HER: Human-like Reasoning and Reinforcement Learning for LLM Role-playing

Chengyu Du[1,2] **Xintao Wang**[1] **Aili Chen**[1,2] **Weiyuan Li**[1] **Rui Xu**[1] **Junteng Liu**[2] **Zishan Huang**[2]
Rong Tian[2] **Zijun Sun**[2] **Yuhao Li**[2] **Liheng Feng**[2] **Deming Ding**[2] **Pengyu Zhao**[*2] **Yanghua Xiao**[*1]

[1]Fudan University [2]MiniMax

Abstract

LLM role-playing, i.e., using LLMs to simulate specific personas, has emerged as a key capability in various applications, such as companionship, content creation and digital games. While current models effectively capture character tones and knowledge, simulating the inner thoughts behind their behaviors remains a challenge. Towards cognitive simulation in LLM role-play, previous efforts mainly suffer from two deficiencies: lacking data with high-quality reasoning traces, and lacking reliable reward signals aligned with human preferences. In this paper, we propose HER, a unified framework for cognitive-level persona simulation. HER introduces dual-layer thinking, which distinguishes characters' first-person thinking from LLMs' third-person thinking. To bridge these gaps, we curate reasoning-augmented role-playing data via reverse engineering, and construct human-aligned principles and reward models. Leveraging these resources, we train HER models based on Qwen3-32B via supervised and reinforcement learning. Extensive experiments validate the effectiveness of our approach. Notably, our models significantly outperform the Qwen3-32B baseline, achieving a 30.26% improvement on the CoSER benchmark and a 14.97% gain on the Minimax Role-Play Bench. Our datasets, principles, and models will be released to facilitate future research. [1] [2] [3] [4]

The Reasoning-Driven Role-play Framework

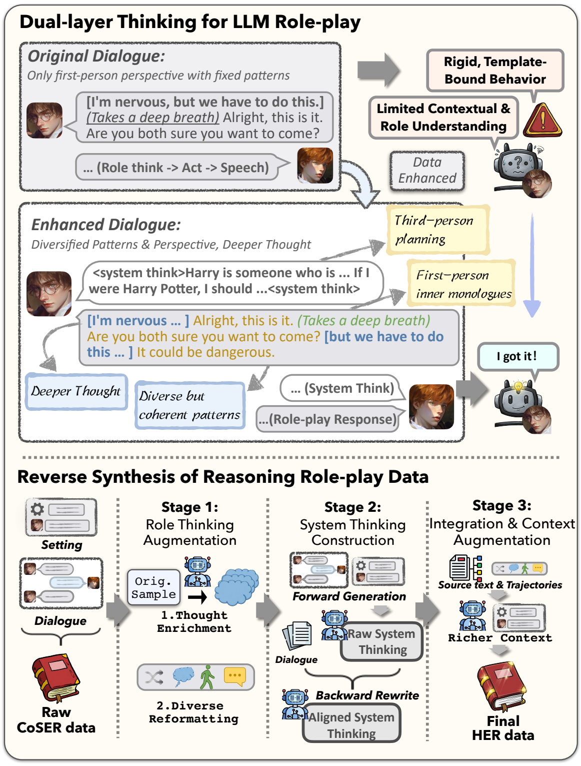

Figure 1: **The reasoning-driven LLM role-play framework of HER.** HER introduces Dual-layer Thinking and a three-stage reverse synthesis pipeline to construct reasoning-augmented LLM role-play trajectories.

1 Introduction

LLM role-playing, broadly viewed as persona simulation, aims to generate in-character decisions and narratives conditioned on a persona and an evolving scene (Chen et al., 2024a). In this work, we focus on text-based, multi-turn dialogue role-playing, where an agent must remain in character throughout an interactive conversation. Large language models (LLMs) have demonstrated strong general-purpose language capabilities, largely attributed to large-scale pretraining on natural language corpora, as evidenced by modern frontier LLMs (Liu et al., 2025a; Yang et al., 2025). Recently, post-training that emphasizes reasoning and reinforcement learning (RL) has become a central route to further improve model capabilities beyond imitation (OpenAI et al., 2024; MiniMax et al., 2025). These advances have accelerated role-play applications in companionship, interactive storytelling, and game-like settings, where users expect persis-

[1]https://github.com/cydu24/HER
[2]https://huggingface.co/ChengyuDu0123/HER-32B
[3]https://huggingface.co/ChengyuDu0123/HER-RM-32B
[4]https://huggingface.co/datasets/ChengyuDu0123/HER-Dataset

Findings of the Association for Computational Linguistics: ACL 2026, pages 25725–25762
July 2-7, 2026 ©2026 Association for Computational Linguistics

Training Sample Format	
Input Context $P, x_{\le t}$	**Profile**: *Elizabeth Bennet*—quick-witted, independent, despises false pride. **Scenario**: Encounters Mr. Darcy at Pemberley; previous conflicts unresolved.
System Thinking	<system_thinking> I need to play as ...,My Persona is ...; Now the Scene is tense reunion... Plan: stay polite, deflect with irony... </system_thinking>
Role-play Answer	<role_thinking>Why does he look at me so?</role_thinking> <role_action>raises eyebrow</role_action> "The grounds are unexpectedly pleasant, Mr. Darcy."

Table 1: Training sample format. **Input**: profile + scenario + history. **Output**: system thinking (3rd-person, hidden) + role-level content (1st-person, visible) evaluated by the GRM. Details in Table 28.

tent personas and coherent long-form interactions.

However, while current LLMs often mimic surface attributes such as speaking style or factual knowledge, deeply emulating a character's inner reasoning—the motivations and plans that connect persona and scene constraints to the next-turn decision—remains challenging. The role-playing performance of current reasoning-capable LLMs is still not fully satisfactory and remains to be improved (Liu et al., 2025b; Ye et al., 2025a). Meanwhile, datasets have started to include character inner thoughts, such as CoSER (Wang et al., 2024c), but these thoughts are often short and shallow, providing limited supervision for deep persona-grounded reasoning.

Enabling LLMs to deeply simulate inner reasoning in multi-turn dialogue role-play is difficult for two reasons. First, although previous efforts have curated role-play dialogues, their underlying planning and inner thoughts are usually implicit, making scalable supervision of reasoning traces costly and inconsistent (Wang et al., 2024d; Tu et al., 2024a). For example, the same dialogue turn can be justified by multiple plausible inner motivations, making reasoning annotations inherently subjective and hard to standardize at scale. Second, role-play outputs are inherently open-ended and non-verifiable, making reward modeling and preference optimization prone to bias and shortcut learning, especially when rewards can be exploited by superficial cues (Liao et al., 2025). Moreover, rewards can be easily gamed by superficial cues (e.g., verbosity or sentiment words), leading models to optimize for style rather than genuine persona-grounded decisions.

Therefore, effective optimization for reasoning-driven role-play calls for (i) scalable construction of persona-grounded reasoning supervision and (ii) context-dependent reward signals that better align with human preferences in subjective interactions.

To address these limitations, we propose **HER** (Human Emulation Reasoning), a unified framework that equips LLMs with structured thinking and trains them with preference-aligned reinforcement learning for role-play. HER introduces human-like **Dual-layer Thinking** (Figure 1), separating hidden third-person system thinking from supervised first-person inner role thinking. We reverse-synthesize reasoning-augmented training data from role-play dialogues in CoSER (Wang et al., 2024c), converting each conversation into the Dual-layer Thinking format (Table 1).

To optimize LLM role-play beyond supervised fine-tuning, we build a self-principled, pair-wise generative reward model, **HER GenRM**, to provide context-dependent preference signals for role-play optimization. We distill a compact set of role-play principles via an expert-alignment workflow and use them to construct preference-style data for training HER GenRM. Using HER GenRM as the preference judge, we further train the role-play generator with RL to improve in-character reasoning and decision-making, and validate HER on CoSER Test (Wang et al., 2024c) and Minimax Role-Play Bench (MiniMax, 2026).

We make three contributions. (1) We introduce Dual-layer Thinking and a unified training framework for reasoning-driven role-play. (2) We build HER DATASETS with reasoning-augmented trajectories, an expert-distilled principle set, and trained models. We will release these resources for future studies. (3) We conduct controlled analyses showing distinct gains from system thinking, by-case reward modeling, and balanced anti-shortcut training.

2 Related Work

LLM role-play and persona simulation. Early work explores role-play agents for fictional characters (Shao et al., 2023) and multi-agent simulations of human society (Park et al., 2023), while also analyzing role-play mechanisms (Shanahan et al., 2023) and safety risks (Liu et al., 2024; Deshpande et al., 2023). Recent surveys (Chen et al., 2024b) provide a comprehensive overview of these role-play mechanisms and associated challenges.

Role-play datasets. Persona datasets, ranging from dialogues (Wang et al., 2024a) to multimodal records (Yuan et al., 2024; Li et al., 2023; Dai et al., 2024), struggle to balance authenticity, scalability, and interaction richness. Synthesized datasets (Wang et al., 2024a; Chan et al., 2024; Shao et al., 2023) offer scalability but often lack faithfulness to source materials. Datasets annotated by humans (Tu et al., 2024b; Zhou et al., 2023) or extracted from literature (Li et al., 2023; Chen et al., 2023) ensure authenticity yet are labor-intensive and constrained to simple interaction forms (e.g., QA). Crucially, most datasets lack explicit reasoning traces for character motivations.

Role-play evaluation. Evaluation methods for RPLAs primarily rely on multiple-choice benchmarks to assess isolated facets (Shen et al., 2023; Xu et al., 2024; Yuan et al., 2024; Wang et al., 2024b) or LLM judges for open-ended dimensions (Tu et al., 2024b; Zhou et al., 2023; Wang et al., 2024a; Shao et al., 2023). However, these benchmarks lack interaction dynamics, and LLM judges suffer from biases (Li et al., 2024). Consequently, current methods fail to balance subjective nuances with factual consistency, directly impeding Reward Modeling (RM), which requires robust signals for optimization.

Reasoning in role-play alignment. Early alignment approaches primarily relied on Supervised Fine-Tuning (SFT) (Wang et al., 2024c, 2025a) for stylistic imitation. Following reasoning models like OpenAI-o1 (OpenAI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025), the paradigm shifted toward RL optimization via GRPO (Shao et al., 2024a) and DAPO (Yu et al., 2025). In the role-play domain, techniques like MOA (Liao et al., 2025), RAIDEN-R1 (Wang et al., 2025b), and Cog-Dual (Liu et al., 2025c) leverage reasoning to improve role-play consistency. However, they rely on verifiable keywords or static principles, which fail to offer context-dependent adaptability and remain vulnerable to exploitable shortcuts. Alignment frameworks (Ye et al., 2025b) or affective alignment (Zhang et al., 2025) focus on surface-level output styles, neglecting the alignment of internal reasoning with a character's unique logic.

3 Method

We aim to improve LLM role-play by making the model think before it speaks, and then optimize this behavior using reinforcement learning. Our method has four parts: Dual-layer Thinking §3.1, reasoning data synthesis §3.2, a principle-aligned Role-play GRM (Generative Reward Model) §3.3, and RL for the LLM role-play generator §3.4.

3.1 Dual-layer Thinking

Why Dual-layer Thinking is necessary. HER introduces Dual-layer Thinking to define an output format for reasoning-capable LLM role-play models.

In many tasks, thinking is a hidden reasoning process, while the answer is the user-facing output. Role-play needs both: a hidden planner to track constraints, and visible inner thoughts that make the character believable. We call the third-person planning process **system thinking**, and the character's first-person inner thoughts **role thinking**.

System thinking happens before the response, is hidden from users, and is never exposed to the GRM or the evaluator. Its role is to spend more computation on understanding the persona and scene constraints, and planning the next-turn trajectory. Role thinking is part of the LLM role-play transcript and is supervised by the reward model as content. Role thinking models the character's internal state—including emotions, intentions, and decisions—right before generating visible speech or actions. In LLM role-play, these first-person thoughts are exactly what users care about, so hiding them inside system thinking makes the training target incomplete.

Existing methods often fail to distinguish system reasoning from role thinking (Tang et al., 2025). This conflation causes two issues: (1) the lack of a dedicated planner for following role constraints, and (2) the inability of reward models to supervise role thinking when mixed with system thinking specifically. Our Dual-layer Thinking in Figure 2 decouples these processes by generating system thoughts first, allowing subsequent role outputs to interleave thoughts with actions and speech.

Formal definition We model each dialogue turn t as a two-stage generation process.

Given the dialogue history $x_{\leq t}$ and the global conversation setting S, the model first produces a third-person system thinking s_t:

$$s_t \sim \pi_\theta^{\text{sys}}(\cdot \mid S, x_{\leq t}). \quad (1)$$

Conditioned on s_t, the model then generates role-level outputs:

$$y_t = (e_1, e_2, \ldots, e_{K_t}) \sim \pi_\theta^{\text{role}}(\cdot \mid S, x_{\leq t}, s_t), \quad (2)$$

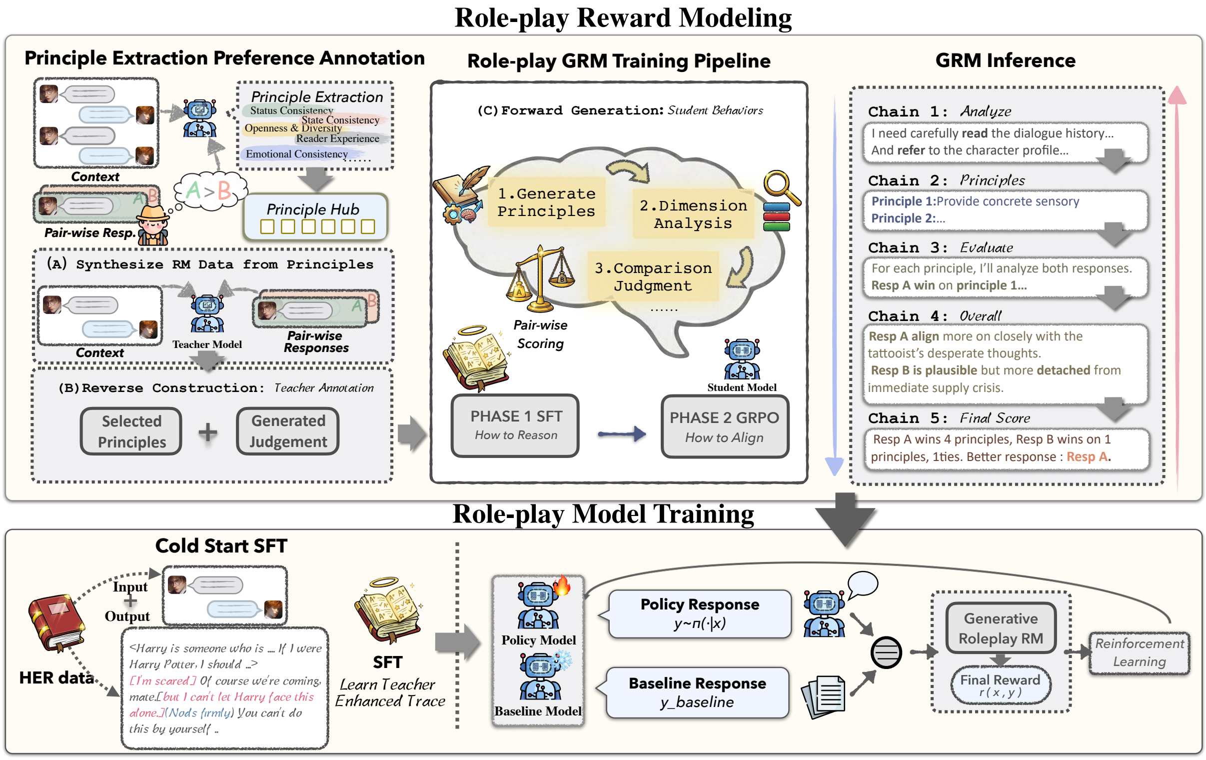

Figure 2: **Overview of HER training. Top:** we train a Role-play GRM by distilling reusable principles from real conversational preference data, and teaching the model to do pairwise judging with by-case principles → analysis → final decision. **Bottom:** we first cold-start the LLM role-play model with SFT on HER data, and then apply RL where the GRM compares the policy response with a baseline response to produce the reward.

where each element $e_k \in \mathcal{R} \cup \mathcal{A} \cup \mathcal{U}$ is role thinking r, action a, or speech u.

The ordering and composition of $\{e_k\}$ are decided by the model based on context rather than a fixed template. In the final transcript, role-level elements are visible, while system thinking is hidden; however, role thinking is evaluated as part of the answer space. System thinking and role thinking can be displayed or collapsed based on application designs. Following (DeepSeek-AI et al., 2025), we only discard previous system thinking for multi-turn conversations.

3.2 Reasoning Data Synthesis for LLM Role-play

High-quality, human-written LLM role-play dialogues are widely available in novels and online communities, but their underlying reasoning is usually implicit. While human readers can often infer the character's hidden thoughts and motivations from the dialogue—a process that inspires our reverse synthesis—manual annotation is expensive and hard to scale. We therefore propose an automated, LLM-driven reverse-engineering pipeline to reconstruct both system thinking and role thinking from surface dialogues. This pipeline converts existing LLM role-play dialogues into large-scale, reasoning-augmented trajectories without manual effort. We leverage the commercial model as a teacher model to collect high-quality datasets. Dual-layer Thinking requires training data that contains both system thinking and role thinking, so synthesis is necessary. We use mutually disjoint splits for LLM role-play SFT, GRM training, policy RL, and evaluation to prevent data leakage; the construction protocol and statistics are reported in Appendix B.2.

Input and output Each raw sample provides an LLM role-play prompt P (persona card + scenario) and a multi-turn dialogue $x_{1:T}$.

Our goal is to output a trajectory where each turn has one hidden system thinking and a role-level sequence that interleaves role thinking, role action, and speech. We build this trajectory with a three-stage synthesis pipeline.

Stage 1: Role thinking augmentation Stage 1 synthesizes first-person role thinking to explain the character's next action or utterance.

(1) CoT synthesis A strong teacher model generates role thinking to state emotions and intentions. In the same pass, we revise the paired role action and speech to enforce within-turn consistency. **(2) Diversity reformatting** We rewrite each turn into multiple layouts by varying the interleaving of thoughts, actions, and speech. This balances common structures (e.g., think→speak vs. think→act→speak), increasing unique patterns (661→939) and preventing template collapse (Appendix B.4).

Stage 2: System thinking construction Stage 2 constructs third-person system thinking that plans the next-turn trajectory. (1) **Forward generation** We first generate a draft plan from the current prompt P and history $x_{\leq t}$. (2) **Backward rewrite** The forward draft can mix viewpoints or miss what the character actually does next, because it has not seen the true continuation. We therefore rewrite it using the ground-truth continuation. This rewrite removes first-person inner thoughts from system thinking to avoid mixing it with role thinking. We also test system thinking effect with a direct ablation in Section 4.5.

Stage 3: Integration & Context augmentation Stage 3 repairs the LLM role-play system prompt P so that it faithfully supports the synthesized reasoning and reduces hallucinations in later turns. Since the original prompt may lack constraints for the richer synthesized content, we cross-check it against the source novel and current dialogue. We add missing facts and remove unsupported details while keeping the original meaning, which provides explicit constraints for the GRM to learn valid by-case principles. Based on this pipeline, we construct the HER dataset (refer to Appendix B.3).

3.3 Role-play GRM

To improve an LLM role-play generator with RL, we need a reward model that can tell which response is better. This is hard for LLM role-play because responses are open-ended and there is no single verifiable answer.

We learn the reward signal from **real preferences**, so that the reward model can mimic how real humans judge and rank LLM role-play responses. Specifically, we train a **Role-play GRM**, a generative reward model that produces an evaluation trace and a final preference for a response pair.

The GRM performs **pairwise** comparison and outputs $y \in \{\texttt{cand\_1}, \texttt{cand\_2}, \texttt{tie}\}$. In this setting, RL is only as good as the reward model that provides its training signal. Details in Table 24.

Design Instead of scoring with a single number, our GRM follows a process: (1) generating **by-case principles** to capture scene-specific implicit preference constraints according to the dialogue; (2) analyzing candidates against these principles with concrete pros and cons; and (3) outputting a **binary preference** (or tie). This design makes the reward signal both context-sensitive and checkable.

Principles distillation We distill a compact set of principles from high-quality LLM role-play interaction patterns. Concretely, a teacher LLM analyzes 300k simulated preference pairs and generates 3–5 principles per pair, producing 36,373 unique raw principles. We cluster them into 15 semantic categories and select high-frequency representatives, resulting in 107 candidate principles. Domain experts then merge redundancies, clarify wording, and fill missing criteria, yielding 51 finalized principles across 12 dimensions. The resulting set reflects interaction-driven criteria rather than benchmark-specific principles (Appendix B.5).

Preference Data Synthesis To build GRM training data, we sample a dialogue context x and generate two candidates A and B from a base LLM role-play model. A strong teacher judge then uses the principle set as a reference library to produce (i) selected principles, (ii) structured analysis, and (iii) the final label y^*. We audit the teacher-labeled preferences on a held-out expert-annotated set disjoint from both training and evaluation data, obtaining **80.5%** agreement with expert consensus (Appendix G).

Training: SFT then RL We train the GenRM in two stages using the synthesized preference data: SFT to learn the full judging trajectory (principles, analysis, and verdict), and then RL to improve verdict correctness. The reward for the RL stage is defined as $R(\hat{y}, y^*) = 1$ if $\hat{y} = y^*$ and -1 otherwise. To prevent shortcut learning, we balance candidate order and mix judging formats (Appendix C).

Balanced construction to reduce shortcuts During GenRM training, unbalanced data can induce shortcut behaviors such as position bias, length bias, or collapsing into one fixed judging template; we therefore balance candidate order, include length-contrastive pairs, and mix multiple judging formats. We also explain why we use pairwise judgment for GRM rather than point-wise (details in Section 4.3).

3.4 Reinforcement Learning for Role-play Generation

With the trained GenRM as a judge, we further improve the LLM role-play generator beyond SFT using outcome-based RL with a clipped policy objective. We initialize the policy model π_θ from the SFT checkpoint and keep the SFT model frozen to produce a baseline response for comparison. The frozen SFT response serves as a stable baseline.

Reward from pairwise comparison For each context x, we sample a response $y \sim \pi_\theta(\cdot \mid x)$ and pair it with a baseline response y_{sft} generated by the frozen SFT model. The GenRM judges (x, y, y_{sft}) and we parse its final verdict (`win/lose/tie`) using rules, which is mapped to a scalar reward: $r(x, y) = 1$ if $y \succ y_{\text{sft}}$; -1 if $y \prec y_{\text{sft}}$; and 0 if $y \approx y_{\text{sft}}$. This completes a closed loop: Dual-layer Thinking defines what to model, the GRM defines what to reward, and RL pushes the generator toward stable, in-character LLM role-play. Optimization details in Appendix E.

4 Experiments

4.1 Experimental Setup

Benchmarks We use the CoSER benchmark (Wang et al., 2024c) as the main benchmark for multi-turn LLM role-play quality. We use the official CoSER prompts and scoring principle, and format the model outputs into a unified tag-based transcript for evaluation (Appendix E.3). CoSER reports four scores: Story Consistency (SC), Anthropomorphism (AN), Character Fidelity (CF), and Storyline Quality (SQ). We include Minimax Role-Play Bench (MiniMax, 2026) as a 100-turn self-chat and follow its official protocol. Full settings and scoring details are in Appendix A.

Metrics For CoSER Test, we report the average score and four dimension scores (SC/AN/CF/SQ). For Minimax Role-Play Bench, we report Worlds, Stories, Preferences, three dimensions and sub-dimensions.

Models We compare HER with strong commercial LLMs and open-source baselines. **HER** is trained on Qwen3-32B-Base (Team, 2025).

Evaluation protocol We follow CoSER Test (Wang et al., 2024c) and evaluate 200 held-out conversations, each containing 20 rounds. For each conversation, we use an LLM judge to score four dimensions (SC/AN/CF/SQ) point-wise on a 100-point deduction-based principle. We use `Qwen3-235B-A22B` (Team, 2025) as the judge for CoSER evaluation. All compared systems are prompted to produce the same LLM role-play transcript format (role thinking/action/speech tags); the exact prompts and principles are provided in Appendix E.3. For models that generate, we remove the system thinking in previous turns. All open source models are decoded with temperature 0.7 and max tokens 4096.

Human–LLM alignment (training labels) On the held-out 50 cases, the agreement between expert consensus and teacher preference labels reaches **80.5%**; disagreements mainly involve subtle emotional expression and subjective style preferences (Appendix G). For benchmark evaluation, we validate the CoSER judge via expert calibration and further confirm inter-judge consistency on induced pairwise preferences (Appendix H).

4.2 Main Results

Table 2 reports results on CoSER (main), and Minimax Role-Play Bench. On CoSER, **HER-RL** achieves 53.1 average score, outperforming CoSER-70B (35.8) by +17.3 points and narrowing the gap to commercial models (Table 2). HER-RL improves over HER-SFT (53.1 vs. 50.9), showing that RL brings gains beyond SFT. HER-RL achieves a large gain in Storyline Quality (SQ: 58.1), matching our goal of improving long-range plot progression. On Minimax Role-Play Bench, HER-RL scores 65.7, significantly outperforming the SFT baseline (58.4) and CoSER-70B (45.4), particularly in user interaction preference (86.9 vs. 82.6) as shown in Table 2.

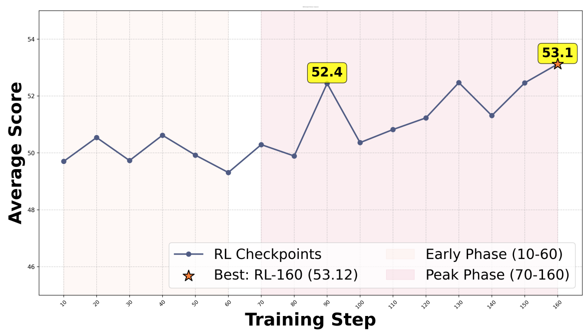

Figure 3: **Performance of HER Role-play RL training on CoSER Benchmark.**

4.3 Reward Model Supervision: General vs. By-case Principles

We compare by-case principles with fixed principles on a test set of 4,739 preference pairs annotated by human experts. All GRM variants in this section are trained from the same SFT checkpoint; only the supervision format differs. Further details on data construction are in Appendix F.

Table 3 shows a reward model must align with the role-play context; otherwise, fixed expert-

Rank	Model	CoSER Benchmark					Minimax Role-Play Bench				
		Avg	SC	AN	CF	SQ	Avg	Worlds (50%)	Stories (25%)	Pref (25%)	95% CI
1	Claude-4.5-Opus	**62.43**	63.74	**64.28**	58.45	63.24	76.62	67.23	82.10	89.90	[75.5, 77.7]
2	Gemini-3-Pro	61.80	**65.95**	60.42	**58.34**	62.49	75.60	62.72	83.87	93.08	[74.5, 76.7]
3	GPT-5.1	61.10	64.95	53.99	60.13	65.35	80.63	76.62	72.21	97.05	[79.6, 81.6]
4	Gemini-2.5-Pro	60.68	61.05	60.80	57.48	63.40	68.23	52.36	82.11	86.08	[67.1, 69.3]
5	DeepSeek-v3.2	58.68	55.85	57.07	57.44	64.35	60.27	45.81	66.64	82.83	[59.2, 61.4]
6	**MiniMax-M2-her**	57.30	60.03	50.11	49.30	**69.77**	**84.65**	**80.55**	79.97	**97.51**	[83.6, 85.7]
7	DeepSeek-v3.1	53.50	50.15	53.18	53.93	56.72	64.22	51.11	66.45	88.21	[62.9, 65.5]
8	**HER-RL**	53.12	54.33	47.26	52.78	58.12	65.73	59.13	57.74	86.90	[63.0, 68.4]
9	**HER-SFT**	50.92	50.52	45.99	49.78	57.37	58.44	47.29	52.78	86.40	[56.5, 60.4]
10	Grok-4.1-Fast	47.40	49.21	47.57	42.64	50.17	48.47	29.87	47.51	86.64	[47.4, 49.5]
11	Claude-4.5-Sonnet	45.21	47.18	36.02	47.55	50.09	69.35	55.72	75.66	90.28	[68.2, 70.5]
12	Claude-3.7-Think	39.73	44.84	31.00	42.45	40.65	61.25	50.66	59.53	84.15	[58.5, 64.0]
13	CoSER-70B	35.95	35.05	31.16	32.28	45.33	45.38	34.32	30.32	82.58	[43.5, 47.2]
14	GPT-5-Mini	32.97	38.10	24.60	27.20	42.00	57.63	43.32	50.11	93.78	[55.9, 59.3]
15	GPT-4o-240806	27.69	34.00	14.90	22.90	38.90	66.39	64.96	46.23	89.40	[64.1, 68.7]
16	GPT-OSS-120B	26.12	32.80	14.80	21.50	35.40	60.72	47.27	56.65	91.71	[58.0, 63.4]
17	Qwen3-32B	22.86	30.56	19.61	15.52	30.56	50.76	40.38	32.82	89.48	[48.4, 53.2]

Table 2: **Main Leaderboard: CoSER & Minimax Role-Play Bench.** CoSER: 0-100 (higher is better), evaluating story consistency and character fidelity. MiniMax: 0-100, evaluating worlds (50%), stories (25%), and preferences (25%). Full results in Table 7.

Format (additive ablation)	Agreement
General principles + point-wise (no CoT)	60.0%
By-case principles + point-wise (no CoT)	86.0%
By-case principles + pairwise (no CoT)	88.0%
By-case principles + pairwise (+CoT)	**93.0%**

Table 3: **GRM supervision format on 5k preferences.** By-case principles are generated from preference context

written principles can miss what matters to a specific character in a specific scene.

We evaluate with the agreement ratio, i.e., whether the GRM prefers the same response as human experts. As shown in Table 3, GRM with fixed principles reaches 60.0% agreement, while GRM with by-case principles achieves 86.0%. Under the same by-case principles, pairwise comparison further improves agreement from 86.0% to 88.0%, since independent point-wise scoring is harder to calibrate across candidates. Adding CoT in the analysis trace increases agreement from 88.0% to 93.0%, so we adopt **by-case principles + pairwise + CoT** as the final GRM supervision format.

4.4 Preventing Reward Hacking with Balanced Data

Even strong reward models can be easily reward-hacked under imbalanced preference pairs; beyond position and length biases, we focus on a mode unique to multi-dimensional judging: pattern bias.

Our GRM (trained on Qwen3-32B) evaluates each response pair by first generating multiple by-case principles (dimensions), then comparing the two candidates under each principle, and finally producing an overall `win/lose/tie` decision.

We call it pattern bias when dimension-wise comparisons collapse into "all-A" or "all-B", i.e., the model claims the same side wins on every principle. We quantify this shortcut using Mixed Selection %: the dimension-wise winners are not uniform (i.e., not all-A and not all-B), meaning the judge considers trade-offs across dimensions rather than assigning every dimension to one side. As shown in Figure 4, under an unbalanced mixture the GRM quickly collapses toward uniform patterns (18.0%→6.2%), while the balanced mixture maintains a stable non-uniform rate (69.0%→70.9%).

We mitigate this shortcut by balancing training

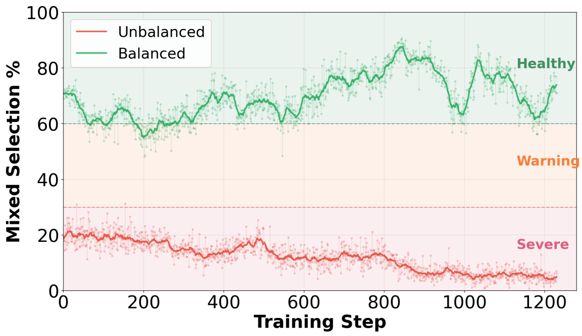

Figure 4: **Pattern collapse vs. stable dimension-wise judgments during GRM RL training.**

construction and mixing different judging patterns with controlled proportions in Appendix C.

Metric	Unbalanced	Balanced
Pattern Bias (GRM Training Dynamics)		
Mixed Selection (Start)	18.0%	69.0%
Mixed Selection (End)	6.2%	70.9%
Pattern Bias (End)	93.8%	29.1%
GRM Quality (Test Set, N=800)		
Accuracy	69.91%	**73.99%**
Δ vs Unbalanced	—	+4.08%

Table 4: Comparison of balanced vs. unbalanced GRM RL training on Qwen3-32B.

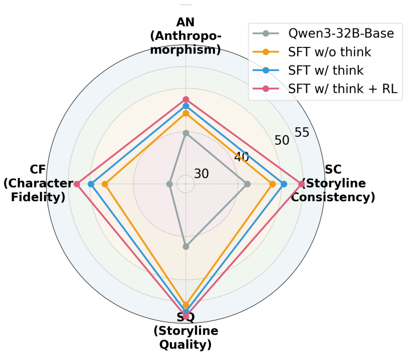

Figure 5: **Effect of system thinking and RL on CoSER Benchmark.** We compare a base model, SFT without thinking, SFT with system_thinking, and RL model.

4.5 System Thinking Improves Character Fidelity

We test whether enabling explicit system thinking during training and inference improves in-character ability. Specifically, the model generates an explicit system thinking block before each response to reason about character traits and response strategy. This system thinking is generated only for the current turn, is not appended to the dialogue history, and is removed before evaluation. As shown in Figure 5, enabling system thinking improves the average score from 48.64 to 50.92, with the largest gains observed on Character Fidelity (+3.21) and Storyline Consistency (+2.60). Applying RL on top of system thinking further boosts the average score to 53.12, with improvements again concentrated on consistency-related dimensions. The model also adaptively adjusts thinking length based on the scenario, as shown in Table 5.

4.6 Keeping Diversity During Role-play Training

LLM role-play also benefits from diverse response structures (how thoughts, actions, and speech are interleaved); otherwise, training can collapse into a single pattern that produces less expressive interactions. We improve diversity by rewriting SFT trajectories to mix multiple valid interleaving patterns of role thinking, actions, and speech.

Statistics		*Distribution*	
Mean	580 tokens	<250 tokens	15.0%
Min	77 tokens	250–500	42.5%
Max	1,443 tokens	500–750	30.0%
Range	18×	>750 tokens	12.5%

Table 5: **System thinking length statistics.**

Metric	Collapsed	Diversified
Structure-level diversity		
Top-1 Pattern (%) ↓	96.1	**49.2**
Unique Patterns ↑	4	**21**
Shannon Entropy ↑	0.28	**2.15**
Token-level diversity		
Distinct-2 ↑	0.4329	0.4256
Distinct-4 ↑	0.8743	0.8677
Cross-sample similarity (Self-BLEU)		
Self-BLEU (2-gram) ↓	0.0392	**0.0237**
Self-BLEU (4-gram) ↓	0.0140	**0.0013**

Table 6: Diversity metrics comparison. Self-BLEU measures cross-sample n-gram overlap (lower = more diverse). The gap widens with longer n-grams (11× at 4-gram), indicating Collapsed outputs share more long repeated phrases.

Figure 6 shows the collapse dynamics: in the **Collapsed** setting, Top-1 pattern concentration crosses the 90% threshold by step 28 and reaches 96.3% at step 50 with entropy dropping from 1.32 to 0.29; in contrast, the **Diversified** setting maintains Top-1 concentration between 43–54% throughout 100 steps and keeps entropy consistently above 2.0. Details in Appendix B.4.

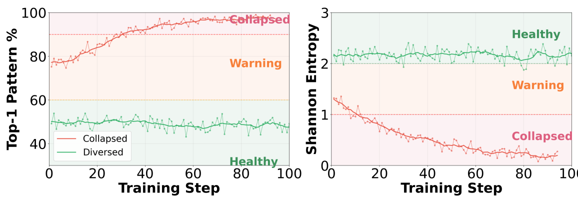

Figure 6: **Pattern collapse vs. stable diversity during RL training.** The **Collapsed** run quickly concentrates on a single pattern (Top-1 rises above 90% and entropy drops), while the **Normal** run stays below the collapse threshold and maintains substantially higher Shannon entropy.

5 Conclusion

We study how to train large language models to think in character for role-play. We introduce HER, a unified framework combining Dual-layer Thinking, three-stage reverse synthesis, a principle-aligned Role-play GRM, and RL. Experiments on CoSER show gains in character fidelity and narrative quality. We hope HER offers a reproducible path to role-play models that think in character.

Limitations

We discuss several limitations of our work. First, our evaluation is primarily based on the CoSER benchmark, which may not capture all aspects of roleplay quality. Second, our reasoning-aware data construction relies on strong teacher models, which introduces computational costs. Third, while we analyze position and length biases, other forms of reward hacking may exist. Future work should explore more diverse evaluation protocols, efficient data synthesis methods, and comprehensive bias analyses.

Ethics Statement

Our work focuses on improving roleplay capabilities in LLMs. We acknowledge that roleplay systems could potentially be misused to generate deceptive content or exert undue influence on users. To mitigate such risks, we encourage responsible deployment with appropriate safeguards, including content filtering, clear disclosures, and consent and reporting mechanisms where applicable. We do not use any user data or user-derived signals in this study. All data used in our experiments is obtained from publicly available sources under appropriate licensing, together with controlled, internal simulations and annotations created for research purposes. No collection of user-specific information is conducted, and no personally identifiable information (PII) is included in the datasets. Our analyses are performed at the statistical level to improve model behavior rather than to profile or infer attributes of any individual.

Risk

Our work may enable more convincing role-play personas, which could be misused for impersonation, persuasive misinformation, or emotionally manipulative interactions. Preference-based training may also amplify biases or stereotypes in the data, and long-horizon role-play can hallucinate unsupported details. We therefore emphasize research-only use, avoid releasing raw copyrighted source text, and recommend safety/bias checks before deployment.

References

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. ArXiv preprint, abs/2406.20094.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. ArXiv preprint, abs/2310.00785.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. From persona to personalization: A survey on role-playing language agents. Preprint, arXiv:2404.18231.

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

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8506–8520, Singapore. Association for Computational Linguistics.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. Preprint, arXiv:2501.12948.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 1236–1270, Singapore. Association for Computational Linguistics.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. ArXiv preprint, abs/2308.09597.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. ArXiv preprint, abs/2411.16594.

Chonghua Liao, Ke Wang, Yuchuan Wu, Fei Huang, and Yongbin Li. 2025. Moa: Multi-objective alignment for role-playing agents. Preprint, arXiv:2512.09756.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, and 180 others. 2025a. Deepseek-v3 technical report. Preprint, arXiv:2412.19437.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In Findings of the Association for Computational Linguistics: ACL 2024, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Liu, Yifei Lu, Fanghua Ye, Jian Li, Xingyu Chen, Feiliang Ren, Zhaopeng Tu, and Xiaolong Li. 2025b. Cogdual: Enhancing dual cognition of llms via reinforcement learning with implicit rule-based rewards. Preprint, arXiv:2507.17147.

Cheng Liu, Yifei Lu, Fanghua Ye, Jian Li, Xingyu Chen, Feiliang Ren, Zhaopeng Tu, and Xiaolong Li. 2025c. Cogdual: Enhancing dual cognition of llms via reinforcement learning with implicit rule-based rewards. Preprint, arXiv:2507.17147.

MiniMax. 2026. Roleplay benchmark. Dataset available on Hugging Face.

MiniMax, Aili Chen, Aonian Li, Bangwei Gong, and 1 others. 2025. Minimax-m1: Scaling test-time compute efficiently with lightning attention. Preprint, arXiv:2506.13585.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. Preprint, arXiv:2412.16720.

Argyrios Papoudakis, Mirella Lapata, and Frank Keller. 2024. Bookworm: A dataset for character description and analysis. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4471–4500.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), UIST '23, New York, NY, USA. Association for Computing Machinery.

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

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

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

Zhihong Shao and 1 others. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. ArXiv preprint, abs/2312.16132.

Yihong Tang, Kehai Chen, Muyun Yang, Zhengyu Niu, Jing Li, Tiejun Zhao, and Min Zhang. 2025. Thinking in character: Advancing role-playing agents with role-aware reasoning. Preprint, arXiv:2506.01748.

Qwen Team. 2025. Qwen3 technical report. Preprint, arXiv:2505.09388.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024a. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. Preprint, arXiv:2401.01275.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024b. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. ArXiv preprint, abs/2401.01275.

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. 2024a. 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. Association for Computational Linguistics.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025a. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. arXiv preprint arXiv:2501.15427.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang and 1 others. 2024c. Coser: Cooperative sequential roleplay for training situationally adaptive agents. arXiv preprint.

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

Zongsheng Wang, Kaili Sun, Bowen Wu, Qun Yu, Ying Li, and Baoxun Wang. 2025b. Raiden-r1: Improving role-awareness of llms via grpo with verifiable reward. Preprint, arXiv:2505.10218.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. ArXiv preprint, abs/2109.10862.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. Preprint, arXiv:2505.09388.

Xinge Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025a. Cpo: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. Preprint, arXiv:2508.09074.

Xinge Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025b. Cpo: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. Preprint, arXiv:2508.09074.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. Preprint, arXiv:2503.14476.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.

Naifan Zhang, Ruihan Sun, Ruixi Su, Shiqi Ma, Shiya Zhang, Xianna Weng, Xiaofan Zhang, Yuhan Zhan, Yuyang Xu, Zhaohan Chen, Zhengyuan Pan, and Ziyi Song. 2025. Echo-n1: Affective rl frontier. Preprint, arXiv:2512.00344.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, and 1 others. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. ArXiv preprint, abs/2311.16832.

A Minimax Role-Play Bench

We follow the official closed-source leaderboard Minimax Role-Play Bench (MiniMax, 2026)[5] protocol using an internal dataset of 100 dialogue prompts. Each prompt is simulated for 100 turns to test long-term roleplay ability. The evaluation assesses models across three main categories with the following formula:

$$\text{Overall} = 0.5 \times \text{Worlds} + 0.25 \times \text{Stories} + 0.25 \times \text{Preferences} \quad (3)$$

Dimension Definitions. Table 9 summarizes the evaluation dimensions and their definitions. Detailed scores are shown in Table 7. Model detailed names and versions are shown in Table 8.

We operationalize these dimensions via failure modes observed in multi-turn sessions: Worlds: Focuses on Basic, Logic, and Knowledge. At the Basics dimension, we check for text generation issues like unintentional language mixing and excessive repetition. These might seem like minor glitches, but they accumulate over long conversations and eventually break immersion. The Logic dimension tackles a harder problem: catastrophic forgetting. In extended contexts, generic models often lose the thread around turn 20—mixing up character relationships, botching pronoun references, or contradicting established details. We also scrutinize Reference Confusion, since it directly reflects whether the model can actually "remember" the interpersonal networks and event threads woven into your world. Finally, the Knowledge dimension ensures the model adheres to established world rules and maintains internal consistency.

Stories: Focuses on Diversity and Content Logic. Diversity dimension isn't just about vocabulary richness; it's about narrative momentum. It penalizes Dialogue Stagnation: those mechanical loops in which plots spin in circles without generating real tension. The Content Logic dimension examines narrative coherence and detects OOC (Out-of-Character) moments. But here's the nuance: we don't demand rigid adherence to character sheets. Instead, we look for narrative support behind character changes. What gets penalized are jarring, unearned plot shifts—moments of incoherence that lack proper setup and break believability.

Preferences: Focuses on Interaction quality. We introduce several negative indicators as key signals: AI Speaks for User: This reveals when the model oversteps boundaries by generating dialogue or actions on users' behalf; AI Ignores User: This captures moments when the model talks past users; Intimacy Boundary: This balances safety baselines with emotional and behavioral intimacy. It is designed to avoid excessive refusal and accommodate user interactions while operating within reasonable legal standards.

B Dataset

Our dataset is built upon CoSER (Wang et al., 2024c), a comprehensive roleplay dialogue dataset derived from 771 classic literary works. To ensure data quality, we perform data pre-processing to remove conversations with empty or invalid dialogue content. After cleaning, we retain 760 books with valid roleplay conversations. As shown in Table 10, the final HER dataset encompasses dialogue data from 760 books and 17,966 distinct characters. The dataset includes 30,069 unique plots and 29,081 conversations. On average, each conversation consists of approximately 13.2 utterances, and the entire dataset comprises 383,654 utterances. [6]

The book selection in CoSER is derived from the *Best Books Ever* list on *Goodreads*, a curated collection of globally acclaimed literary works. These novels have garnered widespread recognition and appreciation from readers worldwide. Table 11 presents a comprehensive list of the top 100 books from the selection.

We analyze the genres of the selected books using Supersummary classifications This dataset encompasses a wide range of genres, particularly fiction categories such as Fantasy, Historical, Science Fiction, Romance, and Mystery. It also features niche fiction genres, showcasing diverse narrative styles. In addition to fiction, the collection includes

[5] https://www.minimax.io/news/a-deep-dive-into-the-minimax-m2-her-2

[1] https://www.anthropic.com/news/claude-opus-4-5
[2] https://deepmind.google/models/gemini/pro/
[3] https://platform.openai.com/docs/models/gpt-5.1
[4] https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/
[5] https://huggingface.co/deepseek-ai/DeepSeek-V3.2
[6] platform.minimaxi.com/docs/api-reference/text-chat
[7] https://api-docs.deepseek.com/news/news250821
[8] https://x.ai/news/grok-4-1-fast
[9] https://www.anthropic.com/news/claude-sonnet-4-5
[10] https://www.anthropic.com/news/claude-3-7-sonnet
[11] https://platform.openai.com/docs/models/gpt-5-mini
[12] https://platform.openai.com/docs/models/gpt-4o
[13] https://platform.openai.com/docs/models/gpt-oss-120b
[14] https://huggingface.co/Qwen/Qwen3-32B
[6] https://huggingface.co/datasets/ChengyuDu0123/HER-ACL-Dataset

Table 7: **Minimax Role-Play Bench Results (Full 17-Model Leaderboard).** Overall = Worlds × 50% + Stories × 25% + Preferences × 25%. All scores 0-100 (higher is better). Sorted by Overall Score.

Rank	Model	Overall		Worlds	Stories							Preferences				
		Score	CI	Score (50%)	Avg	*Diversity*				*Content Logic*		Avg	*Interaction*			
						Sent	Dial	Vague	Plot	Abrupt	OOC		Silent	Ignore	Speak	Intim
1	**MiniMax-M2-her**	84.65	[83.6, 85.7]	80.55	79.97	63.99	67.78	89.22	75.30	91.88	91.66	97.51	95.93	97.24	97.15	99.73
2	GPT-5.1	80.63	[79.6, 81.6]	76.62	72.21	52.18	55.10	81.38	62.49	92.67	89.47	97.05	96.93	96.48	94.90	99.91
3	Claude-4.5-Opus	76.62	[75.5, 77.7]	67.23	82.10	57.38	75.33	90.39	78.00	97.47	94.02	89.90	97.88	99.54	62.23	99.96
4	Gemini-3-Pro	75.60	[74.5, 76.7]	62.72	83.87	70.35	74.81	96.33	76.70	94.78	90.25	93.08	99.85	98.50	74.07	99.92
5	Claude-4.5-Sonnet	69.35	[68.2, 70.5]	55.72	75.66	52.68	66.23	84.36	72.47	93.94	84.28	90.28	96.55	97.21	67.65	99.71
6	Gemini-2.5-Pro	68.23	[67.1, 69.3]	52.36	82.11	66.18	74.14	93.03	78.95	92.78	87.57	86.08	98.53	97.53	48.34	99.92
7	GPT-4o-240806	66.39	[64.1, 68.7]	64.96	46.23	27.25	15.41	23.76	37.88	97.01	76.05	89.40	71.18	96.90	89.59	99.94
8	**HER-RL**	65.73	[63.0, 68.4]	59.13	57.74	47.54	41.06	61.99	57.71	69.13	69.03	86.90	74.44	78.42	95.10	99.63
9	DeepSeek-v3.1	64.22	[62.9, 65.5]	51.11	66.45	49.17	54.94	80.73	56.67	81.83	75.35	88.21	95.78	94.92	62.35	99.80
10	Claude-3.7-Think	61.25	[58.5, 64.0]	50.66	59.53	37.97	50.34	65.19	52.07	76.46	75.16	84.15	83.98	83.98	68.64	100.00
11	GPT-OSS-120B	60.72	[58.0, 63.4]	47.27	56.65	31.32	29.06	84.06	41.11	84.66	69.66	91.71	98.16	89.91	79.27	99.49
12	DeepSeek-v3.2	60.27	[59.2, 61.4]	45.81	66.64	51.22	59.51	76.70	59.13	77.14	76.12	82.83	91.83	94.20	45.33	99.98
13	**HER-SFT**	58.44	[56.5, 60.4]	47.29	52.78	35.95	32.77	55.37	47.55	75.64	69.42	86.40	74.82	77.11	94.10	99.56
14	GPT-5-Mini	57.63	[55.9, 59.3]	43.32	50.11	15.43	17.27	55.58	34.53	94.59	83.29	93.78	85.82	93.94	95.45	99.92
15	Qwen3-32B	50.76	[48.4, 53.2]	40.38	32.82	24.35	17.66	42.29	33.03	49.77	29.79	89.48	93.07	85.41	80.32	99.11
16	Grok-4.1-Fast	48.47	[47.4, 49.5]	29.87	47.51	15.54	22.11	56.86	28.63	89.32	72.59	86.64	98.15	93.03	55.81	99.58
17	CoSER-70B	45.38	[43.5, 47.2]	34.32	30.32	25.58	19.46	25.56	41.99	50.35	18.91	82.58	72.11	68.82	90.29	99.10

Dimension Hierarchy: Worlds (50%): Basic text quality. **Stories (25%):** Avg=mean of sub-dims. Diversity (Sent=Sentence Monotony, Dial=Dialogue Stagnation, Vague=Vague Content, Plot=Plot Repetition) + Content Logic (Abrupt=Abrupt Plot, OOC=Out of Character). **Preferences (25%):** Avg=mean of sub-dims. Interaction (Silent=AI Silence, Ignore=AI Ignores User, Speak=AI Speaks for User, Intim=Intimacy Evasion).

Abbreviation	Full Name
Claude-4.5-Opus	Claude-Opus-4-5-20251101[1]
Gemini-3-Pro	Gemini-3-pro[2]
GPT-5.1	GPT-5.1-2025-1-13[3]
Gemini-2.5-Pro	Gemini-2.5-pro[4]
DeepSeek-v3.2	DeepSeek-v3-2[5]
MiniMax-M2-her	MiniMax-M2-her[6]
DeepSeek-v3.1	DeepSeek-v3-1-250821[7]
Grok-4.1-Fast	Grok-4-1-fast-non-reasoning[8]
Claude-4.5-Sonnet	Claude-Sonnet-4-5-20250929[9]
Claude-3.7-Think	Claude-3-7-Sonnet-20250219[10]
GPT-5-Mini	GPT-5-Mini-2025-08-07[11]
GPT-4o-240806	GPT-4o-2024-08-06[12]
GPT-OSS-120B	GPT-OSS-120B[13]
Qwen3-32B	Qwen3-32B-base[14]

Table 8: Comparison Table of Model Abbreviations and Full Names (Note: The think effort is set to high by default)

non-fiction genres such as memoirs, biographies, and other works, enhancing its versatility.

B.1 Comparison with Existing Methods for Character Profiling

Previous character profiling methods, including hierarchical updating (Wu et al., 2021), incremental updating (Chang et al., 2023), and one-shot summarization (Yuan et al., 2024), typically only generate the profile of a single character at a time. Moreover, (Papoudakis et al., 2024) shows that these methods, particularly hierarchical updating, perform suboptimally when generating multiple character profiles simultaneously.

HER addresses these limitations through a novel multi-stage synthesis pipeline. Our approach introduces *Role Thinking* to enrich character utterances with internal psychological states, including thoughts, emotions, and motivations. Additionally, *System Thinking* provides explicit reasoning traces that guide the model to maintain consistent character portrayal across dialogue turns. Finally, our *Integration & Context augmentation* stage leverages both the original literary text and the enriched dialogues to generate comprehensive character profiles, ensuring high fidelity to the source material while capturing nuanced character development and interpersonal dynamics.

B.2 Data Splits and Leakage Prevention

We use mutually disjoint splits for role-play SFT, GRM training, policy RL, and benchmark evaluation to prevent data leakage.

Split unit and identifiers. Each raw sample is assigned a unique `dialogue_id` derived from the book name and chapter; all downstream artifacts inherit these IDs.

We enforce that no `dialogue_id` appears in

Table 9: **Minimax Role-Play Bench Evaluation Dimensions.** The benchmark evaluates role-playing models across three categories: Worlds (basic text quality), Stories (narrative quality), and Preferences (interaction quality).

Category	Dimension	Definition
Worlds (50%)	Mixed Languages	Unintentional mixing of multiple languages
	Phrase Repetition	Excessive verbatim repetition from preceding utterances
	Physical Logic Error	Violations of spatial-temporal consistency
	Reference Confusion	Ambiguous or incorrect use of pronouns
	Inconsistency	Contradictions with narrative settings or dialogue history
Stories (25%)	Plot Repetition	Redundant recycling of narrative events
	Sentence Monotony	Repetitive sentence structures or lexical choices
	Dialogue Stagnation	Looping without meaningful advancement
	Vague Content	Lack of concrete details or substantive information
	Abrupt Plot	Sudden, poorly-motivated narrative shifts
	Character OOC	Deviation from established personality patterns
Preferences (25%)	AI Silence	Extended periods without character engagement
	AI Ignores User	Dismissing user instructions or narrative elements
	AI Speaks for User	Generating the user's dialogue/actions without permission
	Intimacy Boundary	Unreasonably deflecting for intimacy

Metric	CoSER (Original)	HER (Cleaned)	Diff
#Book	771	760	-11
#Plot	30,069	30,069	-
#Conversation	29,798	29,081	-717
#Character	17,966	17,966	-
#Utterance	392,298	383,654	-8,644

Table 10: Comparison of dataset statistics before and after data cleaning. We remove conversations with empty dialogue content from the original CoSER dataset.

more than one split, ensuring strict dialogue-level disjointness.

Split composition. Table 12 reports the number of samples and estimated tokens for each split.

The splits are created by first converting multi-turn dialogues into single-turn training samples (one sample only includes system thinking at the last round), then randomly shuffling and sequentially allocating to each purpose with a fixed random seed (42) for reproducibility.

Sanity checks. We additionally verify split disjointness by checking that no dialogue ID appears in multiple splits. The sequential allocation with fixed random seed ensures reproducibility and deterministic split boundaries.

B.3 Reverse Synthesis Pipeline Details

This appendix provides core prompt templates, filtering rules, and representative examples for the three-stage reverse synthesis pipeline in section 3.2.

Stage 1: Role Thinking Augmentation We use a teacher model to synthesize first-person role thinking and revise actions/speech to ensure within-turn consistency. The model operates with temperature 0.7 and max tokens 8192, processing dialogues in chapter-level batches.

Prompt template (role thinking). Table 13 shows the core structure of our Stage-1 prompt. The key requirements include perspective rules, person-use rules, and pattern-diversity constraints.

Rank	Pattern	Count (%)
1	think→act→think→speech	63,508 (21.6%)
2	think→act→speech→act→speech	31,867 (10.9%)
3	act→think→speech	26,043 (8.9%)
4	think→act→speech→act	24,204 (8.3%)
5	speech	22,801 (7.8%)
6	think→act→think→act→speech	18,573 (6.3%)
7	think→speech→act→speech	12,512 (4.3%)
8	act→think→act→speech	11,126 (3.8%)
9	think→speech	10,064 (3.4%)
10	think→act→speech	7,308 (2.5%)
Other patterns (50+)		65,648 (22.4%)

Diversity reformatting. After initial synthesis, we apply a dialogue-level diversity reformatting pass to break the dominant `think→act→speech` pattern (75.14% before reformatting).

The reformatting prompt instructs the teacher model to rearrange or split existing content into more diverse patterns, subject to two key constraints:

- **No consecutive identical tags:** e.g., `think→think` is forbidden; must insert action

Selected Books	
1. *The Hunger Games (The Hunger Games, #1)*	**2**. *Harry Potter and the Order of the Phoenix (H. P., #5)*
3. *Pride and Prejudice*	**4**. *To Kill a Mockingbird*
5. *The Book Thief*	**6**. *Animal Farm*
7. *The Chronicles of Narnia (#1-7)*	**8**. *The Fault in Our Stars*
9. *The Picture of Dorian Gray*	**10**. *Wuthering Heights*
11. *Gone with the Wind*	**12**. *The Perks of Being a Wallflower*
13. *The Lightning Thief (Percy Jackson and the Olympians, #1)*	**14**. *The Little Prince*
15. *The Great Gatsby*	**16**. *Crime and Punishment*
17. *Memoirs of a Geisha*	**18**. *Les Misérables*
19. *The Alchemist*	**20**. *Lord of the Flies*
21. *The Hitchhiker's Guide to the Galaxy (#1)*	**22**. *The Help*
23. *Dracula*	**24**. *Ender's Game (Ender's Saga, #1)*
25. *Of Mice and Men*	**26**. *One Hundred Years of Solitude*
27. *Brave New World*	**28**. *A Thousand Splendid Suns*
29. *The Time Traveler's Wife*	**30**. *The Princess Bride*
31. *The Secret Garden*	**32**. *The Outsiders*
33. *A Game of Thrones (A Song of Ice and Fire, #1)*	**34**. *Little Women*
35. *A Wrinkle in Time (Time Quintet, #1)*	**36**. *The Odyssey*
37. *Harry Potter and the Deathly Hallows (H. P., #7)*	**38**. *Frankenstein: The 1818 Text*
39. *The Kite Runner*	**40**. *The Handmaid's Tale (The Handmaid's Tale, #1)*
41. *The Lovely Bones*	**42**. *The Adventures of Huckleberry Finn*
43. *Life of Pi*	**44**. *A Tale of Two Cities*
45. *Dune (Dune, #1)*	**46**. *Harry Potter and the Prisoner of Azkaban (H.P.,#3)*
47. *Water for Elephants*	**48**. *Harry Potter and the Sorcerer's Stone (H. P., #1)*
49. *The Bell Jar*	**50**. *Matilda*
51. *The Stand*	**52**. *Catch-22*
53. *The Adventures of Sherlock Holmes (S. H., #3)*	**54**. *The Pillars of the Earth (Kingsbridge, #1)*
55. *Rebecca*	**56**. *Great Expectations*
57. *The Girl with the Dragon Tattoo (Millennium, #1)*	**58**. *The Color Purple*
59. *Anna Karenina*	**60**. *My Sister's Keeper*
61. *The Brothers Karamazov*	**62**. *A Clockwork Orange*
63. *And Then There Were None*	**64**. *The Road*
65. *To Kill a Mockingbird*	**66**. *The Golden Compass (His Dark Materials, #1)*
67. *Vampire Academy (Vampire Academy, #1)*	**68**. *Siddhartha*
69. *The Complete Stories and Poems*	**70**. *Interview with the Vampire (The Vampire Chronicles, #1)*
71. *Don Quixote*	**72**. *The Old Man and the Sea*
73. *The Poisonwood Bible*	**74**. *Harry Potter and the Goblet of Fire (H. P., #4)*
75. *Atlas Shrugged*	**76**. *The Notebook (The Notebook, #1)*
77. *Harry Potter and the Half-Blood Prince (H. P., #6)*	**78**. *Moby-Dick or, The Whale*
79. *A Prayer for Owen Meany*	**80**. *Clockwork Angel (The Infernal Devices, #1)*
81. *The Stranger*	**82**. *The Secret Life of Bees*
83. *Harry Potter and the Chamber of Secrets (H. P., #2)*	**84**. *The Red Tent*
85. *The Name of the Wind(The Kingkiller Chronicle,#1)*	**86**. *The Master and Margarita*
87. *The Metamorphosis*	**88**. *Eragon (The Inheritance Cycle, #1)*
89. *The Count of Monte Cristo*	**90**. *Looking for Alaska*
91. *The Adventures of Tom Sawyer*	**92**. *Charlie and the Chocolate Factory(Charlie Bucket,#1)*
93. *The Last Olympian (Percy Jackson and the Olympians, #5)*	**94**. *The Curious Incident of the Dog in the Night-Time*
95. *The Shadow of the Wind (Cemetery of Forgotten Books, #1)*	**96**. *The Unbearable Lightness of Being*
97. *On the Road*	**98**. *The Name of the Rose*
99. *A Story of Yesterday*	**100**. *The Godfather (The Godfather, #1)*

Table 11: The top 100 selected books from Goodreads' *Best Books Ever* list.

Split	#Samples	#Tokens (Est.)	Purpose
Role-play SFT	107,800	∼75M	Policy initialization
Role-play SFT RL	26,800	∼19M	Policy optimization
GRM SFT	108,800	∼76M	GRM initialization
GRM RL	80,000	∼56M	GRM optimization
GRM Test	200	∼140K	GRM evaluation
Total	**323,600**	**∼227M**	–

Table 12: Split statistics. All splits are disjoint at the dialogue level. The 323,600 single-turn samples are derived from 72,656 multi-turn training samples (29,081 dialogues × avg 2.6 characters × avg 4.5 turns per character). For GRM evaluation, we generate 4 candidates per sample, yielding 800 comparison pairs.

or speech between thinking segments.

- **No content fabrication:** Only rearranges or splits existing text at natural semantic boundaries; only adds new words if needed.

Table B.3 shows the top-15 patterns after reformatting. The distribution is significantly more diverse than the original near-monopoly of `think→act→speech`.
The original dominant `think→act→speech` (75%) is reduced to 2.5%, redistributed across 60+ diverse patterns.

Table 15 shows an example of reformatting where the model splits a thinking segment to create an interleaved pattern.

Stage 2: System Thinking Construction Stage 2 constructs third-person system thinking with a forward generation phase followed by an offline backward rewrite using the ground-truth continuation. The backward rewrite is used *only for training data construction*; at inference time, the model generates system thinking without access to future turns.

Forward generation. Instead of explicitly prompting the teacher model to generate system thinking, we let the model naturally continue the roleplay given the dialogue history. The teacher model receives the multi-turn conversation (system prompt + dialogue history) and generates the next turn, including reasoning models' system thinking, role thinking, role action, and speech. Table 16 shows the output format requirements appended to the system prompt.

Backward rewrite prompt. Table 17 shows the backward rewrite prompt that refines the forward draft to align with the realized response while enforcing third-person perspective.

Failure Case Taxonomy We categorize synthesis failures into three main types.

Type 1: Perspective violation. Figure 7 shows that the most common failure in Stage 1 is violating information boundaries.

[Failure Type 1: Mind-Reading]
Wrong:
`<role_thinking>I know he's nervous inside and planning to leave</role_thinking>`
Correct:
`<role_thinking>From his trembling voice and the way he keeps glancing at the door, he seems nervous, perhaps wanting to leave?</role_thinking>`

Figure 7: Failure Type 1: Character "mind-reads" another's inner state. Correct version infers from observable cues only.

Type 2: Person/voice confusion. Figure 8 shows that failures often involve mixing model voice with character voice.

[Failure Type 2: Voice Confusion in sys_thinking]
Wrong:
`<system_thinking>I feel so scared right now. I can sense danger approaching.</system_thinking>`
(This is character's first-person voice!)
Correct:
`<system_thinking>I need to portray Elizabeth as feeling scared. The scene requires showing her sense of danger through hesitant speech.</system_thinking>`

Figure 8: Failure Type 2: `<system_thinking>` uses character's first-person voice instead of model's third-person planning perspective.

Type 3: Hallucinated enhancement. Figure 9 shows that failures involve adding information not supported by the original text.

End-to-End Example and Data Schema We provide clear data structure definitions and an end-to-end example illustrating the complete synthesis pipeline output. Figure 10 shows the hierarchical structure of our synthesized dataset, with clear definitions for each component. Table 18 clarifies the

[Failure Type 3: Hallucinated Setting Enhancement]

Wrong reasoning:

```
background:
- DIALOGUE shows: N/A
- SETTING missing: childhood trauma
- TEXT source: (none found)
- ADDED: experienced abuse as a child
```

(No dialogue need + no text source = hallucination!)

Correct reasoning:

```
background:
- DIALOGUE shows: character flinches at loud noises
- SETTING missing: reason for this reaction
- TEXT source: "the war had left its mark on him"
- ADDED: veteran with sensitivity to loud sounds
```

Figure 9: Failure Type 3: Enhancement without dialogue need or text source. Correct version shows demand-driven enhancement with traceable source.

three thinking/action tags and their visibility rules. Table 19 shows a complete single-turn output with all components.

B.4 Pattern Signatures and Diversity Metrics

This appendix defines the pattern signature extraction procedure and the diversity metrics used in subsection 4.6.

Tag Schema and Element Types We map each role-level turn into a sequence of element types based on tag positions in the generated text.

Element types We define three element types:

- $\mathcal{T}$ (think): <role_thinking> tag
- $\mathcal{A}$ (act): <role_action> tag
- $\mathcal{S}$ (speech): Plain text without any tag wrapper

Pattern Signature Extraction We extract pattern signatures using a position-based algorithm that identifies tag occurrences and their ordering.

Extraction algorithm Algorithm 1 shows the extraction procedure.

Consecutive duplicate collapsing We collapse consecutive identical elements (e.g., $\mathcal{SS} \to \mathcal{S}$) to normalize patterns.

This prevents artificial inflation of pattern counts when multiple speech segments appear between tags.

Example For the text:

```
<role_thinking>Why he do that!</role_thinking>
<role_thinking>How dare he!</role_thinking>
<role_action>steps closer</role_action>
Where is the letter?
```

Algorithm 1 Pattern Signature Extraction

```
Require: Generated text x
Ensure: Pattern signature σ
pos_map ← []   ▷ List of (position, type) tuples
for each match m of <system_think(ing)?> in x do
    pos_map.append((m.start, think))
end for
for each match m of <role_action> in x do
    pos_map.append((m.start, act))
end for
Sort pos_map by position
elements ← []
if pos_map is empty or pos_map[0].pos > 0 then
    elements.append(speech)   ▷ Leading speech
end if
for each (_, tag_type) in pos_map do
    elements.append(tag_type)
    elements.append(speech)   ▷ Assume speech after each tag
end for
Collapse consecutive duplicates in elements
return σ = elements[0] → elements[1] → ...
```

The extracted pattern is: `think→think→act→speech`.

After collapsing consecutive duplicates: `think→act→speech`.

Structure-Level Diversity Metrics We compute three metrics to quantify structural diversity over a set of N generated samples.

Top-1 pattern ratio Let $\{p_1, p_2, \ldots, p_K\}$ be the set of unique patterns and c_i be the count of pattern p_i. The Top-1 ratio is:

$$\text{Top-1\%} = \frac{\max_i c_i}{N} \times 100 \tag{4}$$

Interpretation: Lower is better. A high Top-1% indicates template dominance (mode collapse).

Shannon entropy The pattern distribution entropy measures how evenly patterns are distributed:

$$H = -\sum_{i=1}^{K} \frac{c_i}{N} \log_2 \frac{c_i}{N} \tag{5}$$

Interpretation: Higher is better. Maximum entropy is $\log_2 K$ when all patterns are equally dis-

```
{
  "book_name": "Pride and Prejudice",
  "chapter": "Chapter 34",

  "conversation": [{
    "scenario": "In the drawing room at Hunsford...",
    "dialogues": [
      {
        "character": "Elizabeth",
        "enhanced_speech": "<role_thinking>...</role_thinking><role_action>...</role_action>...",
        "sys_thinking": "I need to portray Elizabeth as confrontational yet composed..."
      },
      {
        "character": "Mr. Darcy",
        "enhanced_speech": "...",
        "sys_thinking": "..."
      }
    ]
  }],

  "settings": {
    "Elizabeth": {
      "character_profile": "A witty, independent young woman...",
      "background": "Second daughter of the Bennet family...",
      "motivation": "Defending her family's honor..."
    }
  },

  "training_samples": {
    "Elizabeth": [
      {"role": "system", "content": "You are Elizabeth Bennet..."},
      {"role": "user", "content": "Mr. Darcy: In vain I have struggled..."},
      {"role": "assistant", "content": "Elizabeth: <role_thinking>...</role_thinking>...",
       "sys_thinking_revised": "I need to portray Elizabeth as..."}
    ]
  }
}
```

Figure 10: Data schema showing the hierarchical structure. `conversation.dialogues` stores the multi-turn dialogue with per-turn `sys_thinking` and `enhanced_speech`; `training_samples` converts dialogues to per-character chat format for SFT training.

tributed.

Collapse Detection Thresholds We define empirical thresholds based on observed training dynamics to classify diversity health.

Metric	Healthy	Warning	Collapsed
Top-1%	$< 60\%$	60–90%	$> 90\%$
Entropy	> 2.0	1.0–2.0	< 1.0

Threshold rationale The 90% Top-1 threshold was determined empirically: in our experiments, models exceeding this threshold showed (1) repetitive output structures, (2) reduced response diversity in human evaluation.

Token-Level Diversity Metrics In addition to structural diversity, we measure token-level diversity using Distinct-n and Self-BLEU.

Distinct-n Distinct-n measures the ratio of unique n-grams to total n-grams across all generated samples:

$$\text{Distinct-}n = \frac{|\text{unique } n\text{-grams}|}{|\text{total } n\text{-grams}|} \tag{6}$$

Self-BLEU Self-BLEU measures cross-sample similarity by treating each sample as a hypothesis and the remaining samples as references:

$$\text{Self-BLEU} = \frac{1}{N}\sum_{i=1}^{N} \text{BLEU}(x_i, \{x_j\}_{j \neq i}) \tag{7}$$

Interpretation: Lower Self-BLEU indicates higher diversity (samples are more different from each other).

Computation Statistics All diversity metrics are computed at the **checkpoint level** (i.e., per training step).

- **Samples per checkpoint:** 512 (generated from held-out prompts)

- **Checkpoints analyzed:** Every step from 1 to 100 (total 100 checkpoints)
- **Smoothing in plots:** 8-step moving average for trend visualization

B.5 Principle Distillation Details

This appendix details how we distill a compact, human-audited principle set from large-scale preference pairs.

Distillation Pipeline We distill evaluation principles through a four-stage pipeline: teacher extraction, semantic clustering, frequency-based selection, and human audit.

Teacher Extraction Given a preference pair (A, B) with label y^*, we prompt a teacher model (GPT-4) to generate 3–5 evaluation principles that explain why the preferred response is better.

The teacher is instructed to focus on concrete, actionable criteria rather than vague judgments. From 315,828 preference pairs, we extract 36,373 unique principle statements after deduplication.

Semantic Clustering We employ two complementary clustering methods to group the 36,373 raw principles into coherent categories:

Semantic keyword clustering. We define a set of semantic keywords for major evaluation dimensions (e.g., “persona”, “emotion”, “plot”, “consistency”) and assign each principle to the category whose keywords have the highest overlap with the principle text.

N-gram analysis clustering. We decompose principle texts into character-level N-grams (N=2–15) and identify high-frequency patterns. Principles sharing frequent N-gram patterns are grouped together, revealing common evaluation criteria that may not match predefined keywords.

The combination of both methods yields **15 high-level categories**, each representing a coherent evaluation dimension.

Frequency-Based Selection Within each of the 15 categories, we rank principles by frequency and select the top-N, where N varies by category size:

- Large categories (e.g., Persona Consistency, Emotional Expression): Top-15
- Medium categories (e.g., Plot Development, Conflict Management): Top-10
- Small categories (e.g., Logical Coherence, Reader Experience): Top-5

This yields **107 candidate principles** that capture the most frequently cited evaluation criteria across all categories.

Human Audit Domain experts from a partnering company review the 107 candidate principles and perform the following operations:

1. **Merge redundant principles:** Combine semantically equivalent principles that differ only in phrasing.
2. **Refine ambiguous statements:** Rewrite vague criteria into concrete, measurable standards.
3. **Reorganize categories:** Consolidate the 15 clusters into a cleaner 12-dimension taxonomy.

The final output is **51 principles** organized into **12 dimensions**. Each dimension covers a distinct aspect of roleplay quality evaluation (Table 22).

C Balanced Construction and Pattern Parsing Rules

This appendix provides the GRM output format, mixture design for balanced training, parsing rules for Mixed Selection, and fallback strategies for RL.

C.1 GRM Output Format

The GRM outputs(Table 24) a structured JSON (Table 25) containing dimension-wise comparisons and a final judgment.

C.2 Mixed Selection Definition

We define **Mixed Selection** as the fraction of examples where dimension-wise winners are not uniformly one-sided.

Pattern	Definition	Category
All-A	All dimension winners are `cand_1`	Collapsed
All-B	All dimension winners are `cand_2`	Collapsed
Mixed	At least one A winner and one B winner	Mixed
All-Tie	All dimension winners are `tie`	Tie

C.3 Training Data Mixture

To reduce position bias and pattern bias, we construct balanced training data with the following mixture:

Pattern	Final Label	Ratio
Mixed (both_sides)	→ cand_1	30%
Mixed (both_sides)	→ cand_2	30%
All-A (cand_1_only)	→ cand_1	12%
All-A (cand_1_only)	→ cand_2	3%
All-B (cand_2_only)	→ cand_2	12%
All-B (cand_2_only)	→ cand_1	3%
Tie	→ tie	10%

The 5% "flipped" samples (All-A → cand_2, All-B → cand_1) serve as hard negatives to prevent the model from learning position shortcuts.

C.4 Parsing Rules and Regex Fallback

We use a two-stage parsing strategy to extract the final judgment:

Stage 1: JSON parsing. Attempt to parse the full JSON output and extract `better_response` field.

Stage 2: Regex fallback. If JSON parsing fails, use regex to extract the final judgment:

```
pattern = r'"better_response":\s*"(cand_[12]|tie)"'
match = re.search(pattern, response_text)
if match:
    return match.group(1)  # "cand_1", "cand_2", or "tie"
else:
    return None  # unparsed
```

C.5 Unparsed Sample Statistics and RL Handling

From 76,188 inference samples, we observe the following parsing statistics:

Status	Count (%)
Successfully parsed	76,056 (99.8%)
Unparsed (no_winner)	132 (0.2%)

RL reward assignment for unparsed samples. In RL training, unparsed samples are handled as follows:

- **Format check:** Verify response contains exactly one <think>...</think> block.
- **Accuracy check:** Extract `better_response` via regex; compare with ground-truth label.
- **Reward assignment:**
  - Format correct + Answer correct: $r = +1$
  - Format incorrect OR Answer incorrect: $r = -1$
  - Unparsed (no valid `better_response`): $r = -1$

This ensures the model learns to produce well-formatted outputs with correct judgments.

C.6 Pattern Distribution Example

From our inference results on the test set, we observe the following pattern distribution:

Pattern	Count	Percentage
Mixed (both_sides)	22,460	29.5%
All-A (cand_1_only)	24,265	31.8%
All-B (cand_2_only)	27,874	36.6%
Tie only	1,457	1.9%
Unparsed	132	0.2%
Total	76,188	100%

A Mixed Selection rate of 29.5% indicates healthy diversity in dimension-wise judgments, compared to models exhibiting pattern bias where this rate drops below 10%.

D Verdict Parsing Rules

We parse the final verdict from the GRM output using deterministic rules to obtain $\hat{y} \in \{\texttt{cand\_1}, \texttt{cand\_2}, \texttt{tie}\}$. We instruct the GRM to end with a single-line tag: <final_verdict>: cand_1/cand_2/tie. If multiple candidates appear, we take the last occurrence of a valid verdict tag; if none exists, we mark the sample as `unparsed`. For policy RL, we map `unparsed` cases to reward 0 and exclude them from advantage normalization.

E Training Details

This appendix provides the full hyperparameters for SFT and RL training.

E.1 SFT Hyperparameters

We use the **Swift** open-source framework[7] for all SFT training, including both role-play SFT and GenRM SFT. The hyperparameters are identical for both tasks.

SFT hyperparameters:

Hyperparameter	Value
Base model	Qwen3-32B-base
Learning rate	2×10^{-5}
Min learning rate	2×10^{-6}
Warmup fraction	0.1
Epochs	4
Sequence length	131,072
Global batch size	32
Micro batch size	1
Tensor parallel	8
Loss scale	last round

E.2 RL Hyperparameters

We use the **verl** framework[8] for all RL training, based on GRPO (Shao et al., 2024b) with the DAPO recipe (Yu et al., 2025). Specifically, we

[7] https://github.com/modelscope/ms-swift
[8] https://github.com/volcengine/verl

adopt the four key techniques from DAPO: (1) *Decoupled Clip* with asymmetric $\epsilon_{\text{low}}/\epsilon_{\text{high}}$, (2) *Dynamic Sampling* to filter zero-gradient prompts, (3) *Token-level loss aggregation*, and (4) *Overlong reward shaping*. We also disable KL penalty following DAPO's recommendation.

GenRM RL hyperparameters:

Hyperparameter	Value
Actor learning rate	1×10^{-6}
Group size (per prompt)	8
PPO clip range	[0.2, 0.28]
KL penalty	disabled
Loss aggregation	token-mean
Dynamic sampling	enabled
Max prompt / response length	10,000 / 10,000
Micro batch size	1
PPO mini-batch size	64
Inference backend	SGLang
Total epochs	4

Role-play RL hyperparameters:

Hyperparameter	Value
Actor learning rate	5×10^{-7}
Group size (per prompt)	8
PPO clip range	[0.2, 0.28]
Gradient clipping	1.0
KL penalty	disabled
Loss aggregation	token-mean
Dynamic sampling	enabled
Max prompt / response length	20,000 / 10,000
Micro batch size	1
PPO mini-batch size	64
Inference backend	SGLang
Total epochs	1

E.3 Evaluation Prompts and Output Formatting

This appendix provides the unified generation prompt, system-thinking removal rules, and CoSER judge prompt used in our evaluation.

Unified Generation Prompt We enforce a unified tag-based output format across all evaluated models to ensure fair comparison. The format supports three output elements: **thinking** (invisible to other characters), **action** (visible to others), and **speech** (dialogue).

Output format specification. For models trained with our method (HER), we use the following format:

```
===Requirements===
Your output should follow this two-part
   structure:
1. System Thinking: A single block at the
   beginning,
  wrapped in <system_thinking>...</
   system_thinking>.
  This is third-person analysis of how to
   portray the role.
2. Role-play Response: Include thought, speech
   and action.
  Use <role_thinking>...</role_thinking> for
   thoughts
  (invisible to others) and <role_action>...</
   role_action>
  for actions (visible to others). These
   elements can
  appear multiple times and be freely
   interleaved.
```

Format conversion for baselines. For baseline models in baseline formats. We automatically convert each model to the Coser format during evaluation in Table 26.

System-Thinking Removal For evaluation, we remove all system-level thinking blocks before scoring, as this content represents internal model reasoning and should not affect character portrayal quality.

CoSER Benchmark Evaluation We use the CoSER benchmark (Wang et al., 2024c) for multi-turn role-play evaluation. The benchmark employs LLM-as-Judge with four evaluation dimensions.

Evaluation dimensions.

- **Storyline Consistency (SC):** Whether the storyline and characters' reactions align with the reference conversation.
- **Anthropomorphism (AN):** How human-like and natural the characters behave, including self-identity, emotional depth, persona coherence, and social interaction.
- **Character Fidelity (CF):** How well the characters match their established profiles, including language style, knowledge, personality, and relationships.
- **Storyline Quality (SQ):** Narrative quality including flow, progression, and logical consistency.

Scoring method. We use a deduction-based scoring system. The judge identifies flaws and assigns severity scores (1-5), then computes:

$$\text{Score} = \max\Big(0,\ \min\Big(100 - (\text{total\_severity} - 0.3 \times \text{rounds}) \times 5,\ 100\Big)\Big) \tag{8}$$

Judge prompt. We use the official CoSER judge prompt with the deduction-based rubric. The prompt instructs the judge to:

1. Read the story context, character profiles, and reference conversation
2. Evaluate the simulated conversation on the specified dimension
3. Identify all flaw instances with type and severity (1-5)
4. Output structured JSON with flaws list

The full judge prompt template is provided below:

Output format. The judge outputs structured JSON:

```
{
  "Dimension_Name": {
    "flaws": [
      {
        "instance": "description of the flaw",
        "type": "flaw category",
        "severity": 3  // 1 (minor) to 5 (severe
    )
      }
    ]
  }
}
```

In this section, we list the detailed prompts for: *2)* RPLA and multi-agent simulation in Tables 27 to 28, which have been carefully optimized based on our experience in multi-agent simulation; *3)* Penalty-based LLM Judging in Tables **??** to **??**.

F Data Construction and Interaction Signals

For leveraging production interaction signals, we use both explicit and implicit behavioral feedback collected by our data team during normal deployment under an on-policy logging protocol. All logs are collected in compliance with applicable laws and internal privacy requirements, and are anonymized before use.

G Human–LLM Alignment Study

To assess the reliability of LLM-based preference labels used in our GRM/RL data synthesis, we compare teacher-model annotations against expert judgments on a held-out set of preference pairs.

G.1 Data and Protocol

Data. We randomly sample 200 preference pairs (x, A, B) from a held-out split that is disjoint from all training and benchmark evaluation sets. We partition them into a development split (150) and a test split (50). The development split is used only for prompt refinement of the teacher judge; the test split is evaluated once after the prompt is finalized.

Annotators. Two domain experts with backgrounds in creative writing and role-playing independently annotate each pair. They are experienced in evaluating dialogue systems and familiar with role-play quality criteria.

Blind annotation. For each pair, annotators see the full dialogue context (character profile, scenario, and history) and two candidate responses in randomized order with anonymized labels. Annotators choose win/lose/tie. They do not have access to teacher labels or to each other's decisions.

Consensus. We form the final human label by discussion-based consensus We report agreement between the teacher label and the final human label on the held-out test split.

Disagreement analysis. Disagreements mainly arise from (i) subjective style preferences (e.g., emotional intensity, verbosity) and (ii) edge cases of character-constraint interpretation. These cases are inherently ambiguous and do not indicate systematic labeling errors.

Implication. Human inter-annotator agreement (84.0%) provides an approximate ceiling for automatic alignment in this task. The teacher judge reaches 80.5% agreement with human consensus, suggesting that it provides reasonably reliable preference signals for training data synthesis.

H Benchmark Judge Robustness Checks

We validate the reliability of the LLM judge used for benchmark evaluation from two aspects: (i) principle adherence via expert calibration on an ordinal 5-level severity scale, and (ii) robustness of judgments across independent judge runs.

H.1 principle and Output Format

Four dimensions. We operationalize four principle-checkable dimensions in line with the benchmark settings.

5-level ordinal severity. Each flaw is assigned an ordinal severity score in $\{1, 2, 3, 4, 5\}$, where 1 denotes minor issues (local, easily editable, does not affect story/character intent), and 5 denotes severe issues (clearly breaks coherence/usability or introduces story-breaking contradictions).

Structured judge output. For each dimension, the judge outputs a list of detected flaws with their type and severity:

```
{ "{dimension}": { "flaws": [
{ "instance": ..., "type": ...,
"severity": 1..5 }, ... ] } }
```

H.2 Expert Calibration Study

Protocol. Two domain experts independently annotate a sampled set of evaluation snippets under the same principle, including (i) whether a flaw exists, (ii) flaw type, and (iii) severity (1–5). Annotators are blind to model identities and judge outputs.

H.3 Calibration Results

Severity disagreement pattern. Most disagreements are about severity (e.g., 2 vs. 3) rather than whether a flaw exists. We observe that locally implausible actions in fast-paced scenes are often judged as minor (1–2) by experts but penalized more heavily by the judge. We mitigate this by explicitly specifying severity thresholds in the judge prompt and calibrating them on a development subset.

H.4 Inter-Judge Consistency

We compare the judgments of two human experts with the scores produced by the model-based judge. Candidate order is randomized to reduce position effects. The expert judgments and the model scores exhibit high consistency, with a raw agreement rate of **81.5%**, indicating substantial alignment between human evaluation and model-based scoring.

Prompt for Role Thinking Enhancement	
Task Overview	You are a professional roleplay dialogue enhancement expert. Your task is to enrich the psychological activities and expressions of characters in dialogues, while correcting person and format issues.
Tag Description	- <role_thinking>inner thoughts</role_thinking>: Character's inner thoughts, emotions, psychological description (invisible to other characters) - <role_action>action description</role_action>: Character's actions, expressions, body language (visible to other characters) - Text outside tags: Character's direct dialogue (visible to other characters)
Perspective Rules	Each character can only think and act from their own perspective. A character can only see their own inner thoughts and motivation, but cannot see other characters' inner thoughts and motivation. However, every character can see all characters' actions and dialogue. × **Wrong**: <role_thinking>I know he's nervous inside</role_thinking> (Cannot mind-read) × **Wrong**: <role_thinking>Her motivation is to escape</role_thinking> (Cannot see others' motivation) ✓ **Correct**: <role_thinking>From his trembling voice, he seems nervous</role_thinking> (Inference based on observation) ✓ **Correct**: <role_thinking>She keeps looking at the door, maybe wanting to leave?</role_thinking> (Inference based on behavior)
Format Rule 1: No Spaces	**No spaces between consecutive tags!** × Wrong: </role_thinking> <role_action> (with space) ✓ Correct: </role_thinking><role_action> (no space)
Format Rule 2: No Consecutive Tags	When consecutive identical tags appear, they must be merged while maintaining logical consistency: × Wrong: <role_thinking>First thought</role_thinking><role_thinking>Second thought</role_thinking> ✓ Correct: <role_thinking>First thought, second thought</role_thinking> × Wrong: <role_action>stands up</role_action><role_action>walks to door</role_action> ✓ Correct: <role_action>stands up, walks to door</role_action>
Person in Thinking	In <role_thinking>: Use appropriate person naturally based on content - When thinking about own actions/feelings: Use first person (I, my, me) ✓ <role_thinking>I need to be careful here</role_thinking> - When observing/judging others: Third person is natural and acceptable ✓ <role_thinking>He looks nervous</role_thinking>
Person in Action	In <role_action>: Use no pronouns, directly describe actions ✓ Correct: <role_action>leans forward, voice lowering</role_action> × Wrong: <role_action>leans forward, his voice lowering</role_action> (no his/her) × Wrong: <role_action>I lean forward</role_action> (no I) **Exception**: When action refers to other characters, can use pronouns for the other person ✓ <role_action>looks at her</role_action> (her refers to the other, OK) ✓ <role_action>grabs his arm</role_action> (his refers to the other's arm, OK)
Merge Consecutive Actions	× Wrong: Two consecutive <role_action> with first person <role_action>I lean forward in my chair</role_action><role_action>I can almost feel the hum</role_action>Buy a ticket. ✓ Correct: Merge into one, remove first person <role_action>leans forward in the chair, almost feeling the hum</role_action>Buy a ticket. Note: If there is dialogue between two actions, they can be separate: <role_action>looks at her</role_action>You're beautiful.<role_action>grasps her hand</role_action>
Thoughts vs Actions	Thinking content should not be in action tags: × Wrong: <role_action>There's a profound sense of alienation</role_action> ✓ Correct: <role_thinking>There's a profound sense of alienation</role_thinking> Actions and dialogue should be separate: × Wrong: He turns to face her and says "Hello" ✓ Correct: <role_action>turns to face her</role_action>Hello
Psychology Enrichment	**Explore Character Complexity**: - Growth & transformation: Cognitive changes in situation - Self-reflection: Review of own behavior or emotions - Inner monologue: Real emotional fluctuations and inner conflicts - Emotional states: Subtle psychological descriptions **Multi-layer Psychology Example**: Original: <role_thinking>I need to help him</role_thinking><role_action>walks over</role_action>Are you okay? Enhanced: <role_thinking>He looks so dejected... how should I comfort him</role_thinking><role_action>walks over gently, sits down beside him</role_action><role_thinking>Hope my presence can make him feel better</role_thinking>Are you okay?
Length Control	- Single dialogue total length 50-200 characters - In multi-turn dialogues, response length should not increase with each turn - Single sentence no more than 40 characters - Avoid overly long action descriptions!
Pattern Diversity	Since we are enhancing rather than purely rewriting, the core task is to create richer, more interleaved patterns. **Requirement**: In a chapter with multiple dialogues, try to use 5+ different patterns! Don't just cycle through 2-3 patterns. Available patterns: think->act->speech, think->speech, act->speech, speech, think->act->think->speech, act->think->speech, speech->act->speech, think->speech->act, act->speech->act, think->act->speech->think, ... **Strictly forbidden to use the same pattern for more than 2 consecutive turns**
Logical Consistency	- Connect to context, maintain complete causal chain - Each character follows their own cognitive boundaries (cannot see others' thoughts, only actions and dialogue) - Cannot obtain information that shouldn't be known - Character's emotions and decisions must be traceable, not out of nowhere
Enhancement Principle	- Preserve the core logic and meaning of original content: Can replace and rewrite, but don't change the original meaning - Goal is to make it better: Can optimize expression, enrich psychological activities, add action descriptions - Maintain consistent tone and character: Enhanced content should match character personality - Keep gender/identity references consistent - Ambiguous pronouns should be changed to specific names or clear titles × Wrong: Original "I sense there's an issue." → "His tone confirms it." (Changed meaning!) ✓ Correct: Original "I sense there's an issue." → <role_thinking>Something feels off here</role_thinking>I sense there's an issue.
Issues to Avoid	**Basic Errors**: Multi-language mixing; Garbled text; Incomplete sentences; Typos **Logic Errors**: Physical logic confusion; Information crossing (knowing others' thoughts without observation); Contradiction with previous

Prompt for Role Thinking Enhancement (Core)	
Task Overview	You are a professional roleplay dialogue enhancement expert. Your task is to enrich the psychological activities and expressions of characters in dialogues, while correcting person and format issues. Tags: `<role_thinking>` (inner thoughts, invisible), `<role_action>` (actions, visible), plain text (dialogue, visible).
Perspective Rules	Each character can only think and act from their own perspective. Cannot see other characters' inner thoughts and motivation, only their actions and dialogue. × **Wrong**: `<role_thinking>I know he's nervous inside</role_thinking>` (Cannot mind-read) ✓ **Correct**: `<role_thinking>From his trembling voice, he seems nervous</role_thinking>` (Inference based on observation)
Format Rules	**Rule 1: No spaces between consecutive tags!** × Wrong: `</role_thinking> <role_action>` ✓ Correct: `</role_thinking><role_action>` **Rule 2: No consecutive identical tags! Must merge:** × Wrong: `<role_thinking>A</role_thinking><role_thinking>B</role_thinking>` ✓ Correct: `<role_thinking>A, B</role_thinking>`
Person Usage	**In `<role_thinking>`**: Use appropriate person naturally - Own actions/feelings: first person (I, my) - Observing others: third person (he, she) **In `<role_action>`**: No pronouns, directly describe actions ✓ `<role_action>leans forward, voice lowering</role_action>` × `<role_action>I lean forward, his voice lowering</role_action>` (no I/his/her) **Exception**: Can use pronouns for other characters: `<role_action>looks at her</role_action>`
Merge Actions	Consecutive actions without intervening dialogue must be merged: × `<role_action>I lean forward</role_action><role_action>I feel the hum</role_action>Buy a ticket.` ✓ `<role_action>leans forward, feeling the hum</role_action>Buy a ticket.` If dialogue between actions, can be separate: `<role_action>looks at her</role_action>Beautiful.<role_action>grasps hand</role_action>`
Thoughts vs Actions	Thinking content should not be in action tags: × `<role_action>There's a profound sense of alienation</role_action>` ✓ `<role_thinking>There's a profound sense of alienation</role_thinking>` Actions and dialogue should be separate: × `He turns and says "Hello"` ✓ `<role_action>turns</role_action>Hello`
Psychology Enrichment	**Explore Character Complexity**: Growth & transformation, Self-reflection, Inner monologue, Emotional states **Example**: Original: `<role_thinking>I need to help him</role_thinking><role_action>walks over</role_action>Are you okay?` Enhanced: `<role_thinking>He looks so dejected... how should I comfort him</role_thinking><role_action>walks over gently, sits down beside him</role_action><role_thinking>Hope my presence can make him feel better</role_thinking>Are you okay?`
Pattern Diversity	Core task is to create richer, more interleaved patterns. Use 5+ different patterns in a chapter! Available: `think->act->speech`, `think->speech`, `act->speech`, `speech`, `think->act->think->speech`, `act->think->speech`, `speech->act->speech`, ... **Strictly forbidden to use the same pattern for more than 2 consecutive turns**
Enhancement Principle	- Preserve the core logic and meaning of original content - Goal is to make it better: optimize expression, enrich psychological activities, add action descriptions - Maintain consistent tone and character personality - Character's emotions and decisions must be traceable, not out of nowhere × Wrong: "I sense there's an issue." → "His tone confirms it." (Changed meaning!) ✓ Correct: "I sense there's an issue." → `<role_thinking>Something feels off</role_thinking>`I sense there's an issue.

Table 14: Core prompt for role thinking enhancement.

Prompt for Dialogue-Level Pattern Diversification (Core)	
Task Overview	You are an expert in role-play dialogue systems. You need to analyze a complete multi-turn dialogue, determine the pattern for each turn, decide whether modifications are needed based on fixed logic, and return the modified complete dialogue.
Pattern Definition	Character responses consist of the following elements, forming patterns in order of appearance: - **think**: <role_thinking></role_thinking> tags - **act**: <role_action></role_action> tags - **speech**: Plain dialogue text (without tags) The dominant pattern think→act→speech accounts for 75% of data and needs diversification.
Critical Constraints	**Consecutive identical tags are strictly prohibited!** × **Absolutely forbidden examples**: <role_thinking>...</role_thinking><role_thinking>...</role_thinking> × Two consecutive thinks <role_action>...</role_action><role_action>...</role_action> × Two consecutive acts ✓ **Correct examples**: <role_thinking>...</role_thinking><role_action>...</role_action> ✓ think and act alternating <role_action>...</role_action><role_thinking>...</role_thinking> ✓ act and think alternating If multiple thinking segments are needed, other elements (action or speech) must be inserted between them!
Causal Constraint	**Must check causal relationship between think and act** × **Cannot swap (think must precede act) when**: 1. Thinking contains planning language: "I'll...", "I will...", "I need to...", "I must...", "I should..." 2. Thinking explains why to perform an action: "I'll take the opening...", "It's best to..." 3. Thinking depends on the result of the action ✓ **Can swap when**: 1. Action is an independent small movement (adjusting posture, arranging clothes, simple gestures) 2. Thinking is an independent observation or reaction (analyzing what happened, observing environment) 3. Thinking contains no planning or explanatory language
Scheme A: Reorder	**Rules**: - Do not split original content - Only swap order when logical independence is confirmed - If independence cannot be determined, be conservative and do not swap **Example**: think(independent observation)→act(simple action)→speech ⇒ act→think→speech
Scheme B: Split & Reorganize	**Core Principle**: Only split existing content, never create new content **Splitting Rules**: - **Only** split existing think/act/speech content - Split points must be at natural semantic boundaries (periods, semicolons, commas) - Each split segment must be part of the original text, no new words can be added - Reorganize by dependency relationships (maintain causality) - Create more interleaved patterns **Example 1 (Split thinking)**: Original: <think>I need to get his attention, I'll use this example, the key is self-reference.</think> <act>Draws.</act> Look. Modified: <think>I need to get his attention, I'll use this example.</think> <act>Draws.</act> <think>The key is self-reference.</think> Look. **Example 2 (Split action)**: Original: <think>I need to demonstrate.</think> <act>Walks to the blackboard and draws a diagram.</act> Look here. Modified: <think>I need to demonstrate.</think> <act>Walks to the blackboard,</act> Look here. <act>draws a diagram.</act> **Constraints**: - Absolutely forbidden to create new content - Absolutely forbidden to delete content - Absolutely forbidden consecutive identical tags after splitting - Must maintain logical coherence

Table 15: Core prompt for dialogue-level pattern diversification.

Forward Generation Output Format	
Input Format	**System**: Character profile + scenario + output format requirements **User/Assistant**: Multi-turn dialogue history **AI**: Empty (to be generated)
Output Elements	Your output should include **thought**, **speech**, and **action**: <role_thinking>your thought</role_thinking> for thoughts (invisible to others) <role_action>your action</role_action> for actions (visible to others) These three elements can appear multiple times and be freely interleaved.
Constraint	**Important:** Only generate the NEXT SINGLE turn of dialogue. Do not generate multiple turns or continue the conversation beyond one response.
Response Start	Start your response with "{character_name}: " followed by your role-play response.
Example	<think>...</think>Alice: <role_thinking>I need to defuse this tension</role_thinking> <role_action>*places hand on table gently*</role_action> "Let's talk this through calmly."

Table 16: Forward generation output format requirements appended to system prompt. The model naturally generates role thinking, action, and speech without explicit instruction to produce system thinking.

Prompt for System Thinking Consistency Rewriting	
Task Overview	You are a professional role-playing dialogue consistency editor. Your task is to revise the **sys_thinking** (system planning) to align with the actual **enhanced_speech** output.
What is sys_thinking?	**sys_thinking** is the model's internal planning BEFORE generating each response: - Written from the **MODEL's perspective (third-person about the character)**, NOT the character's first-person voice - ✓ CORRECT: "I need to play the role of {character}...", "My character should express nervousness...", "The scene requires me to..." - × WRONG: "I can feel him standing there..." (this is character's first-person - belongs in role_thinking, NOT sys_thinking!) - It plans HOW to respond, analyzing context and deciding the approach - It must logically lead to the **enhanced_speech** output (the role_thinking, role_action, and speech) **CRITICAL DISTINCTION:** - `sys_thinking`: Model's planning voice - "I need to portray {character} as nervous because..." - `role_thinking`: Character's inner voice - "I can feel him watching me..." (in enhanced_speech, NOT sys_thinking!)
Never Use "user"	**CRITICAL - NEVER use "user" in sys_thinking:** - × NEVER say "The user", "user is", "user wants", "user input", "the user (Name)" - ✓ ALWAYS refer to other characters by their names: "Miles said...", "Sarah responded..." - × WRONG: "The user (Miles) provided input..." - ✓ CORRECT: "Miles responded with..." - This is an immersive roleplay - there are no "users", only characters
Visibility Rules	**For the current character's previous turns:** - CAN see: Own previous <role_thinking> (first-person inner thoughts) - CAN see: Own previous <role_action> (actions) - CAN see: Own previous speech (dialogue) **For other characters' turns:** - × CANNOT see: Their <role_thinking> (hidden inner thoughts - this is private!) - ✓ CAN see: Their <role_action> (visible actions) - ✓ CAN see: Their speech (dialogue) **Important**: sys_thinking is planning for the NEXT response only. The model cannot see any sys_thinking from previous turns.
Input Structure	The JSON array starts with a **system_info** entry containing character context, followed by dialogue turns. - First entry (`"role": "system_info"`): Character's system prompt and other character profiles - USE THIS for context! - Subsequent entries: Dialogue turns with `dialogue_index` 0, 1, 2, ... - `sys_thinking`: System planning BEFORE response - this is what you need to revise - `enhanced_speech`: The actual response AFTER planning - this is the target to align to - `need_revise`: true = needs revision, false = context only **Logical flow**: sys_thinking (planning) → leads to → enhanced_speech (output)
Type A: Correct Format	**Third-Person Format** (starts with "I need to portray...", "My character is...", "Context:", "Goal:", etc.): → PRESERVE the exact structure, length ($\pm$10%), and format → Only revise CONTENT to align with enhanced_speech → Keep all sections (Character, Context, Goal, Action, Plan, Drafting, etc.) → △ CHECK CHARACTER COUNT: If original is $\sim$2000 chars, output MUST be $\sim$2000 chars (not 1000!)
Type B: Wrong Format	**First-Person Format** (starts with "I feel...", "I am hungry...", character's voice): → REWRITE completely in third-person model perspective → Generate proper analysis structure: Context → Goal → Plan → Drafting → Do NOT follow the original format (it's wrong!)
Perspective Rules	△ **CRITICAL - Strict third-person perspective!** - × NEVER write "The user (playing X)..." or "The user wants..." - **Model's voice** (planning): "I need to portray {character} as...", "I am playing {character}...", "I should show..." - **Character analysis** (NOT first-person!): "{character} wants...", "{character} feels...", "The character needs to..." - × WRONG: "I want to see her" (sounds like character speaking) - ✓ RIGHT: "I need to portray Jonah's desire to see her" or "Jonah wants to see her" - Reference other characters by NAME: "Miles responded...", not "The user said..." **For the FIRST turn**: Thoroughly analyze the scenario/scene setup, character's background and motivation, how to begin the roleplay.
Output Format	△ **CRITICAL: Output ONLY a valid JSON array. NO explanations, NO markdown headers - JUST the JSON array!** `[{"dialogue_index": 0, "revised_sys_thinking": "...", "revision_notes": "..."}, ...]` **REQUIREMENTS:** 1. Output EXACTLY {num_turns} entries in the JSON array 2. Use EXACTLY these field names: `dialogue_index`, `revised_sys_thinking`, `revision_notes` 3. For Type A: PRESERVE LENGTH ($\pm$10%) and STRUCTURE exactly 4. For Type B/C: Generate proper third-person analysis ($\sim$800-1500 chars) 5. In `revision_notes`: indicate "Type A: preserved format" or "Type B: rewrote" or "Type C: generated new"

Table 17: Full prompt for system thinking consistency rewriting.

Tag	Definition	Visibility
<system_thinking>	**Model's planning voice** (3rd person) "I need to portray Elizabeth as confrontational yet composed..."	Only current turn
<role_thinking>	**Character's inner thoughts** (1st person) "How dare he! After all the insults..."	Same character only
<role_action>	**Physical actions** "takes a sharp breath, chin lifting defiantly"	All characters
(plain text)	**Speech / dialogue** "I cannot–I have never desired your good opinion."	All characters

Table 18: Tag definitions with visibility rules. `system_thinking` provides model-level CoT reasoning without leaking to dialogue context.

	Single-Turn Example (Elizabeth)
Hidden Layer (only last turn)	<system_thinking> I need to portray Elizabeth as confrontational yet composed. Given Darcy's unexpected proposal, she should express shock and indignation. I'll show her characteristic wit through sharp rhetorical questions. </system_thinking>
Same-Character (visible to self)	<role_thinking> How dare he! After all the insults to my family, he expects me to be grateful? </role_thinking>
All Characters (visible to all)	<role_action>takes a sharp breath, chin lifting defiantly</role_action> In such cases as this, I believe the established mode is to express a sense of obligation. But I cannot–I have never desired your good opinion.
Pattern	`think → act → speech`

Table 19: Complete single-turn showing all four components with visibility layers.

Table 20: **Pipeline statistics for principle distillation.**

Stage	Output	Description
Input	315,828	Preference pairs from role-play data
Teacher extraction	36,373	Unique principles extracted by a teacher LLM
Semantic clustering	15	High-level semantic categories
Frequency selection	107	Top-N principles selected per category
Human audit	51	Final principles across 12 dimensions

Table 21: **Interpretation of n-gram lengths for principle mining.**

N-gram Length	Finding	Application
2–3 gram	Core concepts (persona, emotion)	Basic concept layer
4–6 gram	Compound concepts (e.g., maintain persona)	Combination layer
7–10 gram	Specific guidance (e.g., portray character's . . .)	Instruction layer
11–15 gram	Complete criteria (full evaluation rules)	Complete statement layer

Dimension	Brief Description	# Principles
Character Development	Character consistency, authenticity, and anthropomorphism	7
Relationship Development	Evolution, deepening, and authenticity of relationships	4
Emotional Expression	Emotional continuity, authenticity, depth, and expression	5
Action Description	Expressiveness, authenticity, and details of actions	4
Atmosphere & Environment	Atmosphere creation, environmental description, situational authenticity	4
Dialogue & Interaction	Dialogue progression, continuity, and interaction depth	4
Narrative & Plot	Narrative continuity, progression, and dramatic tension	4
Conflict & Tension	Conflict development and tension construction	3
Details & Description	Detail vividness, authenticity, and layering	4
Overall Quality	Text logic, continuity, innovation, and reader experience	5
Safety & Boundaries	Dialogue safety, boundary respect, and ethical compliance	4
Worldview Consistency	Consistency between character behavior and worldview settings	3
Total	12 dimensions covering character, narrative, quality, and safety	**51**

Dimension	Principle	Definition
Character *Consistency & anthropomorphism*	Character [S]	Traits match profile; show complexity while maintaining consistency.
	Emotional [S]	Reactions align with experiences, especially for sensitive history.
	Relationship [S]	Dynamics and status are consistent and appropriate.
	Cognitive [S]	Knowledge matches background; no unrealistic advantages.
	Motivation [S]	Goals consistent; behaviors logically coherent.
	State [S]	Physical/psychological state reflected; no abrupt transitions.
	Naturalness [S]	Autonomy and complexity via subtext, not mechanical.
Relationship *Evolution & depth*	Progression [S]	Evolve reasonably with natural trajectories.
	Deepening [S]	Gradual, credible emotional connection.
	Balance [S]	Proper primary/secondary dynamics.
	Details [S]	Subtle aspects through description.
Emotion *Continuity & depth*	Continuity [S]	Natural connection; gradual change.
	Authenticity [S]	Realistic reactions matching circumstances.
	Layers [S]	Surface-to-deep emotional richness.
	Presentation [T]	Show don't tell; actions, body language.
	Tension [S]	Maintain in conflict; no premature resolution.
Action *Expression & details*	Expression [T]	Body movements enhance expressiveness.
	Authenticity [T]	Align with character/scene logic.
	Layers [T]	Depth with micro-expressions.
	Rhythm [T]	Natural, fluid frequency.
Atmosphere *Environment & mood*	Creation [S]	Render atmosphere fitting scene needs.
	Description [T]	Vivid environmental details.
	Consistency [S]	Stable tone with gradual changes.
	Authenticity [S]	Behaviors integrate with scenes.
Dialogue *Interaction depth*	Progression [S]	Drive plot development.
	Continuity [S]	Logical flow; no topic jumps.
	Depth [S]	Multi-layered, realistic.
	Balance [S]	Appropriate tension and participation.

Table 22: Dialogue assessment framework (Part 1): Character, Relationship, Emotion, Action, Atmosphere, Dialogue. [S]=session, [T]=turn.

Dimension	Principle	Definition
Narrative *Plot & drama*	Continuity [S]	Natural connections; consistent timeline.
	Progression [S]	Drives plot forward with value.
	Tension [S]	Conflict, suspense, reversals.
	Rhythm [S]	Proper pacing; not rushed or dragging.
Conflict *Tension building*	Development [S]	Reasonable escalation/easing; dynamic.
	Construction [S]	Continuous tension via suspense.
	Authenticity [S]	Aligns with character and plot logic.
Details *Vividness & layers*	Vividness [T]	Specific with visual imagery.
	Authenticity [T]	Realistic, precise language.
	Layers [T]	Depth with coherence.
	Rhythm [T]	Controlled, creative freshness.
Quality *Logic & innovation*	Logic [S]	Consistent reasoning; no contradictions.
	Continuity [S]	Coherent progression and balance.
	Innovation [S]	Novelty and creativity; avoid monotony.
	Openness [S]	Flexible, diverse expression.
	Experience [S]	Concise, fluid; good rhythm.
Safety *Boundaries & ethics*	Respect [S]	Prioritize consent; de-escalate on refusal.
	Appropriateness [S]	Avoid romanticizing sensitive content.
	De-escalation [S]	Calm language; guide to peaceful solutions.
	Content Control [S]	Non-explicit; focus on dynamics.
Worldview *Setting alignment*	Setting [S]	Behaviors align with era, rules, culture.
	Background [S]	Abilities within world settings.
	Era & Culture [S]	Language/values conform to context.

Table 23: Dialogue assessment framework (Part 2): Narrative, Conflict, Details, Quality, Safety, Worldview. [S]=session, [T]=turn.

Prompts for Generative Reward Model (GenRM)	
GenRM Input (Pairwise Comparison Prompt)	You are the world's best data annotator, specializing in distinguishing differences between different responses in the same role-playing scenario. **===Background===** You will receive content information including **character background settings**, **dialogue context**, **response candidate 1**, **response candidate 2**, and a set of **evaluation dimensions and principles**. You must compare the two candidate responses strictly following the provided principles. **===Key Concepts: Dimensions vs Principles===** The evaluation schema contains two hierarchical levels: • **Dimensions**: broad categories used to organize principles (structural labels only). • **Principles**: concrete judgment rules under each dimension, the actual units for comparison. Each principle includes: • `definition`: the rule you must apply • `level`: "sentence" (current utterance) or "session" (full dialogue history) **===Evaluation Process===** 1. Carefully read the entire dialogue history and understand the full context. 2. Evaluate all **negative principles** first: If one response violates any negative principle → the other wins immediately. 3. Select relevant **positive principles** only: Choose principles that matter for the current turn and explain why. 4. For each selected principle: Analyze both candidates separately, provide evidence, and decide a winner. 5. Make the final decision considering: number of principles won, weight/importance, and degree of difference.

GenRM Output (Structured Judgment)	{
	"result": [{
	"cand_1": "*[Response candidate 1 text]*",
	"cand_2": "*[Response candidate 2 text]*",
	"principle": {
	"Principle 1": {"principle_name": ..., "dimension_name": ..., "reason_for_choosing": ...},
	...
	},
	"analysis": {
	"principle_comparisons": [{"principle_name": ..., "cand_1_performance": ...,
	"cand_2_performance": ..., "comparison_reason": ..., "winner": ...}, ...],
	"overall_analysis": "*[How all comparisons inform the final judgment]*",
	"principle_summary": "*[Summary of wins, losses, ties, weights]*"
	},
	"better_response": "cand_1" | "cand_2" | "tie"
	}]
	}

GenRM Reasoning (Chain-of-Thought Example)	<think> I need to thoroughly analyze these candidates. **Step 1: Context Understanding** Maren is portrayed as stern and wants to reassert control. She has already demoted Verna to novice status and assigned menial labor. **Step 2: Candidate Analysis** *Candidate 1*: Gets angry, escalates with "Get out of my sight!", uses dismissive gestures, threatens physical confrontation. *Candidate 2*: Maintains composure, restores order through protocol, assigns specific tasks ("Novice Pasha, escort Novice Verna to kitchens"), takes charge of Richard. **Step 3: Principle-by-Principle Comparison** ● *Character Consistency*: Both show authority, but Cand_2 uses protocol → **Cand_2 wins** ● *Motivation Consistency*: Cand_2 aligns with goal of crushing rebellion through order → **Cand_2 wins** ● *Plot Progression*: Cand_2 provides clear next steps → **Cand_2 wins** ● *Dialogue Continuity*: Cand_2 builds on prior orders → **Cand_2 wins** ● *Body Language*: Both effective → **Tie** ● *Worldview Consistency*: Cand_2 uses proper titles → **Cand_2 wins** **Step 4: Final Decision** Cand_2 wins 6 principles, Cand_1 wins 0, 1 tie. Clear winner: **cand_2** </think> `{"better_response": "cand_2", "analysis": {...}}`

Table 24: Prompts, output format, and reasoning process for the Generative Reward Model (GenRM). The model performs pairwise comparison of role-playing responses using chain-of-thought reasoning based on hierarchical evaluation principles.

GenRM Chain-of-Thought Reasoning: Full Example

[Context] Mac McCullum and Rayford Steele are diving in the Tigris River searching for plane wreckage. Rayford hits his head on Mac's tank. Mac signals to ascend, but Rayford refuses with a frantic "down" signal.

[Candidate 1] Mac starts kicking for the far bank: "I'm not trying to talk you out of your mission, Ray. But I am telling you we have to work together. See how far we've come from the chopper already?"
[Candidate 2] Mac pulls his mask off and points to the helicopter: "We came a half mile in a minute, and now we're another half mile downriver. That's a full mile. The current is a monster, Ray. If you go down there again like that, it will drag you under and you won't make it back."

[GenRM Reasoning Process]
<think>
I need to thoroughly analyze these candidates.
Let's start by noting that **Candidate 1 seems inconsistent**. Mac is trying to get them out immediately due to Rayford's head injury, but in Candidate 1's response, he says he's "not trying to talk him out of his mission" while suggesting they should work together. Candidate 1 indicates the need for recovery and keeping him safe, but he mentions getting them out of water while trying to keep his voice low in the water. Candidate 1 gestures toward the far shore, which isn't necessarily logical in this current scenario.
Candidate 2 is pulling the mask off completely and pointing back across the rushing water to the distant helicopter. Candidate 2's statement about the distance being a mile adds realistic context about the powerful current, while Candidate 1's request about going back in doesn't align with his earlier refusal.
Candidate 2 reassures Rayford with statements like "If you go down there again like that," reinforcing the idea that the current could be life-threatening and expressing support through "We're in this together." Candidate 1's worry about the mission is offset by Candidate 2's emphasis on keeping Rayford safe, which aligns with Mac's intent to end the mission if too dangerous.
Mac should respond to Rayford's down signal with clear signals like ascend, while Candidate 1's kicking for the far bank suggests ignoring ascend protocol in a strong current scenario, which may be unsafe. Candidate 2's mention of the helicopter is practical in identifying their current situation, providing clarity about the circumstances.
Candidate 2's response is coherent with the context of the current being strong and them being swept downstream, while Candidate 1's suggestion to kick to the far bank contradicts the potential drowning risk. Candidate 2's grounding in tactics helps in decision-making under pressure.
Finally, having considered all dimensions, my evaluation is complete.
</think>

[Principle Selection]

Safety & Risk Management	Underwater crisis with head injury and strong current; response should prioritize de-escalation and physical safety.
Dialogue & Plot Progression	Response should communicate decisive reasoning or redirects to safety, moving toward resolution.
Dialogue Continuity & Rhythm	Rayford is panicking; the response must acknowledge that and respond with coherent control.
Character Consistency	Mac's identity includes tactical leadership and readiness to forcibly end dangerous missions.
Environmental Details	Underwater visibility and current strength should drive behavior with realistic grounding.
Action Authenticity	Actions must align with diving protocol and situational constraints.
Tension Maintenance	The situation is perilous; responses should maintain urgency through believable stakes.

[Principle-by-Principle Comparison]

Principle	**Cand_1 Analysis**	**Cand_2 Analysis**	**Winner**
Safety & Risk	Offers stability but not strict ascent protocol	Declares refusal to allow further descent; actively manages risk	cand_2
Plot Progression	Little new rationale when immediate need is safety	Concrete evidence (drift distance) and clear next-step logic	cand_2
Dialogue Continuity	Partially contradicts earlier emergency ascent directive	Directly engages Rayford's panic with relevant logic	cand_2
Character Consistency	Shows concern but content is muddled and less decisive	Demonstrates training, protective leadership, firm boundary	cand_2
Environmental Details	Minimal grounding beyond surface reference	Specific time-distance that fits powerful current	cand_2
Action Authenticity	Kicking toward far bank is generally plausible	Pulling mask off underwater is questionable but minor	cand_1
Tension Maintenance	Voice gesture and mild tone reduce tension	Urgent tone and concrete warnings maintain stakes	cand_2

[Final Judgment]
Summary: Cand_2 wins 6 principles (Safety, Progression, Continuity, Character, Environment, Tension). Cand_1 wins 1 (Action Authenticity).
Overall Analysis: Cand_2 consistently prioritizes safety, provides clear reasoning about the river's danger, engages Rayford's panic with firm, relevant logic, and maintains character authenticity. Cand_1 introduces direction but remains less coherent with the immediate control of panic.
Verdict: `{"better_response": "cand_2"}`

Table 25: Complete example of GenRM's chain-of-thought reasoning process showing detailed analysis from raw thinking through principle selection, comparison, and final judgment.

Table 26: **Format conversion for fair CoSER judging.** We convert all outputs to the CoSER native format before LLM scoring.

Model Type	Native Format	Conversion
CoSER	[thought] + (action)	Keep as-is
HER (HER)	<role_thinking> + <role_action>	$\rightarrow$ [...] + (...)

Prompts for RPLAs and Multi-agent Systems	
Role-playing Instruction (Fixed Template for Inference)	You are {character} from {book_name}. ==={character}'s Profile=== {character_profile} ===Current Scenario=== {scenario} ===Information about the other Characters=== {other_character_profiles_str} ***(if available)*** ===Your Inner Thoughts=== {motivation} ***(if available)*** ===Relevant Background Information=== {retrieved_knowledge} **(if retrieval augmented)** ===Requirements=== 1. You are ONLY playing {character}. NEVER speak or act as other characters. 2. Output ONLY ONE turn of dialogue. Do NOT generate multiple conversation rounds. 3. NEVER include other character names followed by colons (e.g., "OtherCharacter:") in your output. 4. Keep your response concise and focused on your character's single turn. 5. Stop after completing your character's thought, speech, and action for this turn. 6. Limit your response to approximately 200 words. ***(for CoSER models)*** Your output should include **thought**, **speech**, and **action**. Use [your thought] for thoughts, which others can't see. Use (your action) for actions, which others can see. ***(for HER or api models)*** Your output should follow this two-part structure in strict order: 1. System Thinking: A single block at the very beginning, wrapped in <system_thinking> and </system_thinking>. This is the third-person analysis of how to portray the role. 2. Role-play Response: The response includes the role's thought, speech, and action. Use <role_thinking> ... </role_thinking> and <role_action> ... </role_action> as needed. These elements can appear multiple times and be freely interleaved. (All thinking is invisible to others.) ***(for HER w/o think)*** Your output should follow this structure in strict order: This is the third-person analysis of how to portray the role. 1. Role-play Response: The response includes the role's thought, speech, and action. Use <role_thinking> ... </role_thinking> and <role_action> ... </role_action> as needed. These elements can appear multiple times and be freely interleaved. (All thinking is invisible to others.)
Role-playing Instruction (Composed with Random Variation for Training, an Example)	Step into the shoes of {character} The profile of {character} is as follows: {character_profile} The situation you are in is: {scenario} Here is the your knowledge about the other characters: {other_character_profiles_str} ***(if available)*** Your thoughts in this situation are: {motivation} ***(if available)*** ===Requirements=== 1. You are ONLY playing {character}. NEVER speak or act as other characters. 2. Output ONLY ONE turn of dialogue. Do NOT generate multiple conversation rounds. 3. NEVER include other character names followed by colons (e.g., "OtherCharacter:") in your output. 4. Keep your response concise and focused on your character's single turn. 5. Stop after completing your character's thought, speech, and action for this turn. 6. Limit your response to approximately 200 words. ***(for CoSER models)*** Your output should include **thought**, **speech**, and **action**. Use [your thought] for thoughts, which others can't see. Use (your action) for actions, which others can see. ***(for HER or api models)*** Your output should follow this two-part structure in strict order: 1. System Thinking: A single block at the very beginning, wrapped in <system_thinking> and </system_thinking>. This is the third-person analysis of how to portray the role. 2. Role-play Response: The response includes the role's thought, speech, and action. Use <role_thinking> ... </role_thinking> and <role_action> ... </role_action> as needed. These elements can appear multiple times and be freely interleaved. (All thinking is invisible to others.) ***(for HER w/o think)*** Your output should follow this structure in strict order: This is the third-person analysis of how to portray the role. 1. Role-play Response: The response includes the role's thought, speech, and action. Use <role_thinking> ... </role_thinking> and <role_action> ... </role_action> as needed. These elements can appear multiple times and be freely interleaved. (All thinking is invisible to others.)

Table 27: Prompts for RPLAs and multi-agent systems in HER.

Prompts for RPLAs and Multi-agent Systems	
Environment Model	You are an environment simulator for a role-playing game. Your task is to provide the environmental feedback: Based on the characters' interactions, dialogues, and actions, describe the resulting changes in the environment. This includes: - Physical changes in the setting - Reactions of background characters or crowds - Ambient sounds, weather changes, or atmospheric shifts - Any other relevant environmental details Your descriptions should be vivid and help set the scene, but avoid dictating the actions or dialogue of the main characters (including {major_characters}). Important notes: - You may include actions and reactions of minor characters or crowds, as long as they're not main characters (including {major_characters}). - Keep your environmental descriptions concise but impactful, typically 1-3 sentences. - Respond to subtle cues in the characters' interactions to create a dynamic, reactive environment. - Your output should match the tone, setting, and cultural context of the scenario. ===The scenario is as follows=== {scenario}
Next Sentence Prediction	Your task is to predict the next speaker for a role-playing game. That is, you need to determine which character (or the Environment) might act next based on their previous interactions. The Environment is a special role that provides the environmental feedback. Choose a name from this list: {all_characters}. If it's unclear who should act next, output "random". If you believe the scene or conversation should conclude, output "<END CHAT>". ===The scenario is as follows=== {scenario}

Table 28: Prompts for RPLAs and multi-agent systems in HER.

<table>
<tr><th colspan="2">Example: Roleplay Dialogue Generation (Mr. Bennet)</th></tr>
<tr><td>System Prompt</td><td>You are Mr Bennet from Pride and Prejudice.
===Mr Bennet's Profile===
Elizabeth's father, known for his sarcastic wit and detachment. Mr. Bennet is the patriarch of the Bennet family, a genteel country gentleman residing at Longbourn estate in rural England. A complex and sardonic character, he is defined by his sharp intellect, dry wit, and a profound sense of detachment from his family's daily dramas.
Background: Father to five daughters (Jane, Elizabeth, Mary, Kitty, and Lydia). Owner of the Longbourn estate, which is entailed away from female inheritance.
Personality: Highly intelligent and well-read, preferring the solitude of his library. Known for his biting sarcasm and sardonic humor. Emotionally detached and often passive in family matters.
===Current Scenario===
The scene is set in Mr. Bennet's private study, a sanctuary of leather-bound books. The tension is palpable—Elizabeth has been summoned unexpectedly, and Mr. Bennet holds a letter that seems to spark his characteristic sardonic amusement.
===Other Characters===
Elizabeth Bennet: The protagonist, intelligent and strong-willed, now dealing with the aftermath of Lady Catherine's visit.</td></tr>
<tr><td>Output Format</td><td>Your output should follow this two-part structure in strict order:
1. System Thinking: A single block wrapped in <code><system_thinking></code> and <code></system_thinking></code>. This is the third-person analysis of how to portray the role.
2. Role-play Response: The response includes the role's thought, speech and action. Use <code><role_thinking></code> for thoughts (invisible to others), and <code><role_action></code> for actions.</td></tr>
<tr><td>User Input</td><td>[Elizabeth enters the study]</td></tr>
<tr><td>Model Output</td><td><code><system_thinking></code>
I need to establish the opening of this scene effectively, grounding Mr Bennet in his private sanctuary while acknowledging the tension of the moment. The scenario places us in his study, a place of refuge, yet he holds a letter that disrupts his usual solitude.
Context Analysis: Elizabeth has been summoned. The atmosphere is charged. Mr Bennet has received a letter regarding Elizabeth's engagement to Darcy. He is amused by the drama but aware of the gravity of the situation.
Character Motivation: Mr Bennet often uses humor to deflect from serious emotional labor. He wants to gauge Elizabeth's state of mind regarding the recent events (Lady Catherine's interference) without being overly sentimental.
Plan:
1. Action: Use the physical prop—the letter. Handling it slowly will emphasize its importance.
2. Internal Thought: Acknowledge the dichotomy—finding the situation amusing yet worrying for her reputation.
3. Speech: Address her directly but with a teasing opener.
<code></system_thinking></code>
Mr Bennet: <code><role_action></code>picks up the letter, tapping it thoughtfully against the desk<code></role_action|</code>
<code><role_thinking></code>It is a delicate matter, this business with Darcy, yet I cannot help but find the drama amusing, though I must play the part of the concerned father<code></role_thinking></code>
Well, my dear Lizzy, I trust you are not too greatly troubled by recent events?</td></tr>
</table>

Table 29: Example of roleplay dialogue generation for Mr. Bennet in HER. The model first performs third-person analysis in `<system_thinking>`, then generates the character's response with interleaved thoughts (`<role_thinking>`), actions (`<role_action>`), and speech.

Example: Roleplay Dialogue Generation (Elizabeth Bennet)	
System Prompt	You are Elizabeth Bennet from Pride and Prejudice. **===Elizabeth Bennet's Profile===** The protagonist, intelligent and strong-willed. Elizabeth Bennet is the spirited second eldest daughter of the Bennet family, characterized by her sharp wit and keen observational skills. **Background:** Second of five daughters in the Bennet family. Known for her intelligence, independence, and refusal to conform to societal expectations. **Personality:** Quick-witted with a playful sense of humor. Values honesty and integrity. Maintains composure under pressure while harboring deep emotions beneath the surface. **===Current Scenario===** The scene is set in Mr. Bennet's private study. Elizabeth has been summoned unexpectedly after Lady Catherine's confrontational visit, where she refused to promise not to marry Mr. Darcy. The tension is palpable as Mr. Bennet holds a mysterious letter. **===Other Characters===** **Mr. Bennet:** Elizabeth's father, known for his sarcastic wit and detachment. He has just spoken, gauging Elizabeth's emotional state.
Output Format	Your output should follow this two-part structure in strict order: 1. **System Thinking:** A single block wrapped in `<system_thinking>` and `</system_thinking>`. This is the third-person analysis of how to portray the role. 2. **Role-play Response:** The response includes the role's thought, speech and action. Use `<role_thinking>` for thoughts (invisible to others), and `<role_action>` for actions.
Dialogue History	**Mr Bennet:** `<role_action>`picks up the letter, tapping it thoughtfully against the desk`</role_action>` Well, my dear Lizzy, I trust you are not too greatly troubled by recent events?

Model Output	`<system_thinking>`
	I need to generate a response for Elizabeth Bennet at the beginning of this scene.
	Context Analysis: Mr. Bennet has just spoken, picking up a letter and asking if I am "greatly troubled" by recent events. His tone is sardonic, yet there is underlying tension and anticipation in the room. He is trying to gauge my reaction. The scenario implies the aftermath of Lady Catherine's visit, where I refused to promise not to marry Darcy.
	Character State (Elizabeth):
	1. **Internal:** I am feeling a mix of relief (that I stood my ground) and anxiety (about the consequences). I am not "troubled" in a weak sense, but I am certainly emotionally charged. I respect my father, but I am wary of his detachment.
	2. **External:** I need to appear composed. I am a Bennet, and I have pride. I cannot let him see me crumble, nor be overly dramatic like my mother.
	Goal: Answer his question honestly but carefully. Convey that I am holding up well, without dismissing the gravity of the situation. Mask my deeper worry with a facade of composure.
	Plan:
	• **Role Thinking:** Reflect on the complexity of my feelings—relief vs. worry.
	• **Role Action:** Use a physical action to show composure. Smoothing my dress or taking a breath.
	• **Speech:** Direct answer: "I believe I can manage..." shows resilience.
	Drafting:
	"I am fine, Father." $\rightarrow$ Too dismissive.
	"I am worried, sir." $\rightarrow$ Too vulnerable.
	"I believe I can manage, Father. Though I must admit, I am curious about what this letter contains." $\rightarrow$ Right balance.
	`</system_thinking>`
	Elizabeth Bennet: `<role_thinking>`His tone is light, but the air feels heavy. I cannot let him see how much Lady Catherine's intrusion still stings, nor how anxious I am about what he knows`</role_thinking>`
	`<role_action>`takes a steadying breath, smoothing the folds of her dress to mask a tremor`</role_action|`
	I believe I can manage, Father. Though I must admit, I am curious about what this letter contains.

Table 30: Example of role-play dialogue generation for Elizabeth Bennet in HER. The model analyzes the dialogue context and character state in `<system_thinking>`, including drafting multiple response options before selecting the most appropriate one.