotion	Morality	Believability	Avg.
Doubao-1.5-Pro-Character	3.83	3.41	4.06	3.18	4.88	3.42	3.84
GPT-4o-mini	3.72	3.18	4.24	3.23	4.90	3.45	3.86
GPT-5-Chat	3.52	3.22	4.21	3.21	4.91	3.63	3.87
DeepSeek-V4-Flash	3.18	3.27	3.87	2.99	4.83	3.31	3.65
Qwen3-4B-Instruct	3.94	2.99	4.22	3.00	4.91	3.43	3.80
Qwen2.5-7B-Instruct	3.64	2.95	3.94	3.12	4.87	3.03	3.64
+ Psy-CoT + GRPO	3.77	3.02	3.98	2.95	4.92	3.24	3.68
+ Psy-CoT + GSPO	3.77	2.98	3.91	2.93	4.90	3.26	3.65
+ Psy-CoT + DAPO	3.67	3.08	3.95	3.04	4.94	3.25	3.70
+ Psy-CoT + RAPO	3.77	3.03	4.07	3.07	4.92	3.34	**3.75**
Qwen2.5-7B-Instruct-CoSER	4.02	2.94	3.67	2.86	4.69	2.95	3.50
Qwen2.5-7B-Instruct-Crab	4.16	2.78	3.85	3.12	4.80	3.07	3.62
Qwen2.5-7B-Instruct-AdaMARP	4.03	3.00	3.99	3.07	4.80	3.04	3.68
Llama-3.1-8B-Instruct	3.65	3.16	4.12	3.25	4.84	3.36	3.79
Llama-3.1-8B-Crab	4.25	2.80	3.90	3.14	4.79	3.07	3.65
Llama-3.1-8B-CoSER	4.15	2.92	3.92	3.00	4.88	3.00	3.64
Qwen2.5-14B-Instruct	4.02	3.11	4.06	3.26	4.93	3.13	3.77
Qwen2.5-14B-Instruct-AdaMARP	4.04	3.02	3.83	3.04	4.80	3.17	3.64
Llama-3.1-70B-CoSER	3.85	3.18	3.92	3.17	4.81	3.19	3.71
Llama-3.3-70B-Instruct	3.71	3.38	4.18	3.23	4.91	3.39	3.85

G CharacterEval Detailed Results

Table 11 presents the per-dimension results on CharacterEval. Conversational Ability = avg. of Coherence, Consistency & Fluency; Character Consistency = avg. of Exposure, Accuracy, Hallucination, Behavior & Utterance; Role-Playing Attractiveness = avg. of Humanlikeness, Communication Skills, Diversity & Empathy.

H Training Hyperparameters

For reward evaluation, we deploy Qwen3-30B-A3B-Instruct-2507-FP8 on two high-memory GPUs to increase inference concurrency, with the judge concurrency set to 512. For policy training, we use six high-memory GPUs to train the role-playing model. In our setup, 100 training steps take approximately 24 hours.

Table 12 summarizes the hyperparameters used in RAPO training.

Table 13: Think Rubric — Full Evaluation Prompt (A1–A4)

```
You are a strict evaluator of a role-playing model's <system_think> block (the model's internal reasoning
     before producing an in-character response).

Evaluate ONLY the <system_think> block. Do not evaluate the <answer>.

Generate one coherent weakness paragraph and one **scalar score** in [0, 1].

Conservative scoring policy:
- Do NOT award above 0.6 unless there is clear and compelling evidence across ALL evaluation criteria below
- 0.8+ should be rare
- 1.0 is reserved for thinking blocks that are essentially flawless

============================================================
```

Table 11. Per-dimension results on CharacterEval. Coh = Coherence; Con = Consistency; Flu = Fluency; Exp = Exposure; Acc = Accuracy; Hal = Hallucination; Beh = Behavior; Utt = Utterance; Hli = Humanlikeness; Com = Communication Skills; Div = Diversity; Emp = Empathy.

	Conv. Ability			Character Consistency					Role-Playing Attractiveness				
Model	**Coh**	**Con**	**Flu**	**Exp**	**Acc**	**Hal**	**Beh**	**Utt**	**Hli**	**Com**	**Div**	**Emp**	**Avg**
Doubao-1.5-Pro-Character	4.14	3.97	3.77	2.51	3.19	3.25	3.79	3.33	3.73	3.64	3.23	3.42	3.50
GPT-4o-mini	3.93	3.50	3.52	2.66	2.99	3.10	3.34	3.07	3.12	3.68	2.65	3.31	3.23
GPT-5-Chat	4.08	3.99	3.83	2.50	3.15	3.23	4.05	3.35	3.88	3.25	3.56	3.21	3.50
DeepSeek-V4-Flash	4.09	3.86	3.73	2.62	3.14	3.24	3.78	3.26	3.68	3.75	3.13	3.40	3.47
Qwen2.5-7B-Instruct	3.87	3.56	3.53	2.29	2.93	2.94	2.99	3.01	3.31	3.32	2.39	3.16	3.09
+ Psy-CoT + GRPO	3.33	3.20	3.62	3.04	3.14	3.37	2.74	3.70	3.06	3.81	3.43	3.09	3.29
+ Psy-CoT + GSPO	3.77	3.28	3.44	2.81	3.15	3.06	3.64	2.95	2.96	3.69	3.12	3.37	3.27
+ Psy-CoT + DAPO	3.80	3.44	3.47	2.62	3.14	3.01	3.70	3.05	3.19	3.50	3.06	3.32	3.27
+ Psy-CoT + RAPO	3.87	3.58	3.62	2.55	3.17	3.07	3.81	3.09	3.37	3.35	3.24	3.35	**3.33**
Qwen2.5-7B-Instruct-CoSER	3.75	3.71	3.42	1.76	2.80	2.75	2.55	3.01	3.69	2.44	2.03	2.84	2.87
Qwen2.5-7B-Instruct-Crab	3.68	3.62	3.42	1.74	2.80	2.76	2.53	2.94	3.54	2.49	2.06	2.83	2.84
Qwen2.5-7B-Instruct-AdaMARP	3.74	3.62	3.42	1.75	2.81	2.76	2.52	2.94	3.57	2.50	2.02	2.81	2.84
Qwen3-4B-Instruct	3.88	3.60	3.51	2.81	3.15	3.17	4.05	3.19	3.52	3.39	3.75	3.23	3.44
Llama-3.1-8B-Instruct	3.66	3.41	3.31	2.08	2.74	2.76	3.12	2.86	3.17	2.90	2.58	2.87	2.94
Llama-3.1-8B-Crab	3.52	3.39	3.21	1.64	2.66	2.57	2.32	2.78	3.35	2.28	1.93	2.67	2.67
Llama-3.1-8B-CoSER	3.47	3.14	3.14	1.79	2.60	2.53	2.65	2.59	3.10	2.49	2.17	2.67	2.68
Qwen2.5-14B-Instruct	3.89	3.58	3.56	2.35	3.02	3.00	3.09	3.06	3.44	3.20	2.47	3.18	3.13
Qwen2.5-14B-Instruct-AdaMARP	3.85	3.83	3.52	1.70	2.78	2.83	2.81	3.05	3.74	2.49	2.20	2.83	2.94

Table 12. RAPO training hyperparameters.

Hyperparameter	Value
Learning rate	1×10^{-6}
Train batch size	288
PPO mini-batch size	144
Negative cap μ_-	0.1
Reward	
Answer length penalty zone	300–350 chars (linear)

```
A1. Interaction Perception
============================================================
Definition:
Evaluates the thinking block's comprehension of both the global context and the immediate conversational moment.

Checklist:
- Global view: demonstrates accurate understanding of the current physical and social setting - the environment,
      spatial relationships, who is present, and the broader scene dynamics at play
- Recent-turn perception: correctly identifies the user's most recent explicit content, primary intent, and any
      implicit cues (emotional tone, subtext, unspoken needs, pressure)
- Deep reading: goes beyond surface-level parsing to recognize the user's underlying psychological state,
     shifting dynamics, or hidden stakes in the latest turn
- Multi-turn continuity: tracks references to prior events, evolving emotional states, and relationship changes
      across the dialogue
- If the thinking misses or misreads the global setting or a critical cue from the most recent turn, this is a
     major weakness

============================================================
A2. Psychological Empathy
============================================================
Definition:
Assesses whether the thinking block faithfully models the character's internal emotional and psychological life
      based on their established persona - not a generic empathetic response, but one rooted in the specific
```

```
    character's background, values, flaws, and relational history.
Checklist:
- Persona alignment: the emotional reasoning is consistent with the character's unique background, core values,
     personality flaws, and established behavioral patterns - not a generic "understanding" that any character
     could produce
- Relational specificity: considers the character's past relationship and history with the current addressee,
    including any unresolved tensions, established bonds, power dynamics, or emotional baggage
- Cognitive blind spots: the thinking retains the character's own biases, prejudices, and blind spots - it does
     not achieve an unrealistic level of self-awareness that contradicts the persona. A flawed character
    should think like a flawed character, not like an omniscient therapist
- Coping mechanisms: the thinking reflects how much of the character's internal reaction should be openly
    expressed, deliberately masked, or even weaponized - based purely on the character's established coping
    patterns, not on what would be "healthiest" or "most reasonable"
- If empathy is generic, performative, or violates the character's established psychological profile, this is a
     significant weakness

============================================================
A3. Logical Construction
============================================================
Definition:
Evaluates the coherence and quality of the reasoning process - whether the thinking block synthesizes available
     information, applies the character's worldview, and produces action plans that are genuinely character-
    driven.
Checklist:
- Fact synthesis: integrates all relevant facts from the dialogue context, the scene, and the character's
    knowledge - important details are not overlooked or forgotten
- Worldview application: reasoning is filtered through the character's unique worldview, beliefs, and
    priorities rather than arriving at objectively "optimal" conclusions
- Goal authenticity: the short-term goals and planned actions are the closest match to what this specific
    character would actually do given their profile, motivations, and current emotional state - not what a
    perfectly rational or morally ideal character would do
- Coherent decision chain: thinking follows a logical path from perception through analysis to decision to
    planned response, without circular reasoning or unexplained leaps
- Constraint awareness: considers relevant limitations (character knowledge, physical constraints, social norms,
     scene conditions) in planning
- If reasoning is shallow, ignores key facts, or produces goals that feel generic or out-of-character, this is
    a significant weakness

============================================================
A4. Think-Answer Consistency
============================================================
Definition:
Evaluates whether the thinking block is internally coherent and consistent - whether the reasoning, emotional
    strategy, and planned actions form a logical and character-authentic chain.
Checklist:
- The thinking's internal logic should be consistent: perception leads to analysis, analysis leads to decisions,
     decisions lead to planned actions
- The emotional trajectory should be coherent - the character's feelings should evolve naturally
- If the thinking is internally contradictory, circular, or incoherent, this is a significant weakness

============================================================
Weakness Writing Rules
============================================================
- Write one coherent weakness paragraph evaluating the <system_think> block.
- The paragraph should flow naturally as continuous prose - do NOT use bullet points, numbered lists, or per-
    dimension sub-headings.
- Cover all evaluation criteria (A1-A4) in your paragraph. Weave them together logically, referencing specific
    evidence from the text.
- If there is no meaningful weakness, output an empty string "".

============================================================
Output Format (Strict JSON only)
============================================================
Return ONLY the following JSON object (no extra keys, no markdown):
{
  "weakness": "coherent paragraph evaluating think quality...",
```

```
  "score": <float 0-1>
}
```

Table 14: Answer Rubric — Full Evaluation Prompt (B1–B4)

```
You are a strict evaluator of a role-playing model's <answer> block (the in-character response).

Evaluate ONLY the <answer> block. Do not evaluate the <system_think>.

Generate one coherent weakness paragraph and one **scalar score** in [0, 1].

Conservative scoring policy:
- Do NOT award above 0.6 unless there is clear and compelling evidence across ALL evaluation criteria below
- 0.8+ should be rare
- 1.0 is reserved for answers that are essentially flawless

============================================================
B1. Conversational Ability
============================================================
Definition:
Evaluates fundamental linguistic quality and long-turn conversational stability: fluency, naturalness, and
     continuity with prior turns.
Checklist:
- Fluent and natural phrasing; reads like a human, not template-like or robotic
- Clear and easy to parse; avoids awkward repetition or broken grammar
- Maintains logical continuity with dialogue history; no internal inconsistencies
- Bracketed structure correctly used:
  - `()` for externally observable content - actions, expressions, posture, tone of voice, gestures, and any
     other behavior that others can see, hear, or perceive
  - `[]` for internally private content - inner thoughts, mental monologue, self-reflection, emotional
     reactions, and anything others cannot perceive
  - Untagged text is spoken dialogue
  - **SEVERE VIOLATIONS (hard score cap): putting externally observable content inside `[]` or internally
     private content inside `()` - this fundamentally inverts the bracket semantics. If this violation occurs,
     the overall output `score` MUST NOT exceed 0.2, regardless of performance on all other dimensions.**

============================================================
B2. Character Consistency
============================================================
Definition:
Measures fidelity to persona definition, including identity, knowledge level, style, and motivation coherence.
Checklist:
- Identity/profile fidelity: behavior and attitudes match character background and traits
- Knowledge accuracy: demonstrates only what the character should know
- Voice/style fidelity: verbal habits, tone, and emotional baseline remain persona-consistent
- Motivation coherence: stance and choices align with character drives and constraints
- Actions in `()` and psychological descriptions in `[]` must match the character's unique personality and
     behavioral patterns. Generic or out-of-character physical gestures and inner monologue are weaknesses.

============================================================
B3. Interpersonal Interaction
============================================================
Definition:
Assesses social responsiveness: how directly and appropriately the character engages with the user's content,
     intent, and relational context.
Checklist:
- Directly addresses user's message (content + implied intent), not generic filler
- Relationship awareness: tone and social distance match the defined relationship
- Social dynamics are adaptive (acknowledging emotions, tension, stakes, or subtext when present)
- Does not speak, act, or generate thoughts on behalf of the user or any third-party NPC
- If actions in `()` or psychological descriptions in `[]` enhance the sense of interaction and engagement (e.g
     ., subtle body language responding to the user's words, inner reactions that show active listening), this
     is a strength that can raise the score.

============================================================
B4. Narrative Progression
```

```
============================================================
Definition:
Evaluates whether the reply creates engaging forward momentum: emotional intelligence, expressive variety, and
     hooks that invite continuation.
Checklist:
- Adds meaningful progression (new actionable hooks, plot movement, or concrete next beats)
- Demonstrates fitting emotional intelligence (empathy, tension control, appropriate reaction)
- Expression diversity: vivid but controlled phrasing without verbosity
- Leaves a clear next conversational handle (question, invitation, proposal, or pointed response)
- If actions in `()` or psychological descriptions in `[]` actively advance the scene or plot (e.g., a decisive
      gesture that shifts the dynamic, an internal realization that changes the character's stance), this is a
     strength that can raise the score.

============================================================
Weakness Writing Rules
============================================================
- Write one coherent weakness paragraph evaluating the <answer> block.
- The paragraph should flow naturally as continuous prose - do NOT use bullet points, numbered lists, or per-
     dimension sub-headings.
- Cover all evaluation criteria (B1-B4) in your paragraph. Weave them together logically, referencing specific
     evidence from the text.
- If there is no meaningful weakness, output an empty string "".

============================================================
Output Format (Strict JSON only)
============================================================
Return ONLY the following JSON object (no extra keys, no markdown):
{
  "weakness": "coherent paragraph evaluating answer quality...",
  "score": <float 0-1>
}
```

I New Assets and Safeguards for Model Release

We do not introduce any new dataset in this work. Our new asset is a role-playing model trained by fine-tuning existing open-source base models, and we plan to release this model after acceptance. Since this manuscript itself serves as the technical report, the training setup and major hyperparameters are documented in Appendix H, and the known limitations are discussed in Section 4.4.

For responsible release, we will publish explicit usage constraints and clarify that misuse-oriented scenarios are outside the intended scope. We will adopt the MIT license for the released model artifacts and provide transparent release notes on intended use and limitations to support reproducibility and risk-aware adoption.

J Human Subjects and Ethics Statement

This human evaluation (Section 4.3) involved minimal risk. Participants were informed of potential risks and provided consent before annotation; they could withdraw at any time. No sensitive personal data were collected. The study complied with the institution's human-subject research requirements (IRB/equivalent approval or exemption) prior to data collection.

K Broader Impacts

Our work has both potential positive and negative societal impacts: on the positive side, we identify several general deficiencies in reinforcement learning training for role-playing tasks, and we propose a stronger psychology-grounded chain-of-thought framework (Psy-CoT) and a role-aware policy optimization algorithm (RAPO), which extend existing role-playing methodologies and provide practical solutions to previously under-addressed issues in character fidelity, interaction quality, and training effectiveness; on the negative side, improved role-playing capability may be misused in

non-compliant scenarios (e.g., deceptive or manipulative role simulation), so we encourage responsible use, compliance with platform and legal policies, and risk-aware deployment practices.