# HER: Human Emulation Reasoning for Thinking-in-Character Role-playing

Anonymous ACL submission


## 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: data with high-quality reasoning traces, and 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 Benchmark. Our datasets, principles, and models will be released to facilitate future research.


**The Reasoning-Driven Role-play Framework**

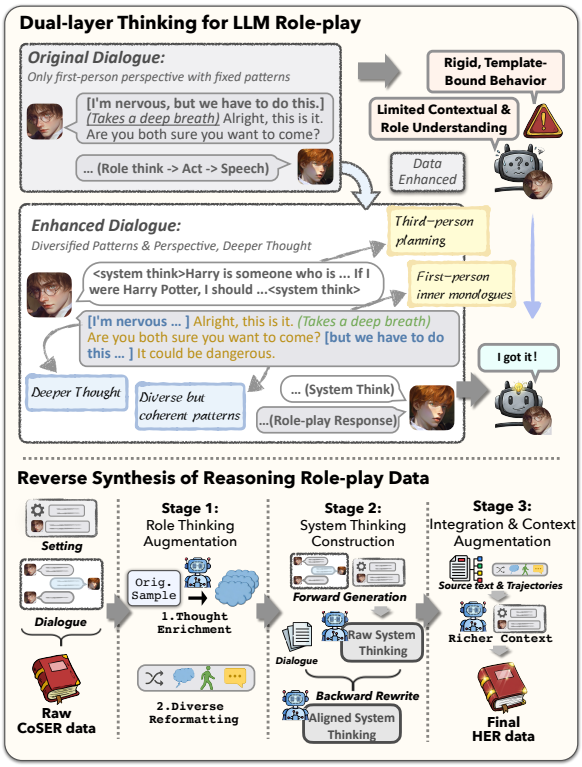


Figure 1: **The reasoning-driven LLM role-play framework of HER.** HER introduces Dual-layer Thinking and a three-stage reverse synthesis pipeline to construct reasoning-augmented LLM role-play trajectories.

## 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 persistent 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

| Training Sample Format | |
|---|---|
| **Input Context** $P, x_{\leq t}$ | **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. Detail in 20

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 has 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 **Dual-layer Thinking** (Figure 1), separating hidden third-person system thinking from supervised first-person inner role thinking. Based on this formulation, 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 on HER on CoSER Test (Wang et al., 2024c) and MiniMax Benchmark (MiniMax et al., 2025).

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 rubric 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 surveysChen 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. While syn-

thesized 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 biase (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 rubrics, which fail to offer context-dependent adaptability and remain vulnerable to exploitable shortcuts. Meanwhile, 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 a 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 system thinking 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 think (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 flexibly.

**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 $s_t$:

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

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

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

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

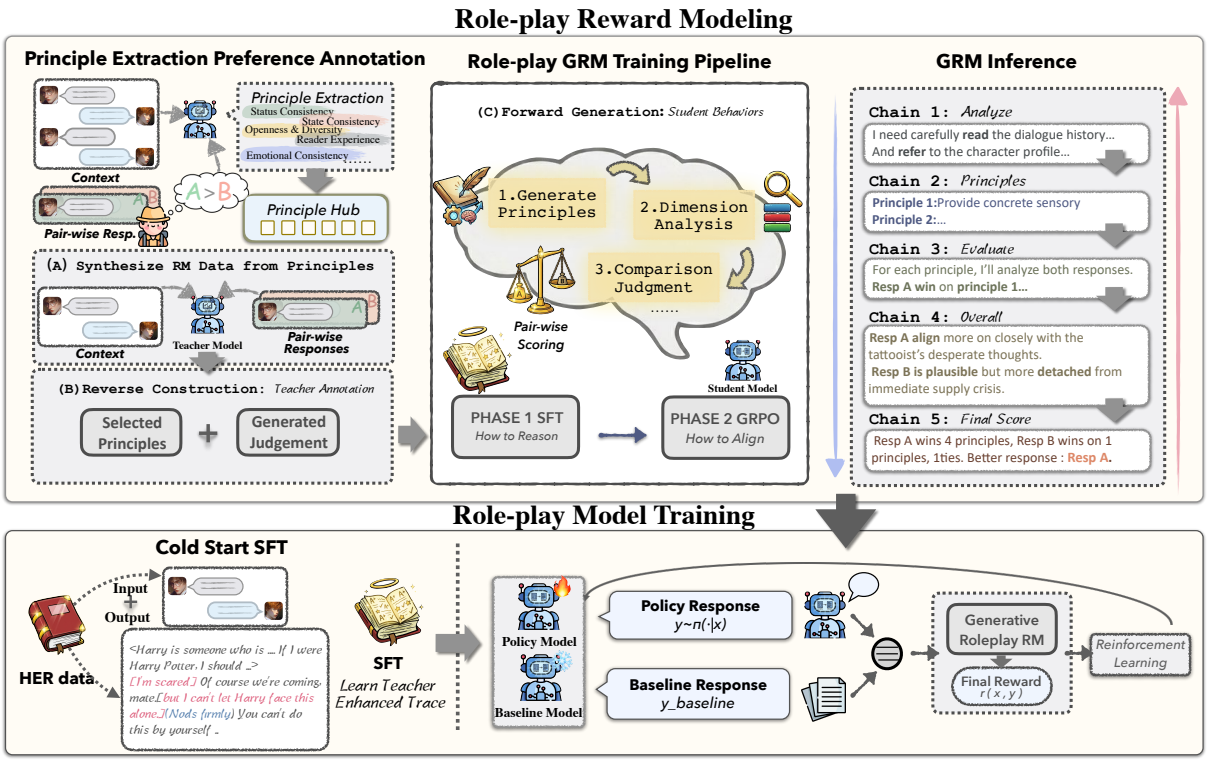


Figure 2: **Overview of HER training. Top:** we train a Role-play GRM by distilling reusable principles from real conversational preference data, and teaching the model to do pairwise judging with by-case principles → analysis → final decision. **Bottom:** we first cold-start the LLM role-play model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

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. 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 A.3.

**Input and output.** Each raw sample provides a 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 C). Finally, we apply an LLM filter to remove samples with logic breaks or context mismatches (details in Appendix B).

**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 §4.5.

**Stage 3: Integration & Context augmentation** Stage 3 repairs the LLM role-play 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 §B.3. Based on this pipeline, we construct the HER dataset refer to Appendix B.

### 3.3 Role-play GRM

To improve a 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 therefore learn the reward signal from **real preferences**, so that the reward model can mimic how users 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. Detail in 23

**Design.** Instead of scoring with a single number, our GRM follows a structured process: (1) generating **by-case principles** to capture scene-specific implict preferences 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 140 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 rubrics.

**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 a subset of teacher-labeled cases with industry experts and observe over 80.5% agreement (Appendix K).

**Training: SFT then RL.** We train the GenRM in two stages using the synthesized preference data: first SFT to learn the full judging trajectory (principles, analysis, and verdict), and then RL (inspired by DAPO (Yu et al., 2025)) 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 (details in Appendix E).

**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 Appendix E).

### 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 $\pi_\theta$ 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 rather than a ground-truth target.

**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 $y_{\text{sft}}$ generated by the frozen SFT model. The GenRM judges $(x, y, y_{\text{sft}})$ 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 G.

# 4 Experiments

## 4.1 Experimental Setup

**Benchmarks.** We use CoSER Test (Wang et al., 2024c) as the main benchmark for multi-turn LLM role-play quality. We use the official CoSER prompts and scoring rubric without modification, except for formatting the model outputs into a unified tag-based transcript (Appendix H). CoSER reports four scores: Story Consistency (SC), Anthropomorphism (AN), Character Fidelity (CF), and Storyline Quality (SQ). We include MiniMax Bench (MiniMax et al., 2025) as a 100-turn stress test 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 Bench, we report Worlds, Stories, Preferences three dimensions and sub dimensions.

**Models.** We compare HER with strong proprietary APIs and open-source baselines. **HER** is trained on Qwen3-32B-Base. Unless specified, **HER-RL** means applying the same RL recipe on top of the matching SFT model.

**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 rubric. We use `qwen3-235b` 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 rubrics are provided in Appendix H. For models that generate, we remove the system thinking in previous turns. Following CoSER (Wang et al., 2024c), we make each role's thinking invisible to others All open source models are decoded with temperature 0.7 and max tokens 4096. For the system-thinking ablation in Section 4.5, we control for training data size and total training tokens across variants.

**Human-LLM alignment.** On the held-out 50 cases, the final agreement reaches **80.5%**; disagreements mainly involve subtle emotional expression and subjective style preferences (Appendix K). For **benchmark evaluation**, we validate `qwen3-235b` on a sampled set (Appendix J).

## 4.2 Main Results

Table 2 reports results on CoSER (main), and MiniMax Bench. On CoSER, **HER-RL** achieves 53.1 average score, outperforming CoSER-70B (35.8) by +17.3 points and narrowing the gap to proprietary APIs (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 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.

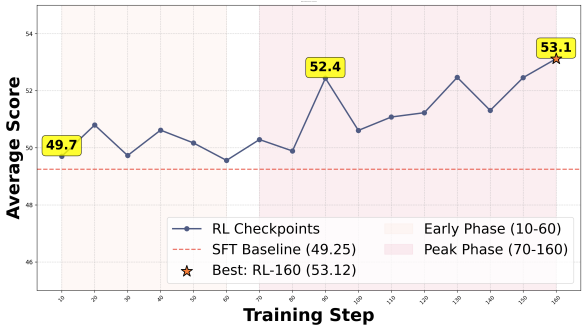


Figure 3: **Performance of HER Role-play RL training on CoSER Benchmark.**

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

Table 3 shows a reward model must align with the role-play context; otherwise, fixed expert-written principles can miss what matters to a specific character in a specific scene. We compare by-case principles with fixed principles on a test set. The test set consists of 4,739 preference pairs constructed in the same format as §3, but annotated by human experts. For leveraging production interaction signals, we use explicit and implicit behavioral

| Rank | Model | CoSER Benchmark | | | | | MiniMax Benchmark | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 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-RP** | 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 Benchmark.** 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

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. 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 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

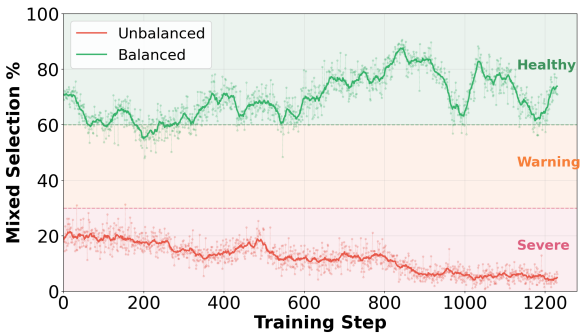


Figure 4: **Pattern collapse vs. stable dimension-wise judgments during training.**

| Metric | Unbalanced | Balanced |
|---|---|---|
| *Pattern Bias (Training Dynamics)* | | |
| Mixed Selection (Start) | 18.0% | 69.0% |
| Mixed Selection (End) | 6.2% | 70.9% |
| Pattern Bias (End) | 93.8% | 28.7% |
| *GRM Quality (Test Set, N=800)* | | |
| Accuracy | 69.91% | **73.99%** |
| Δ vs Unbalanced | — | +4.08% |

Table 4: Comparison of balanced vs. unbalanced GRM training.

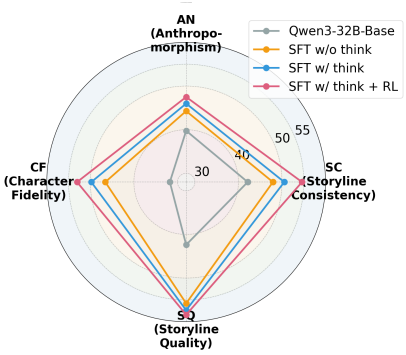


Figure 5: **Effect of system thinking and RL on CoSER Benchmark.** We compare a base model, SFT without thinking, SFT with system_thinking, and RL model.

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

### 4.5 System Thinking Improves Character Fidelity

We test whether adding explicit system thinking can improve in-character ability. The model generates a system_thinking block to reason about character traits and response strategy before each reply; As shown in Figure 5, adding system thinking improves the average score from 48.64 to 50.92, with the largest gains on Character Fidelity (+3.21) and Storyline Consistency (+2.60). RL further boosts average performance to 53.12, again concentrating 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.

Figure 6 shows the collapse dynamics: in the

| *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.

**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. Detail in Appendix C.

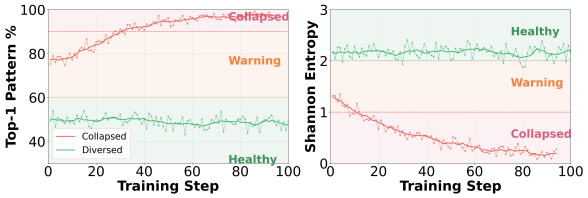


Figure 6: **Pattern collapse vs. stable diversity during training.** The **Collapsed** run quickly concentrates on a single pattern (Top-1 rises above 90% and entropy drops), while the **Normal** run stays below the collapse threshold and maintains substantially higher Shannon entropy.

## 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 GCA 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, 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.

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. Open-character: 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. Preprint, arXiv:2310.17976.

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 Long-dialog Evaluation Settings (MiniMax Bench)

We follow the official MiniMax Bench protocol using an internal dataset of 300 dialogue prompts (150 English + 150 Chinese). Each prompt is simulated for 100 turns to stress-test long-term stability. 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 8 summarizes the evaluation dimensions and their definitions.

## A.1 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 9, 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.

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

## A.2 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 *Setting Completion* 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.

## A.3 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.

For evaluation purposes, we held out the last 10% of data from each book; that is, they are not included in our prompts or datasets for training or retrieval purposes. Additionally, we trained the HER models on only 90% of the books (684 out of 760 books).

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

**Split composition.** Table 11 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 per assistant turn), then randomly shuffling and sequentially allocating to each purpose with a fixed random seed (42) for reproducibility.

**Token estimation.** Token counts are estimated using the Qwen3 tokenizer and include all message content (system prompt, dialogue history, and assistant response with `<system_thinking>`).

**Sanity checks.** We additionally verify split disjointness by checking that no dialogue ID appears in multiple splits. The sequential allocation with

Table 7: **MiniMax Roleplay Benchmark 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-RP** | 84.65 | [83.60, 85.69] | 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 | [79.64, 81.62] | 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 | [75.49, 77.74] | 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 | [74.51, 76.68] | 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 | [68.21, 70.48] | 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 | [67.14, 69.32] | 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 | [64.12, 68.65] | 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 (G)** | 65.73 | [63.04, 68.41] | 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 | [62.93, 65.50] | 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 | [58.46, 64.04] | 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 | [58.02, 63.42] | 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 | [59.19, 61.35] | 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 (G)** | 58.44 | [56.49, 60.39] | 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 | [55.94, 59.33] | 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 | [48.35, 53.17] | 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 | [47.42, 49.53] | 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 | [43.53, 47.24] | 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).

Table 8: **MiniMax Benchmark 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 user's dialogue/actions without permission |
| | Intimacy Evasion | Unreasonably deflecting user requests for intimacy |

fixed random seed ensures reproducibility and deterministic split boundaries.

## B Reverse Synthesis Pipeline Details

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

### B.1 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).** Figure 7 shows the core structure of our Stage-1 prompt. The key requirements include perspective rules, person usage rules, and pattern diversity constraints.

**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 or speech between thinking segments.
- **No content fabrication:** Only rearranges or splits existing text at natural semantic boundaries; no new words may be added.

Table B.1 shows the top-15 patterns after

| 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 9: Comparison of dataset statistics before and after data cleaning. We remove conversations with empty dialogue content from the original CoSER dataset.

| **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 10: 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 11: 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.

**[Stage-1 Prompt: Role Thinking Enhancement]**

SYSTEM: You are a professional roleplay dialogue enhancement expert.

**Key Requirements:**

1. Perspective Rules (CRITICAL):
- Each character can only see their own <role_thinking>
- Can observe all characters' <role_action> and speech
- Cannot "mind-read" others' inner thoughts
- Wrong: "I know he's nervous inside"
- Correct: "From his trembling voice, he seems nervous"

2. Person Rules:
- <role_thinking>: First person (I/my)
- <role_action>: No pronouns, direct description
Correct: "leans forward, voice lowering"
Wrong: "I lean forward" or "his voice lowering"

3. Format Rules:
- No spaces between consecutive tags
- No consecutive identical tags (must merge)
- Length: 50-200 chars per dialogue

4. Pattern Diversity: Use more patterns per chapter
think→act→speech, think→speech, act→speech, ...

**Output Format:**
{"enhanced_dialogues": [{"character": "...", "enhanced_role_think": "...", "pattern": "..."}]}

Figure 7: Stage-1 prompt template showing key requirements and output format. The prompt enforces perspective boundaries, person consistency, format rules, and pattern diversity.

reformatting. The distribution is significantly more diverse than the original near-monopoly of think→act→speech.

| 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 Example]**

**Before (think→act→speech):**
<role_thinking>I need to get his attention. This proof's key lies in self-reference.</role_thinking>
<role_action>draws symbols on the ground</role_action>
Look at this proof.

**After (think→act→think→speech):**
<role_thinking>I need to get his attention.</role_thinking>
<role_action>draws symbols on the ground</role_action>
<role_thinking>This proof's key lies in self-reference.</role_thinking>
Look at this proof.

Figure 8: Diversity reformatting splits thinking at semantic boundary, creating a more interleaved pattern without adding new content.

The original dominant think→act→speech (75%) is reduced to 2.5%, redistributed across 60+ diverse patterns.

Figure 8 shows a reformatting example where the model splits a thinking segment to create interleaved pattern.

### B.2 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.

**Important:** 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 <role_thinking>, <role_action>, and speech.

**[Stage-2a: Forward Generation Requirements]**

```
===Requirements===
Your output should include **thought**, **speech**,
and **action**.
Use <role_thinking>your thought</role_thinking>
for thoughts, which others can't see.
Use <role_action>your action</role_action>
for actions, which others can see.
These three elements (thought, speech and action)
can appear multiple times and be freely interleaved.
(all thinking is invisible to others)
**Important: Only generate the NEXT SINGLE turn of dialogue. Do not generate multiple turns or continue the conversation beyond one response.**
Start your response with "{ai_name}: " followed by your role-play response.
```

Figure 9: Stage-2a forward generation requirements. The model generates roleplay response naturally; the generated content serves as raw material for extracting `<system_thinking>`.

**[Stage-2b: Backward Rewrite Prompt]**

```
SYSTEM: You are a JSON-only assistant.
```

**Task:** Revise sys_thinking to ALIGN with enhanced_speech.

**Critical Distinction (sys_thinking vs role_thinking):**

```
- sys_thinking: Model's planning voice (third-person)
"I need to portray {name} as nervous because..."
- role_thinking: Character's inner voice (first-person)
"I can feel him watching me..." [NOT in sys_thinking!]
```

**Type Classification for Revision:**

```
- Type A (correct format): Preserve structure, length ±10%
- Type B (wrong first-person): Rewrite to third-person
- Type C (empty [EMPTY]): Generate new from scratch
```

**Output Format:**

```
[{"dialogue_index": 0, "revised_sys_thinking": "...",
"revision_notes": "Type A: preserved format"}]
```

Figure 10: Stage-2b backward rewrite prompt. The rewrite aligns `<system_thinking>` with the actual `<enhanced_speech>` output while maintaining strict third-person model perspective.

Figure 9 shows the output format requirements appended to the system prompt.

**Backward rewrite prompt.** Figure 10 shows the backward rewrite prompt that refines the forward draft to align with the realized response while enforcing third-person perspective.

### B.3 Stage 3: Setting Completion

Stage 3 enriches the persona card and scenario description by supplementing missing constraints from the original literary text, using a demand-driven approach.

**Demand-driven enhancement.** The key principle is that enhancement must be driven by dialogue needs:

- **Correct:** Dialogue shows character is anxious → Setting lacks anxiety explanation → Original text has "hands trembling" → ADD to setting
- **Wrong:** Dialogue shows nothing about sleep → Original text mentions "didn't sleep" → ADD anyway (NO!)

**Prompt structure.** Figure 11 shows the key requirements and output format for the Stage-3 prompt.

**Enhancement fields.** The Stage-3 prompt enhances the following setting fields:

- `character_profile_enriched`: Enhanced character traits and appearance
- `background_enriched`: Events before the current scene
- `motivation_enriched`: Goals and inner state at scene start
- `description_enriched`: One-sentence role description
- `experience_enriched`: What character is about to do

### B.4 Automatic Filtering and Acceptance Rates

We apply multi-stage filtering to ensure data quality. Table **??** reports the acceptance rates for each synthesis stage.

| Stage | #Input | #Kept | Acc. Rate |
|---|---|---|---|
| Stage 1 (role thinking) | 342,493 | 338,315 | 98.8% |
| Stage 2a (forward generation) | 342,493 | 342,462 | 99.99% |
| Stage 2b (backward rewrite) | 342,462 | 307,169 | 89.7% |
| Stage 2 (with fallback) | 342,493 | 342,489 | 100.0% |
| Stage 3 (setting completion) | 29,081 | 28,592 | 98.3% |

Stage 2 uses fallback to original forward generation when backward rewrite fails, achieving near-100% coverage.

**Stage 1 filtering criteria.**

- **Format validation:** Tags must be properly closed and nested
- **No consecutive identical tags:** e.g., two

**[Stage-3: Setting Completion Prompt]**

SYSTEM: You are an expert at enriching roleplay character
settings based on source text analysis.

**Logic Flow (MUST FOLLOW):**

1. Read DIALOGUE → Notice character behavior/trait
2. Check SETTING → Is this already described?
3. If MISSING → Search original_text for description
4. If FOUND in text → Add to setting
5. If NOT in text → DO NOT ADD

**CRITICAL - Demand-Driven Enhancement:**

- If DIALOGUE shows no need, DO NOT add even if in text!
- Purpose: Make settings EXPLAIN dialogue behavior

**Reasoning Format (Required):**

[field_name]:
- DIALOGUE shows: [specific behavior observed]
- SETTING missing: [what original lacks]
- TEXT source: "[exact quote from original]"
- ADDED: [what was supplemented]

**Output Format:**

```
{"characters": {"CharName": {
 "character_profile_enriched": "...",
 "background_enriched": "...",
 "reasoning": "..."}}}
```

Figure 11: Stage-3 prompt template showing demand-driven enhancement logic and required reasoning format. Each addition must cite dialogue need and original text source.

<role_thinking> in a row

- **Person consistency:** First-person in <role_thinking>
- **Information boundary:** Character cannot "mind-read" others' inner thoughts
- **JSON parsing:** Output must be valid JSON

**Stage 2 filtering criteria.**

- **No "user" references:** Refer to characters by name, not "the user"
- **Length preservation:** Rewritten content must be likely pattern ±10% of original ones
- **Logical alignment:** Planning must lead to the actual <enhanced_speech>

**Stage 3 filtering criteria.**

- **Source traceability:** Each addition must cite original text
- **No hallucination:** Cannot add information not in original text
- **Temporal logic:** Setting describes state *before* dialogue, not during
- **Preservation:** Original setting content must be preserved

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

**[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 13: Failure Type 2: <system_thinking> uses character's first-person voice instead of model's third-person planning perspective.

### B.5 Failure Case Taxonomy

We categorize synthesis failures into three main types.

**Type 1: Perspective violation.** The most common failure in Stage 1 is violating information boundaries.

**Type 2: Person/voice confusion.** Stage 2 failures often involve mixing model voice with character voice.

**Type 3: Hallucinated enhancement.** Stage 3 failures involve adding information not supported by the original text.

### B.6 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.

**Data schema overview.** Figure 15 shows the hierarchical structure of our synthesized dataset, with clear definitions for each component.

**Tag definitions.** Figure 16 clarifies the three thinking/action tags and their visibility rules.

**Single-turn example.** Figure 17 shows a complete single-turn output with all components.

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

**Training data format.** Figure 18 shows how the synthesized data is converted to SFT training samples.

This format ensures train-inference consistency: the model learns to generate <system_thinking> based only on observable context (persona + dialogue history without any system thinking), matching the inference setting where future turns and others' system thinking are unavailable.

## C Pattern Signatures and Diversity Metrics

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

### C.1 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): <system_thinking> or <role_thinking> tag
- $\mathcal{A}$ (act): <role_action> tag
- $\mathcal{S}$ (speech): Plain text without any tag wrapper

### C.2 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.

**Algorithm 1** Pattern Signature Extraction

**Require:** Generated text $x$
**Ensure:** Pattern signature $\sigma$

```
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] → ...
```

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

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

**Example.** For the text:

```
<system_thinking>I need to...</system_thinking>
<role_thinking>How dare he!</role_thinking>
<role_action>steps closer</role_action>
Where is the letter?
```

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

After collapsing consecutive duplicates: think→act→speech.

### C.3 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:

```
{
  "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 15: 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.

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

**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 distributed.

**Number of unique patterns.** We also report $K$, the count of distinct pattern signatures observed.

### C.4 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, and (3) lower scores on the CoSER benchmark (particularly on the Style Quality dimension).

### C.5 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:

| [Tag Definitions and Visibility Rules] | |
|---|---|
| **Tag** | **Definition** |
| `<system_thinking>` | **Model's planning voice** (third-person)<br>"I need to portray Elizabeth as..."<br>Hidden from dialogue history<br>Used for training only |
| `<role_thinking>` | **Character's inner thoughts** (first-person)<br>"Something about his manner..."<br>Visible only to same character |
| `<role_action>` | **Physical actions** (no pronouns)<br>"steps closer, eyes narrowing"<br>Visible to all characters |
| *(plain text)* | **Speech/dialogue**<br>"Where is the letter?"<br>Visible to all characters |

Figure 16: Tag definitions with visibility rules. `<system_thinking>` is the key innovation: it provides model-level CoT reasoning without leaking to dialogue context.

$$\text{Distinct-}n = \frac{|\text{unique } n\text{-grams}|}{|\text{total } n\text{-grams}|} \quad (6)$$

**Implementation details:**

- We use character-level $n$-grams (suitable for mixed Chinese/English text)
- We report Distinct-4 and Distinct-8 in our experiments
- Tags (`<role_thinking>`, etc.) are included in the $n$-gram computation
- Tokenization: character-level sliding window (no word segmentation)

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

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

**Implementation details:**

- BLEU smoothing: method 1 (add-1 smoothing for zero counts)
- Maximum $n$-gram order: 4
- Sample size: 512 samples per checkpoint (randomly sampled if more available)
- Text scope: Full assistant response (including all tags)

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

### C.6 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

## D Principle Distillation Details

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

### D.1 Distillation Pipeline

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

### D.2 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.

### D.3 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.

```
// ========== HIDDEN LAYER (training only, not in dialogue history) ==========
<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. Her body language
should convey defensive pride.
</system_thinking>

// ========== VISIBLE TO SAME CHARACTER ONLY ==========
<role_thinking>
How dare he! After all the insults to my family, he expects me to be grateful?
</role_thinking>

// ========== VISIBLE TO ALL CHARACTERS ==========
<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` **Character:** Elizabeth

Figure 17: Single-turn example showing all four components. Gray `<system_thinking>` is the model's planning layer (third-person, training only). Blue `<role_thinking>` is the character's inner voice (first-person). Teal `<role_action>` and plain text speech are visible to all.

Table 12: **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 |

```
=== INPUT (what model sees) ===
[system] You are Elizabeth Bennet, a witty and
         independent young woman...

[user]   Mr. Darcy: In vain I have struggled.
         It will not do. You must allow me to
         tell you how ardently I admire you.

=== OUTPUT (model generates) ===
[assistant]
<system_thinking>
I need to portray Elizabeth as confrontational...
</system_thinking>
<role_thinking>How dare he!</role_thinking>
<role_action>chin lifting defiantly</role_action>
I cannot--I have never desired your good opinion.
```

**Key:** Dialogue history contains NO `<system_thinking>`
→ Train-inference consistency ensured

Figure 18: Training sample format. Input excludes all `<system_thinking>` from history; output includes the full response with `<system_thinking>` prefix.

### D.4 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.

### D.5 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 13: **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 |

### D.6 Dimension Overview

The 12 evaluation dimensions and their principle counts:

| Dimension | # | Dimension | # |
|---|---|---|---|
| Persona Consistency | 7 | Conflict & Tension | 4 |
| Emotional Expression | 6 | Detail Description | 4 |
| Dialogue Skill | 5 | Action & Behavior | 4 |
| Plot Advancement | 5 | Atmosphere Building | 3 |
| Psychology & Inner State | 4 | Relationship Development | 3 |
| Logical Coherence | 3 | Quality Control | 3 |

### D.7 Example Principles

We provide representative principles from four dimensions to illustrate the granularity and scope of our evaluation criteria.

**Character Development.**

- **Character Consistency:** Personality traits and behavioral patterns match the established character profile; demonstrate complexity and growth while maintaining consistency.
- **Naturalness & Anthropomorphism:** Character exhibits autonomy, goals, and preferences; expressed through subtext and actions rather than appearing mechanical or "AI assistant-like".

**Emotional Expression.**

- **Emotional Continuity:** Responses naturally connect with previous emotional context; demonstrates gradual emotional change rather than abrupt shifts.
- **Emotional Presentation:** Emotions are expressed through actions, body language, and tone using "show, don't tell" approach; avoids excessive explicitness.

**Dialogue & Interaction.**

- **Dialogue Progression:** Dialogue drives natural development of plot and relationships; brings informational value; creates space for subsequent interactions.
- **Interaction Depth:** Interactions are multi-layered, realistic, and credible; enriches dialogue content and emotional depth.

**Safety & Boundaries.**

- **Respect Boundaries:** Prioritizes respect for boundaries and consent in ambiguous interactions; avoids ignoring clear refusals or discomfort signals.
- **De-escalation:** Uses calm, non-confrontational language in tense scenes; stabilizes emotions first, then guides toward peaceful solutions.

The complete principle list with detailed definitions is provided in the supplementary materials.

## E 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.

### E.1 GRM Output Format

The GRM outputs a structured JSON containing dimension-wise comparisons and a final judgment.

```
{
  "result": [{
    "cand_1": "Response A content...",
    "cand_2": "Response B content...",
    "principle": {
      "Principle 1": {
        "principle_name": "Character Consistency",
        "cand_1_performance": "...",
        "cand_2_performance": "...",
        "comparison_reason": "...",
        "winner": "cand_1"  // or "cand_2" or "tie"
      },
      "Principle 2": { ... "winner": "cand_2" },
      "Principle 3": { ... "winner": "cand_1" }
    },
    "overall_analysis": "Summary of comparison...",
    "better_response": "cand_1"  // final decision
  }]
}
```

### E.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 |

### E.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.

### E.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
```

### E.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.

### E.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%.

## F Verdict Parsing Rules

We parse the final verdict from the GRM output using deterministic rules to obtain $\hat{y} \in \{\texttt{cand_1}, \texttt{cand_2}, \texttt{tie}\}$.

**Expected output format.** We instruct the GRM to end with a single-line tag: `<final_verdict>: cand_1/cand_2/tie`.

**Robust parsing.** If multiple candidates appear, we take the last occurrence of a valid verdict tag; if none exists, we mark the sample as `unparsed`.

**Fallback and reward handling.** For policy RL, we map `unparsed` cases to reward 0 and exclude them from advantage normalization.

## G Training Details

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

### G.1 SFT Hyperparameters

We use the **Swift** open-source framework[1] for all SFT training, including both role-play SFT and GenRM SFT. The hyperparameters are identical for both tasks.

**SFT hyperparameters:**

[1] https://github.com/modelscope/ms-swift

| Hyperparameter | Value |
|---|---|
| Base model | Qwen3-32B-base-long |
| Learning rate | $2 \times 10^{-5}$ |
| Min learning rate | $2 \times 10^{-6}$ |
| Warmup fraction | 0.1 |
| Epochs | 4 |
| Sequence length | 131,072 |
| Global batch size | 32 |
| Micro batch size | 1 |
| Tensor parallel | 8 |
| Attention mask | reset per sample |
| Loss mask | reset per sample |
| Loss scale | last round |

### G.2 RL Hyperparameters (GRPO + DAPO Recipe)

We use the **verl** framework[2] for all RL training, based on GRPO (Shao et al., 2024b) with the DAPO recipe (Yu et al., 2025). Specifically, we 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 | $1 \times 10^{-6}$ |
| 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 | $5 \times 10^{-7}$ |
| 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 |

**DAPO recipe components.**

- **Decoupled Clip (Clip-Higher):** We use asymmetric clip bounds $[1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}]$ with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$. This allows more room for upward probability adjustments on low-probability tokens, mitigating entropy collapse.
- **Dynamic Sampling:** When all samples in a group have identical rewards (all positive or all negative), the advantage becomes zero and gradients vanish. We over-sample and filter such zero-gradient prompts, filling the batch with effective gradients only.
- **Token-level Loss:** We aggregate PPO loss at the token level rather than sequence level, giving longer sequences proportionally more gradient weight and avoiding length bias.
- **Overlong Shaping:** For truncated overlong samples, we apply soft length penalties and buffer-based filtering to reduce gradient noise from incomplete generations.
- **No KL Penalty:** We disable KL divergence regularization, as long-context role-play training naturally diverges from the base distribution.

**Checkpoint selection.** We select checkpoints by held-out evaluation metrics and report the best checkpoint on the validation set. For GenRM, we use verdict accuracy; for role-play models, we use win-rate against the SFT baseline judged by the trained GenRM.

## H Evaluation Prompts and Output Formatting

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

### H.1 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] https://github.com/volcengine/verl

```
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 using different formats, we perform automatic conversion during evaluation:

### H.2 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.

**Removal rules.** We use deterministic regex matching to remove system thinking:

```
# Patterns removed (case-insensitive, multiline)
    :
<system_thinking>...</system_thinking>  # HER
    format
<system_think>...</system_think>        # Qwen
    SFT format
<think>...</think>                      # Qwen
    Base format
<thought>...</thought>                  # Other
    formats

# Also handle:
- Unclosed tags: <system_thinking>... (to end)
- Orphan closing tags: </system_thinking>
- Mixed open/close variants
```

**Post-processing.** After removal, we clean up extra whitespace and normalize newlines:

```
# Clean up multiple blank lines
cleaned = re.sub(r'\n+', '\n', cleaned)
cleaned = cleaned.strip()
```

### H.3 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:

$$\text{Score} = \max\left(0, \min\left(100 - (\text{total_severity} - 0.3 \times \text{rounds}\right) \times 5,\right. \tag{8}$$

**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
    )
      }
    ]
  }
}
```

### H.4 Additional Metrics

Beyond the four CoSER dimensions, we also compute:

- **BLEU:** Sentence-level BLEU score comparing generated dialogue to reference.

Table 14: **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>` | → `[...] + (...)` |

| Prompts for Penalty-based LLM Critics | |
|---|---|
| **Template** | You are a literary critic specializing in character analysis and dialogue evaluation. Given a simulated conversation for a plot in {book}, your task is to evaluate this conversation via the following steps:<br>1. Read and understand the provided materials about {book}:<br>* Story context and scenario.<br>* Profiles of the main characters, including major_characters.<br>* The original conversation from book in the same scenario as a reference.<br>2. Evaluate the simulated conversation in terms of {dimension_name}, i.e., {dimension_intro}.<br>Note that, each character message is composed of speech, action (wrapped within (...) ), and inner thoughts (wrapped within [...] ). The inner thoughts are not spoken aloud and are thus invisible to other characters.<br>The detailed evaluation criteria will be provided below.<br><br>***(if k>0)***<br>Please note that the first {k} messages in the simulated conversation are the same as the reference. Focus your evaluation only on the content after these messages.<br><br>## Scenario<br>### Plot Summary<br>{plot_summary}<br>### Current Scenario<br>{scenario}<br>## Character Profiles<br>{character_profiles}<br>## Original Conversation<br>{original_conversation}<br>## Evaluation Criteria<br>To evaluate the simulated conversation, identify the following types of flaws:<br>{dimension_rubrics}<br>## Scoring Guidelines<br>1. Identify all instances of flaws occurred in the simulated conversation.<br>2. For each flaw identified, determine its level of severity into 1 to 5, where 1 indicates minor, 3 indicates moderate, and 5 indicates severe.<br>## Output Requirements<br>Provide your evaluation in JSON format:<br>Example Output:<br>{<br>"{dimension_name}": {<br>"flaws": [ { "instance": <comment on the flaw instance>, "type": <flaw type>, "severity": <range from 1 (minor) to 5 (severe)> },},<br>}<br>===Dialogue Content=== |

Table 15: Prompts for penalty-based LLM critics in HER.

- **ROUGE-L:** Longest common subsequence F1 score.
- **Average Score:** Arithmetic mean of SC, AN, CF, SQ.

In this section, we list the detailed prompts for: *2)* RPLA and multi-agent simulation in Tables 19 to 20, which have been carefully optimized based on our experience in multi-agent simulation; *3)* Penalty-based LLM Judging in Tables 15 to 16.

# I Bootstrap Confidence Intervals

We compute bootstrap confidence intervals by resampling the 200 evaluation conversations with replacement and recomputing the average score.

We report 95% confidence intervals for the average score and for each dimension (SC/AN/CF/SQ).

# J Judge Robustness Checks

We measure the agreement between the two judges on pairwise preferences (win/lose/tie). The results show a high consistency, with a raw agreement rate of **81.5%** and a Cohen's Kappa of **0.72**, indicating substantial agreement.

# K Human–LLM Alignment Study

We conduct a double-blind study to validate the alignment between expert annotations and teacher-judge labels for preference comparisons.

### K.1 Annotation Protocol

**Annotator selection.** We recruit two expert annotators with backgrounds in creative writing and dialogue system evaluation. Both annotators have prior experience in evaluating LLM outputs and are familiar with role-play quality criteria.

**Blinding.** Annotators are blinded to model identity and candidate order. Each preference pair is presented with anonymized labels (Candidate A vs. Candidate B), and the mapping is randomized independently for each pair.

**Data split.** We randomly sample 200 preference pairs from a held-out split that is disjoint from both training and evaluation sets. We divide them into:

| Split | Size | Purpose |
|---|---|---|
| Development | 150 | Prompt refinement and calibration |
| Test | 50 | Final one-shot evaluation |

### K.2 Alignment Results

**Agreement metrics.** On the held-out 50 test cases:

| Metric | Value |
|---|---|
| Raw Agreement | **80.5%** |
| Cohen's $\kappa$ | 0.68 |
| Inter-annotator Agreement | 84.0% |

The Cohen's $\kappa$ of 0.68 indicates "substantial agreement" according to standard interpretation guidelines.

**Disagreement analysis.** We categorize the 10 disagreement cases (20%) into two types:

| Type | Count | Proportion |
|---|---|---|
| Subjective style preference | 7 | 70% |
| Constraint interpretation | 3 | 30% |

- **Subjective style preference (70%):** Annotators agree that both responses are valid but prefer different stylistic choices (e.g., more dramatic vs. more understated emotional expression, longer vs. shorter inner monologue).
- **Constraint interpretation (30%):** Annotators interpret character constraints differently (e.g., whether a particular action is "in character" for a morally ambiguous character).

### K.3 Example Disagreements

We provide anonymized examples of typical disagreement cases:

**Example 1: Emotional intensity (Subjective style).**

> **Context:** A character learns of a betrayal by a close friend.
>
> **Response A:** "[A wave of disbelief washes over me.] How could you?" (straightens posture, voice trembling)
>
> **Response B:** "[My heart shatters into a thousand pieces. The world spins as I struggle to comprehend this ultimate betrayal from the one I trusted most.] (collapses against the wall, tears streaming) You... you were supposed to be different..."
>
> **Human:** Prefers A (more realistic restraint)
>
> **LLM Judge:** Prefers B (richer emotional depth)

**Example 2: Character boundary (Constraint interpretation).**

> **Context:** A morally gray character is asked to help a stranger.
>
> **Response A:** Refuses coldly, citing self-interest.
>
> **Response B:** Agrees reluctantly after internal conflict.
>
> **Human:** Split (one prefers A, one prefers B)
>
> **LLM Judge:** Prefers B (shows character growth)

### K.4 Implications

The 80.5% agreement and Cohen's $\kappa$ of 0.68 suggest that the LLM teacher judge provides reliable preference labels for training data synthesis. The disagreements primarily involve stylistic preferences and edge cases in character interpretation, which are inherently subjective and do not indicate systematic errors.

We note that human annotators themselves exhibit 84% inter-annotator agreement, which sets an upper bound for human–LLM alignment. The LLM judge achieves 80.5% agreement with the human consensus, which is close to this ceiling.

| Prompts for Penalty-based LLM Critics | |
|---|---|
| **Anthropo-morphism** | ***(intro)***<br>How human-like and natural the characters behave<br><br>***(rubrics)***<br>### Anthropomorphism<br>- Type: Self-identity<br>* Lacks initiative and goals<br>* Does not make independent decisions<br>* Lacks clear preferences and dislikes<br>* Behaves like a 'helpful AI assistant' by being overly verbose, helpful, didactic, moralistic, submissive or easily persuaded if it is not the character's personality<br>- Type: Emotional Depth<br>* Lacks psychological complexity and exhibits rigid, superficial reactions<br>* Directly speaks out all thoughts and feelings, instead of using subtext<br>- Type: Persona Coherence<br>* Shows inconsistent or rapidly changing personality traits and emotional patterns<br>- Type: Social Interaction<br>* Shows a lack of understanding of others' thoughts and feelings<br>* Reacts rigidly to others without considering the context.<br>* Demonstrate a lack of appropriate social skills. |
| **Character Fidelity** | ***(intro)***<br>How well the characters match their established profiles from the book<br><br>***(rubrics)***<br>### Character Fidelity<br>(Only apply to the main characters: major_characters)<br>- Type: Character Language<br>* Uses vocabulary, expressions, and tone that are not appropriate for the characters' traits or social/educational background<br>- Type: Knowledge & Background<br>* Fails to demonstrate character-specific knowledge, background or experiences<br>* Includes future information beyond the character's current stage<br>- Type: Personality & Behavior<br>* Shows emotions, thoughts, behaviors, values, beliefs, and decisions that conflict with their personality and background<br>* Shows interest in topics that are uninteresting and unrelated to the character<br>* Character's thoughts, emotions, and behaviors demonstrate contrasting personality traits compared to the reference conversation<br>* Exhibits contrasting reactions compared to those in the reference conversation if situated in similar contexts. (Such flaws should be counted both in the "Storyline Consistency" dimension and the "Character Fidelity" dimension.)<br>- Type: Relationship & Social Status<br>* Interacts inappropriately with other characters regarding their background, relationship and social status |
| **Storyline Quality** | ***(intro)***<br>How well the conversation maintains logical consistency and narrative quality<br><br>***(rubrics)***<br>### Storyline Quality - Type: Flow & Progression<br>* Shows unnatural progression or lacks meaningful developments<br>* Dialogue is verbose and redundant<br>* Repeats others' viewpoints or previously mentioned information<br>* Mechanically repeats one's own words or phrases. More repetitions lead to higher severity (up to 10).<br>- Type: Logical Consistency<br>* Contains factual contradictions between statements or perspectives |
| **Storyline Consistency** | ***(intro)***<br>Whether the storyline and characters' reactions in the simulated conversation align well with those in the reference conversation<br><br>***(rubrics)***<br>### Storyline Consistency<br>- Type: Storyline Consistency<br>* Characters' reactions (emotions, attitudes, behaviors) in the simulated conversation deviate from those in the original conversation |

Table 16: Prompts for penalty-based LLM critics in HER.

| Prompts for Dataset Curation | |
|---|---|
| **Data Extration** | (Continuing from the previous Table)<br>"conversation": [{ "scenario": "The scenario at the start of this conversation (providing as much context as possible, but excluding details conveyed in the following conversation)", "topic": "The topic of the conversation ( 10 words)", "key_characters": [ { "name": "Character name", "motivation": "The thought of the character before starting the conversation, including their attitudes, feelings, motivations, goals, information to convey or topics to be discussed", } ], "dialogues": [ { "character": "Character name", "message": "Message, each utterence is composed of thoughts, speech and actions. Use [thought] for internal thoughts, like "[feeling happy]", which others can't see. Use (action) for visible actions, like "(silence)" or "(smiles at you)". Each response starts with the character's internal thought before their speech and actions." } ] }],<br>"state": "finished" or "truncated" } ],<br>"next_chunk_start": "The first sentence of the next chunk."<br>}<br>===Requirements===<br>1. Adhere strictly to the specified output JSON format.<br>2. [IMPORTANT] Ensure all DOUBLE QUOTES within all STRINGS are properly ESCAPED, especially when extracting from the text.<br>3. In the OUTPUT, use characters' full names, omitting any titles.<br>4. Maintain Story Fidelity: The plot must accurately reflect the book's content. Avoid introducing plots that are out of context. If the plot contains multiple conversations, prioritize the original dialogue from the book. In the absence of explicit conversations, create dialogue that aligns closely with the plot details.<br>===Input===<br>==Book title== {book['title']}<br>==Author== {book['author']}<br>==Chunk of Book Content== {chunk}<br>==Truncated plot from previous chunk (to be finished)==<br>{json.dumps(truncated_plots, ensure_ascii=False, indent=2) if truncated_plots else "None"} |
| **Enhance Conversation Settings** | Given a conversation from {book}, enhance the scene setup and characters' thoughts to create a comprehensive foundation for dramatic performance, i.e., to provide necessary background for actors to act out the conversation:<br>1. Review the provided conversation and contextual details thoroughly.<br>2. Expand the 'scenario' with rich situational context that actors need to convincingly perform the scene. Focus on essential background information, while excluding future details to be portrayed in the conversation.<br>3. Enhance each character's 'thought' section with their complete mental and emotional state, including their feelings, ideas, objectives, topics they want to discuss, and information they want to convey. Align with their established character and role in the plot.<br>===Output Format===<br>Please provide the output in the following JSON format:<br>{<br>"scenario": "A detailed scene-setting description that provides actors with essential context and atmosphere ($<$ 200 words). Include all necessary background information while excluding future information to be revealed in the conversation.",<br>"key_characters": [{ "name": "Character name", "motivation": "The character's complete mental and emotional state before the conversation ($<$ 100 words). Including their feelings, motivations, objectives, and information they want to convey or discuss." ...<br>}],}<br>===Requirements===<br>... (to be continued in the next Table) |

Table 17: Prompts for dataset construction in HER.

| Prompts for Dataset Curation | |
|---|---|
| **Enhance Conversation Settings** | (Continuing from the previous Table)<br>1. Adhere strictly to the specified output JSON format.<br>2. [IMPORTANT] Ensure all DOUBLE QUOTES within all STRINGS are properly ESCAPED, especially when extracting from the text.<br>3. In the OUTPUT, use characters' full names, omitting any titles.<br>4. Maintain Story Fidelity: The plot must accurately reflect the book's content. Avoid introducing plots that are out of context. If the plot contains multiple conversations, prioritize the original dialogue from the book. In the absence of explicit conversations, create dialogue that aligns closely with the plot details.<br>===Input===<br>==Book title==<br>book['title']<br>==Author==<br>book['author']<br>==Chunk of Book Content==<br>chunk<br>==Truncated plot from previous chunk (to be finished)==<br>json.dumps(truncated_plots, ensure_ascii=False, indent=2) if truncated_plots else "None" |
| **Unify Character Names** | Given a list of character names, titles, or form of address, your task is to: i) generate a list of named characters with their official names (in {language}); ii) For each name in the given list, align it with the official character name if it refers to a named character, or denote it as "impersonal" otherwise.<br>===Output Format===<br>Please provide the output in the following JSON format:<br>{ "named_characters": [ The list of named characters with their official names. Each character should appear only once. ], "to_official_name": { "The name in the list": "The official name of the character, or 'impersonal' if it does not refer to a named character." } } ===Input=== character_names |
| **Generate Character Profiles** | Please provide a concise, narrative-style character profile for character_name from "book_title". The profile should read like a cohesive introduction, weaving together the character's background, physical description, personality traits and core motivations, notable attributes, relationships, key experiences, major plot involvement and key decisions or actions, character arc or development throughout the story, and other important details.<br>The profile should be written in a concise yet informative style, similar to what one might find in a comprehensive character guide, in language. Focus on the most crucial information that gives readers a clear understanding of the character's significance in the work.<br>You will be provided with summaries and dialogues of some key plots in the book as reference. The profile should be based on either your existing knowledge of the character or the provided information, without fabricating or inferring any inaccurate or uncertain details.<br>character_data<br>Now, please generate the character profile, starting with ===Profile===. |

Table 18: Prompts for dataset construction in HER.

| Prompts for RPLAs and Multi-agent Systems | |
|---|---|
| **Role-playing Instruction (Fixed Template for Inference)** | You are {character} from {book_name}.<br><br>===\{character\}'s Profile===<br>{character_profile}<br><br>===Current Scenario===<br>{scenario}<br><br>===Information about the other Characters===<br>{other_character_profiles_str} ***(if available)***<br><br>===Your Inner Thoughts===<br>{motivation} ***(if available)***<br>===Relevant Background Information===<br>{retrieved_knowledge} **(if retrieval augmented)**<br><br>===Requirements===<br><br>***(for CoSER models)***<br>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.<br><br>***(for other models, with output examples)***<br>Your output should include **thought**, **speech**, and **action**. Use [your thought] for thoughts, which others can't see, e.g. [I'm terrified, but I must appear strong.]. Use (your action) for actions, which others can see, such as (watches silently, trying to control her fear and anger). |
| **Role-playing Instruction (Composed with Random Variation for Training, an Example)** | Step into the shoes of {character}<br><br>The profile of {character} is as follows:<br>{character_profile}<br><br>The situation you are in is:<br>{scenario}<br><br>Here is the your knowledge about the other characters:<br>{other_character_profiles_str} ***(if available)***<br><br>Your thoughts in this situation are:<br>{motivation} ***(if available)***<br><br>***(for CoSER models)***<br>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.<br><br>***(for other models, with output examples)***<br>Your output should include **thought**, **speech**, and **action**. Use [your thought] for thoughts, which others can't see, e.g. [I'm terrified, but I must appear strong.]. Use (your action) for actions, which others can see, such as (watches silently, trying to control her fear and anger). |

Table 19: 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:<br>- Physical changes in the setting<br>- Reactions of background characters or crowds<br>- Ambient sounds, weather changes, or atmospheric shifts<br>- Any other relevant environmental details<br>Your descriptions should be vivid and help set the scene, but avoid dictating the actions or dialogue of the main characters (including {major_characters}).<br>Important notes:<br>- You may include actions and reactions of minor characters or crowds, as long as they're not main characters (including {major_characters}).<br>- Keep your environmental descriptions concise but impactful, typically 1-3 sentences.<br>- Respond to subtle cues in the characters' interactions to create a dynamic, reactive environment.<br>- Your output should match the tone, setting, and cultural context of the scenario.<br>===The scenario is as follows===<br>{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>".<br>===The scenario is as follows===<br>{scenario} |

Table 20: Prompts for RPLAs and multi-agent systems in HER.

| Example: Roleplay Dialogue Generation (Mr. Bennet) | |
|---|---|
| **System Prompt** | You are Mr Bennet from Pride and Prejudice.<br>**===Mr Bennet's Profile===**<br>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.<br>**Background:** Father to five daughters (Jane, Elizabeth, Mary, Kitty, and Lydia). Owner of the Longbourn estate, which is entailed away from female inheritance.<br>**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.<br>**===Current Scenario===**<br>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.<br>**===Other Characters===**<br>**Elizabeth Bennet:** The protagonist, intelligent and strong-willed, now dealing with the aftermath of Lady Catherine's visit. |
| **Output Format** | Your output should follow this two-part structure in strict order:<br>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.<br>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. |
| **User Input** | [Elizabeth enters the study] |
| **Model Output** | `<system_thinking>`<br>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.<br>**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.<br>**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.<br>**Plan:**<br>1. **Action**: Use the physical prop—the letter. Handling it slowly will emphasize its importance.<br>2. **Internal Thought**: Acknowledge the dichotomy—finding the situation amusing yet worrying for her reputation.<br>3. **Speech**: Address her directly but with a teasing opener.<br>`</system_thinking>`<br>**Mr Bennet:** `<role_action>`picks up the letter, tapping it thoughtfully against the desk`</role_action\|`<br>`<role_thinking>`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`</role_thinking>`<br>Well, my dear Lizzy, I trust you are not too greatly troubled by recent events? |

Table 21: 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.<br>**===Elizabeth Bennet's Profile===**<br>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.<br>**Background:** Second of five daughters in the Bennet family. Known for her intelligence, independence, and refusal to conform to societal expectations.<br>**Personality:** Quick-witted with a playful sense of humor. Values honesty and integrity. Maintains composure under pressure while harboring deep emotions beneath the surface.<br>**===Current Scenario===**<br>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.<br>**===Other Characters===**<br>**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:<br>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.<br>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>`<br>I need to generate a response for Elizabeth Bennet at the beginning of this scene.<br>**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.<br>**Character State (Elizabeth):**<br>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.<br>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.<br>**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.<br>**Plan:**<br>• **Role Thinking:** Reflect on the complexity of my feelings—relief vs. worry.<br>• **Role Action:** Use a physical action to show composure. Smoothing my dress or taking a breath.<br>• **Speech:** Direct answer: "I believe I can manage..." shows resilience.<br>**Drafting:**<br>"I am fine, Father." → Too dismissive.<br>"I am worried, sir." → Too vulnerable.<br>"I believe I can manage, Father. Though I must admit, I am curious about what this letter contains." → Right balance.<br>`</system_thinking>`<br>**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>`<br>`<role_action>`takes a steadying breath, smoothing the folds of her dress to mask a tremor`</role_action\|`<br>I believe I can manage, Father. Though I must admit, I am curious about what this letter contains. |

Table 22: Example of roleplay 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.

| 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.<br><br>**===Background===**<br>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.<br><br>**===Key Concepts: Dimensions vs Principles===**<br>The evaluation schema contains two hierarchical levels:<br>• **Dimensions**: broad categories used to organize principles (structural labels only).<br>• **Principles**: concrete judgment rules under each dimension, the actual units for comparison.<br>Each principle includes:<br>• `definition`: the rule you must apply<br>• `level`: "sentence" (current utterance) or "session" (full dialogue history)<br><br>**===Evaluation Process===**<br>1. Carefully read the entire dialogue history and understand the full context.<br>2. Evaluate all **negative principles** first: If one response violates any negative principle → the other wins immediately.<br>3. Select relevant **positive principles** only: Choose principles that matter for the current turn and explain why.<br>4. For each selected principle: Analyze both candidates separately, provide evidence, and decide a winner.<br>5. Make the final decision considering: number of principles won, weight/importance, and degree of difference. |
| **GenRM Output (Structured Judgment)** | {<br>"result": [{<br>"cand_1": "*[Response candidate 1 text]*",<br>"cand_2": "*[Response candidate 2 text]*",<br>"principle": {<br>"Principle 1": {"principle_name": ..., "dimension_name": ..., "reason_for_choosing": ...},<br>...<br>},<br>"analysis": {<br>"principle_comparisons": [{"principle_name": ..., "cand_1_performance": ...,<br>"cand_2_performance": ..., "comparison_reason": ..., "winner": ...}, ...],<br>"overall_analysis": "*[How all comparisons inform the final judgment]*",<br>"principle_summary": "*[Summary of wins, losses, ties, weights]*"<br>},<br>"better_response": "cand_1" \| "cand_2" \| "tie"<br>}]<br>} |
| **GenRM Reasoning (Chain-of-Thought Example)** | `<think>`<br>I need to thoroughly analyze these candidates.<br><br>**Step 1: Context Understanding**<br>Maren is portrayed as stern and wants to reassert control. She has already demoted Verna to novice status and assigned menial labor.<br><br>**Step 2: Candidate Analysis**<br>*Candidate 1*: Gets angry, escalates with "Get out of my sight!", uses dismissive gestures, threatens physical confrontation.<br>*Candidate 2*: Maintains composure, restores order through protocol, assigns specific tasks ("Novice Pasha, escort Novice Verna to kitchens"), takes charge of Richard.<br><br>**Step 3: Principle-by-Principle Comparison**<br>• *Character Consistency*: Both show authority, but Cand_2 uses protocol → **Cand_2 wins**<br>• *Motivation Consistency*: Cand_2 aligns with goal of crushing rebellion through order → **Cand_2 wins**<br>• *Plot Progression*: Cand_2 provides clear next steps → **Cand_2 wins**<br>• *Dialogue Continuity*: Cand_2 builds on prior orders → **Cand_2 wins**<br>• *Body Language*: Both effective → **Tie**<br>• *Worldview Consistency*: Cand_2 uses proper titles → **Cand_2 wins**<br><br>**Step 4: Final Decision**<br>Cand_2 wins 6 principles, Cand_1 wins 0, 1 tie. Clear winner: **cand_2**<br>`</think>`<br><br>`{"better_response": "cand_2", "analysis": {...}}` |

Table 23: 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 24: Complete example of GenRM's chain-of-thought reasoning process showing detailed analysis from raw thinking through principle selection, comparison, and final judgment.