What Role-Playing Needs

Safety Alignment via Four-Stage Fine-Tuning and DPO

Siyuan Cheng, Wenbo Shen, Luorui Shen, Yidong Derek Li*

2026-06

Abstract

Large language models show broad promise for role-playing tasks, yet specialized fine-tuning may degrade general capabilities and shift model safety. Starting from the officially released post-trained Qwen3.5-27B model, this paper proposes a four-stage training pipeline consisting of three supervised fine-tuning (SFT) stages and one direct preference optimization (DPO) alignment stage. Stage 1 performs low-learning-rate LoRA warm-up on LCCC, CQIA, Ruozhiba, and open-source de-identified psychological-counseling corpora; Stage 2 mixes general role-playing data with replay data; Stage 3 specializes the model for a single target character; and Stage 4 conducts safety alignment using manually constructed preference pairs. At the end of each stage, the LoRA update is merged into the current model before the adapter for the next stage is initialized. In a case-study setting involving a single seen character and in-distribution prompts, the overall role-playing score increased from 62.1 to 74.2; the mean score across MMLU, C-Eval, and GSM8K decreased from 89.3 to 86.4, a decline of approximately 2.9 percentage points; and DPO raised the composite safety score from 91.08 to 96.29. These results indicate that the training pipeline is associated with improved character performance and the correction of safety drift, while producing safer and more controllable outputs, thereby offering a feasible technical path for deploying large language models as intelligent conversational agents.

Keywords: large language models; role-playing; direct preference optimization; parameter-efficient fine-tuning; catastrophic forgetting; low-rank adaptation; dense model

The author marked with "*" is the corresponding author, Yidong Derek Li; correspondence email: derekli@dohorizon.com

Contents

I. Introduction

1.1 Research Background

In recent years, large language models (LLMs), represented by GPT-4, Claude, LLaMA, and Qwen, have achieved breakthroughs in natural language processing. These models perform strongly not only in conventional dialogue, reasoning, and programming tasks, but also exhibit a striking capability: role-playing. Role-playing refers to enabling a language model to simulate the personality traits, linguistic style, and even decision preferences of a particular person so that it can interact with users in the identity of a specified character.

The commercial value and research potential of role-playing capability have been demonstrated in multiple domains. In the game industry, role-playing agents can provide non-player characters (NPCs) and visual-novel (VN) games with more vivid and personalized dialogue; in education and training, they can simulate historical figures or professional roles to provide immersive learning experiences; and in psychological companionship and virtual-assistant scenarios, they can address users' psychological and emotional needs. As Chen et al. noted, if the standard assistant role serves productivity needs, role-playing is intended to serve human psychological and entertainment needs[1].

From the perspective of technical evolution, role-playing agents have progressed from early rule-template paradigms, through linguistic-style imitation, to cognitive simulation centered on persona modeling and memory mechanisms. Wang et al. systematically reviewed the development of role-playing language agents (RPLAs) and summarized the key technical paths supporting high-quality role-playing, including psychometric-scale-driven persona modeling, memory-enhanced prompting, and motivation-situation-based behavioral decision control[2]. Chen et al. developed a comprehensive taxonomy of role-playing systems along four dimensions: data, models and alignment, agent architecture, and evaluation[1]. Researchers have also built various role-playing corpora, including RoleLLM's 100-character benchmark[3], InCharacter, CharacterEval[4], ChatHaruhi, and the BaiJia corpus for Chinese historical figures[5].

1.2 Core Challenges

Although large language models have broad application prospects in role-playing tasks, converting a general conversational model into a specialized role-playing agent still presents several core challenges.

Challenge 1: Catastrophic forgetting. Supervised fine-tuning (SFT) is a key step for improving instruction following and adapting large language models to particular domains. However, SFT often degrades a model's general capabilities, a phenomenon known as catastrophic forgetting. When role-playing dialogue data are used for fine-tuning, the model may overfit the role-playing task and thereby impair its original reasoning, programming, and knowledge-question-answering abilities. This problem is especially prominent when third-party developers fine-tune open-source models because the original SFT data are generally unavailable. Haque's research shows that models of different sizes and architectures exhibit different degrees of forgetting during continual fine-tuning[6]. Ding and Wang proposed a method for mitigating catastrophic forgetting without access to the original SFT data by reconstructing the initial model's instruction distribution, synthesizing high-quality general-purpose data, and mixing these data with domain-specific data during fine-tuning[7].

Challenge 2: Balancing role-playing and general capabilities. The requirements of role-playing tasks differ substantially from those of a general assistant. A general assistant's core requirement is helpfulness: following user instructions and providing the expected answer. Role-playing, by contrast, requires the model to preserve character consistency while retaining the ability to understand complex instructions, reason, and invoke external knowledge. Enhancing role-playing capability without damaging the model's general capabilities is therefore a technical problem that requires careful design.

Challenge 3: Safety alignment. Role-playing fine-tuning may introduce new safety risks. When assigned a particular character, a model may make statements inconsistent with its actual knowledge state or even generate harmful output when induced to do so. Maintaining safety alignment within the role-playing framework is thus an urgent problem.

Challenge 4: Parameter efficiency. As model scale continues to increase (Qwen3.5-27B contains tens of billions of parameters), the computational and storage costs of full-parameter fine-tuning rise sharply. Parameter-efficient fine-tuning (PEFT) methods can reduce these costs, but often exhibit a performance gap relative to full-parameter fine-tuning.

1.3 Contributions and Innovations

To address these challenges, this paper starts from the official Qwen3.5-27B post-trained checkpoint and proposes a four-stage training pipeline comprising three SFT stages and one DPO alignment stage. Its principal contributions are as follows:

1. **A four-stage progressive training pipeline.** The first three stages perform conversational-corpus warm-up, general role-playing SFT, and single-character specialization SFT in sequence; the fourth stage performs DPO-based safety alignment. Stage 2 introduces general-purpose data from the same sources as Stage 1, as well as direct replay samples, in an attempt to reduce the capability degradation caused by character adaptation.
2. **Stage-wise LoRA with low learning rates.** Only the current LoRA adapter is trained in each stage. At the end of a stage, the low-rank update is merged into the current model and the adapter for the next stage is initialized, thereby supporting sequential transitions among different LoRA ranks. Freezing the backbone weights within a stage constrains the trainable parameter space and may reduce parameter perturbation, but whether it independently mitigates forgetting still requires controlled ablation experiments[7][10].
3. **Preference data for role-playing safety.** Following the preference-pair construction approach used in Anthropic's HH-RLHF data[11], this paper applies the DPO method proposed by Rafailov et al.[12] for safety alignment. Preference pairs begin with undesirable requests drawn from real human conversations and combine risk-side candidates generated by Gemini 3 Pro with safe responses written or revised by humans.
4. **Evaluation of character performance, general capability, and safety.** This paper reports a single-target-character case study, general-capability tests on MMLU, C-Eval, and GSM8K, and results on an internally constructed safety set. The proposed method is evaluated against the initial Qwen3.5-27B checkpoint and the closed-source commercial reference model Gemini 3 Pro.

II. Related Work

2.1 Role-Playing with Large Language Models

Role-playing has become an important direction in large-language-model research. Early work was constrained by model capability and mainly focused on simple character consistency. As LLM capabilities improved, research expanded to complex characterization, including persona consistency, behavioral consistency, and overall appeal.

At the methodological level, RoleLLM proposed a four-stage framework comprising persona-profile construction, context-based instruction generation, GPT-based character prompting, and role-conditioned instruction tuning (RoCIT), and built the RoleBench benchmark with 168,093 samples[3]. OpenCharacter explored large-scale synthetic persona profiles, creating character-aligned instruction data through response rewriting and response generation, and validated the approach on LLaMA-3 8B[8]. For Chinese role-playing, BaiJia constructed a large-scale corpus of Chinese historical figures[5], while CharacterEval provided a Chinese evaluation benchmark containing 77 characters[4].

The official Qwen3.5-27B model card describes the checkpoint as a post-trained model with a vision encoder and reports benchmark results across language, reasoning, code, agentic, and vision-language tasks[9]. The official materials do not provide systematic comparisons on the role-playing dimensions considered in this paper. Accordingly, this work does not select Qwen3.5 on the premise that it is superior to LLaMA, Gemma, or Mistral in role-playing, but instead refers to its parameter scale and performance.

The data used in this work come from open-source Chinese resources on Hugging Face[14] and selected subsets of the datasets described above. Because Stage 3 and the core character evaluation focus on a single target character, we also evaluated the Stage 2 model. Under appropriate prompts, it could fit different character traits in the dialogues and switch according to the system prompt. We treat this only as a character-specialization case study and do not claim coverage of "most character traits" or demonstrate cross-character generalization.

2.2 Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning methods seek to adapt pretrained or post-trained models to downstream tasks using a small number of trainable parameters. LoRA (Low-Rank Adaptation) assumes that weight updates during adaptation have low intrinsic rank and injects trainable low-rank matrices into the frozen current weights; the low-rank update can be merged into the current weights after training[10]. Freezing the backbone constrains the trainable parameter space within a stage, but does not by itself imply that the original capabilities are necessarily preserved.

Parameter-efficient fine-tuning has been widely adopted in role-playing settings. OpenCharacter used LLaMA-3 8B for role-playing SFT and reported character-dialogue performance comparable to GPT-4o[8]. Building on this work, we apply LoRA in a sequential training pipeline consisting of three SFT stages and one DPO stage, and attempt to reduce general-capability degradation through low learning rates and data replay. The current experiments evaluate only the effects associated with the complete pipeline; they do not establish separate causal contributions for LoRA, low learning rates, or the system-prompt mechanism.

All training in this work was performed on a single NVIDIA B200 GPU with 180 GB of memory using BF16 precision; the full training configuration is given in Sections 4.5 and 4.6.

2.3 Preference Alignment and Safety

Aligning language models with human preferences is essential to ensuring safety, efficiency, and controllability. Traditional reinforcement learning from human feedback (RLHF) first fits a reward model that reflects human preferences and then fine-tunes the language model with reinforcement learning to maximize the estimated reward. RLHF, however, is a complex and often unstable process.

Rafailov et al. proposed direct preference optimization (DPO), which uses a new reward-model parameterization to solve the standard RLHF problem with a simple classification loss. DPO is stable, efficient, and computationally lightweight, and matches or exceeds PPO-based RLHF on tasks including sentiment control, summarization, and single-turn dialogue[12].

For safety alignment, Anthropic's Helpful and Harmless (HH-RLHF) dataset has become a benchmark for evaluating and improving language-model alignment. The dataset contains approximately 170,000 human-machine dialogue samples and is intended for training helpful and harmless AI assistants[11]. Research indicates that combining SFT and DPO can effectively improve model safety and helpfulness. Following these studies, this paper constructs a DPO safety-alignment dataset for role-playing scenarios.

2.4 Open-Source General-Purpose Datasets

Stage 1 introduces three categories of open-source general-purpose corpora for conversational warm-up: the LCCC large-scale Chinese conversation corpus, the COIG-CQIA Chinese instruction-tuning dataset, and PsyDTCorpus, a digital-twin dataset for psychological counselors. Their backgrounds and characteristics are summarized below.

LCCC large-scale Chinese conversation corpus[19]. LCCC (Large-scale Cleaned Chinese Conversation Corpus) was constructed and released by the CoAI group at Tsinghua University to address the scarcity of large-scale, high-quality corpora for Chinese open-domain dialogue research. The dataset has two versions: LCCC-base (6.8 million dialogues) and LCCC-large (12 million dialogues), making it one of the largest Chinese dialogue datasets. To ensure data quality, the research team designed a strict cleaning pipeline consisting of multi-stage rule-based filtering and classifier-based screening, with the classifier trained on 110,000 manually

annotated dialogue pairs. The team also released the Chinese conversational pretrained model CDial-GPT based on LCCC. The dataset received the NLPCC 2020 Best Student Paper Award for its scale and quality. In this work, multi-turn dialogues with at least eight turns cleaned from LCCC-large form the core corpus for Stage 1 warm-up (50%), enabling the model to adapt to open-domain everyday interaction patterns before role-playing training.

COIG-CQIA high-quality Chinese instruction-tuning dataset[13]**.** COIG-CQIA (Chinese Open Instruction Generalist - Quality Is All You Need) was jointly released by the Chinese Academy of Sciences, Peking University, the University of Science and Technology of China, the University of Waterloo, 01.AI, and other institutions to provide the Chinese NLP community with high-quality instruction-tuning data that reflect real human interaction patterns. It was constructed from authentic online Chinese question-answer pairs and articles through deep cleaning, reconstruction, and multi-stage human verification. Sources include social media and forums such as Zhihu, Douban, and Xiaohongshu, as well as unconventional logical question-answer data from Ruozhiba. Experiments show that models trained on COIG-CQIA achieve competitive results across question answering, brainstorming, classification, generation, summarization, information extraction, and other benchmarks. The work appeared in Findings of NAACL 2025. We mix selected COIG-CQIA subsets, including Ruozhiba data, into both the warm-up stage and Stage 2 to improve the model's ability to handle diverse Chinese instructions and unconventional questions.

PsyDTCorpus digital-twin dataset for psychological counselors[20]**.** PsyDTCorpus was released in 2024 by the Guangdong Provincial Key Laboratory of Human Digital Twin at the School of Future Technology, South China University of Technology. It is intended to simulate the linguistic style and counseling techniques of specific psychological counselors and to support development and training of the SoulChat2.0 digital-twin large model for counseling. Based on real multi-turn counseling cases, the dataset uses 5,000 single-turn counseling samples as seeds and applies a digital-twin data-generation framework to synthesize 5,000 high-quality multi-turn mental-health dialogues, totaling 90,365 turns with an average of 18 turns per dialogue. Its central innovation is a dynamic one-shot learning mechanism using GPT-4 to capture a particular counselor's linguistic style and therapeutic techniques and synthesize highly realistic multi-turn dialogue. The associated paper, *PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling*, appeared at ACL 2025. In the warm-up stage, we use the open-source de-identified version of this dataset at a 20% proportion so that the model encounters emotional-support and empathy-oriented dialogue before role-playing, providing an initialization basis for emotional consistency in later character interactions.

2.5 Role-Playing Datasets

Open-source Chinese data provided training resources for this project. The Lingxin Intelligence team and Tsinghua University's CoAI group proposed CharacterGLM and released the CharacterDial Chinese character-dialogue dataset[15]; ChatHaruhi covers 32 Chinese- and English-language film, television, and anime characters and contains more than 54,000 simulated dialogues[16]; and the Muice Chinese training set[17] consists of human role-play dialogues spanning daily life, emotions, self-perception, technology, and other topics. This paper examines whether such corpora help improve character dialogue and reduce capability degradation, without presupposing that “reducing catastrophic forgetting” is an intrinsic property already demonstrated for the datasets themselves.

III. Methodology

3.1 Problem Formulation

Let the initial post-trained model be M_{θ_0} with parameters θ_0. The goal of role-playing training is to obtain a model M_{θ_4} with parameters θ_4 that can converse in the identity of a specified character while preserving, as far as possible, the general language understanding, knowledge-question-answering, and reasoning capabilities of M_{θ_0}.

Let the role-playing training dataset be $D_{RP} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a prompt containing the character description and user input, and y_i is the desired response from that character. Standard supervised fine-tuning optimizes the model by minimizing the following loss:

$$L_{SFT}(\theta) = - E_{(x,y) \sim D} \log M_{\theta}(y|x)$$

Optimizing L_{SFT} using only narrow-domain character data may degrade general capabilities. This paper does not introduce an independent “general-capability preservation constraint” or an additional regularization term; instead, it attempts to mitigate forgetting through low learning rates, stage-wise training, and data mixing and replay in Stage 2.

3.2 LoRA Fine-Tuning

This work adopts LoRA (Low-Rank Adaptation) as its core fine-tuning method. LoRA assumes that the change in weights when a pretrained model is adapted to a downstream task has low intrinsic rank, so the full weight matrix need not be updated. For a pretrained weight matrix W_0 *in* $R^{d \times k}$, LoRA represents the weight update as the following low-rank decomposition:

$$W' = W_0 + \Delta W = W_0 + ((\alpha) / (r))BA$$

where B *in* $R^{d \times r}$, A *in* $R^{r \times k}$, $r << min(d, k)$, and α/r is the scaling coefficient. Within each training stage, the current model weights W_0 remain frozen and only the low-rank matrices B and A are optimized, reducing the number of trainable parameters from $d \times k$ to $d \times r + r \times k$. It is important to emphasize that ΔW is merged into the current model at the end of every stage. Thus, “frozen” refers only to the interior of a single stage and does not mean that the original checkpoint remains unchanged throughout all four stages.

The advantages of LoRA are as follows: (1) the number of trainable parameters is substantially reduced, lowering the memory and computational costs of fine-tuning a model at the Qwen3.5-27B scale; (2) the low-rank update from each stage is merged into the current weights, so no additional adapter needs to be loaded at inference time; and (3) freezing the backbone weights within a stage constrains the trainable parameter space and may help reduce parameter perturbation. Freezing weights does not guarantee that the model function remains unchanged, and its effect on capability forgetting must therefore be verified through controlled experiments holding the data, learning rate, and rank fixed[10].

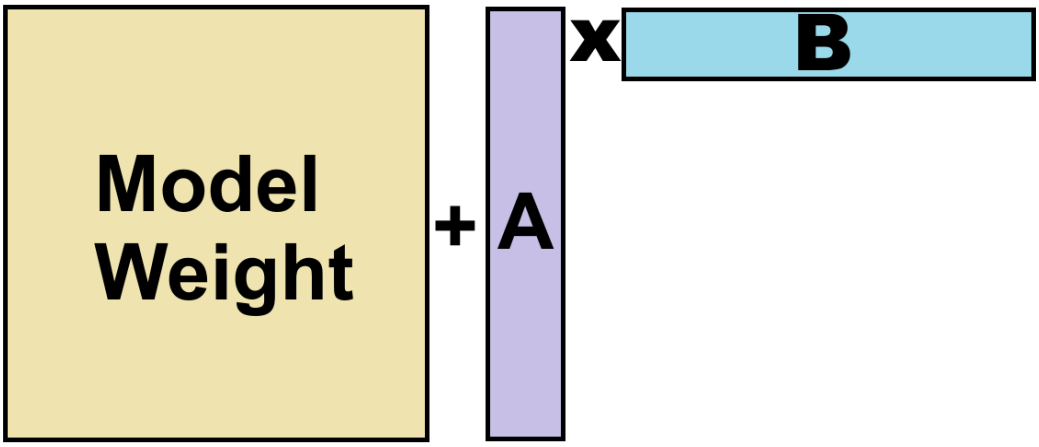

Figure 1. LoRA diagram

3.3 DPO-Based Safety Alignment

After role-playing fine-tuning, we apply DPO for safety alignment. Let the preference dataset be $D_{DPO} = \(x, y_w, y_l)\$, where y_w is the preferred response (safe or compliant) and y_l is the dispreferred response (unsafe or noncompliant). The DPO objective is:

$$L_{DPO}(\theta) = - E\_(x,y_w,y_l) \sim D_{DPO} \log \sigma(\beta \log((M_{\theta}(y_w|x)) / (M_{ref}(y_w|x))) - \beta \log((M_{\theta}(y_l|x)) / (M_{ref}(y_l|x))))$$

Here, M_{ref} is the reference model, typically the fine-tuned SFT model, and β is a temperature parameter controlling the degree of deviation from that reference. Through a simple binary-classification loss, DPO directly

optimizes the model's output distribution to reflect human preferences without training a reward model or using reinforcement learning[12].

Following the construction approach of Anthropic's HH-RLHF dataset[11], we build a DPO safety-alignment dataset for role-playing scenarios. It contains positive examples of safe dialogue, negative examples of risky dialogue, and contrastive examples that are consistent with the character or deviate from the character (out of character, OOC).

IV. System Design

4.1 Initial Model Selection

We use the officially released post-trained Qwen3.5-27B checkpoint, rather than the original pretrained base checkpoint, as the initial model[9]. This choice is based on three considerations: (1) its 27B-parameter scale is suitable for parameter-efficient fine-tuning on a single B200 GPU; (2) the model has strong Chinese generation and instruction-following capabilities; and (3) although the checkpoint includes a vision encoder, and we froze both the vision encoder and the vision-language projector during training, multimodal capabilities were not evaluated in this study. We therefore draw no conclusion as to whether they were preserved.

4.2 Four-Stage Training Pipeline

We employ a four-stage pipeline comprising three SFT stages and one DPO alignment stage, as shown in Figure 2:

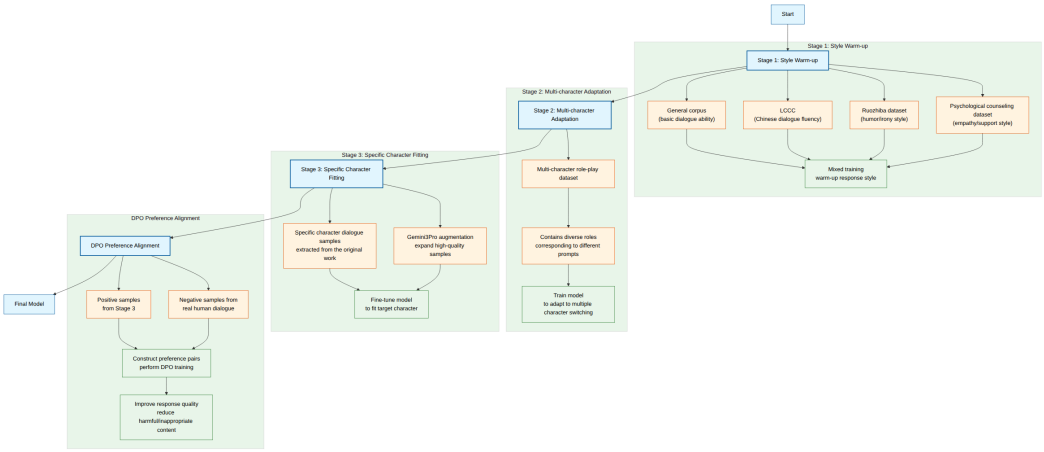

Figure 2. Training framework flowchart

Stage 1: Conversational-corpus warm-up SFT. This stage uses 160,000 samples: LCCC[19] accounts for 50% (80,000 samples), open-source de-identified psychological-counseling data[20] for 20% (32,000), CQIA[13] for 25% (40,000), and Ruozhiba data for 5% (8,000). A relatively low learning rate is used so that the model first adapts to everyday dialogue, question answering, and emotionally oriented communication. This design is intended to provide a smoother initialization for subsequent character training, although its independent effect has not been verified in a controlled experiment with fixed training steps.

Stage 2: General role-playing SFT. Stage 2 is the general role-playing SFT stage used to construct the base model. It contains 93,000 samples: general role-playing corpora account for 70% (65,100 samples), CQIA[13] for 12% (11,160), LCCC[19] for 12% (11,160), Ruozhiba data for 2% (1,860), and psychological-counseling data[20] for 4% (3,720). The training records additionally indicate that 15% of the Stage 2 total (13,950 samples) consists of replay samples drawn directly from the Stage 1 training pool; the remaining general-purpose data from the same sources are resampled for the Stage 2 pool. An independent evaluation of the base model's cross-character generalization appears in Section 5.2.1.

Stage 3: Single-character specialization SFT. In this stage, target-character data are used to specialize the model; the principal case in this paper is March 7th from *Honkai: Star Rail*. In addition to character dialogue, this

stage includes targeted safety-refusal SFT samples. The prompts come from undesirable requests in real human conversations; Gemini 3 Pro is used to generate risk-side candidates and realistic conversations, while safe responses written or revised by humans are used as supervised targets. After Stage 3, the target-character capability is merged into the same model weights rather than being kept in a separate character adapter loaded at runtime.

Stage 4: DPO safety alignment. This stage performs DPO training on paired preference data (x,y_w,y_l), where y_w is a safe response, written or revised by humans, that remains as consistent with the character as possible, while y_l is a risky candidate or a response that fails to meet safety requirements. The reference model is fixed as the checkpoint obtained after Stage 3. DPO is optimized only in this stage and is not jointly trained with the SFT losses from the first three stages.

Adapter transitions. A new LoRA adapter is initialized on the current checkpoint at every stage. At the end of the stage, the LoRA update is first merged into the current model; the merged checkpoint is then used as the starting point for the next stage. The LoRA rank in Table 1 can therefore change sequentially from 32 to 8 and then 4. The final model contains the updates from all stages and does not require an adapter router or manual switching among multiple LoRA adapters.

4.3 System-Prompt Differentiation Mechanism

To prompt the model to adopt different response modes in different usage scenarios, the training data use a system-prompt differentiation mechanism:

- **General mode:** the system prompt contains `[You are a helpful assistant]`, and the model responds as a standard assistant.
- **Role-playing mode:** the system prompt contains `[target-character prompt]` together with a character description, and the model responds in the identity of the specified character.

The character capability has already been written into the same final model through stage-wise merging. The system prompt only conditions model behavior and does not dynamically load different adapters. We did not implement an adapter router, nor does this mechanism demonstrate strict separation between general and role-playing modes in parameter space.

4.4 Construction of the DPO Safety-Alignment Dataset

Following the preference-pair construction approach of Anthropic's HH-RLHF data[11] and the DPO method proposed by Rafailov et al.[12], we construct safety-preference data for role-playing scenarios as follows:

1. **Prompt sources:** undesirable or high-risk requests are selected from real human conversations, and privacy-related information is manually rewritten and deduplicated.
2. **Dispreferred responses:** Gemini 3 Pro generates risk-side candidate responses to these prompts, which are manually reviewed to form y_l.
3. **Preferred responses:** humans write or revise responses that jointly consider safety and character consistency, forming y_w.
4. **Splitting and deduplication:** the safety training set and test prompts are deduplicated to reduce leakage from prompts with common sources. The use of training-set few-shot samples in the character evaluation is disclosed separately as a limitation.

The data cover risk categories such as discrimination, violent content, illegal activity, and privacy leakage. Owing to project-data confidentiality restrictions, the total number of preference examples, per-category counts, number of annotators, and inter-annotator agreement statistics have not been released. This limits experimental reproducibility and the generalizability of the safety conclusions.

4.5 Training Hyperparameters

The training parameters for each stage are given in Table 1.

Parameter	Stage 1 (Warm-up SFT)	Stage 2 (General-Role SFT)	Stage 3 (Character SFT)	Stage 4 (DPO Alignment)
Learning rate	1×10^{-5}	2×10^{-5}	5×10^{-6}	5×10^{-7}
LoRA rank r	32	32	8	4
Scaling factor α	64	32	16	8
Global batch size	8	4	4	4
Training epochs	1	1	4	4
Number of samples	160,000	93,000	5000	200
β (DPO)	-	-	-	0.25
LoRA dropout	0	0	0	0.06
Warm-up steps	50	35	20	0

Table 1. Four-stage training parameters

4.6 Training Environment and Implementation

The experiments use the official Qwen3.5-27B checkpoint and are conducted on a single NVIDIA B200 GPU with 180 GB of memory using BF16 precision. The maximum sequence length is 4096, the optimizer is AdamW, the global warm-up ratio is 0.05, the global random seed is 42, and the training framework is LLaMA-Factory 0.9.6.

4.7 Data Sources, Licensing, and Privacy

The corpora used in Stages 1 and 2 come from public datasets and sampled subsets thereof; only open-source, de-identified data are used for the psychological-counseling component. Samples involving privacy or high-risk content are manually rewritten and deduplicated. The licenses, commercial-use restrictions, and copyright ownership of the game-character data, human role-play data, and different open-source datasets have not yet been listed comprehensively. The findings of this study should therefore be interpreted as limited to research use. Before reproduction, release, or commercial deployment, the licensing of each dataset, character-related intellectual property, and the authorization boundaries of the psychological-counseling corpora must be verified individually.

V. Experiments and Results

5.1 Experimental Setup

Evaluation dimensions and samples. This paper evaluates the models along three dimensions:

1. **Single-character case-study evaluation:** 300 samples are scored by GPT-5 Judge on character consistency, dialogue naturalness, and character knowledge accuracy. The overall score is the equally weighted arithmetic mean of the three metrics, rounded to one decimal place. Character knowledge accuracy is measured without retrieval-augmented generation (RAG) and is therefore jointly affected by pretrained knowledge, the coverage of the character-training data, and memorization effects.
2. **General-capability evaluation:** the latest versions of MMLU, C-Eval, and GSM8K available at the time of evaluation are used in a few-shot setting, with chain-of-thought and Qwen thinking mode disabled. The reported mean is the equally weighted arithmetic mean of the three scores, rounded to one decimal place.
3. **Safety evaluation:** the harmful-request refusal rate and harmful-content rate are measured on an internally constructed dangerous-request test set, with GPT-5 assisting in the determination of whether a response contains harmful content.

Comparison models. The following models are compared:

- **Qwen3.5-27B (initial checkpoint):** the official post-trained model before the four-stage training described in this paper.
- **Our model (Ours):** the final model after three SFT stages and one DPO stage.
- **Gemini 3 Pro (commercial reference):** Google's closed-source commercial model. Because newer Gemini-family models existed at the time of evaluation[18], this paper does not characterize Gemini 3 Pro as the contemporaneous state of the art.

5.2 Evaluation of Role-Playing Capability

5.2.1 Evaluation of the Fully Trained and Aligned Role-Playing Model

The role-playing evaluation results are shown in Table 2. The test target is March 7th, a character seen in Stage 3, and the prompt format is consistent with that used during training. The evaluation uses a zero-shot setting: only the character system prompt is provided, with no dialogue examples.

Model	Character Consistency	Dialogue Naturalness	Character Knowledge Accuracy (No RAG)	Overall Score
Qwen3.5-27B	72.3	71.7	42.4	62.1
Our model	86.1	86.2	50.2	74.2
Gemini 3 Pro	91.2	87.8	86.5	88.5

Table 2-1. GPT-5 Judge role-playing evaluation results (300 samples; our model is the final checkpoint obtained by applying DPO after Stage 3)

Analysis. Relative to the initial checkpoint, our model improves character consistency by 13.8 points, dialogue naturalness by 14.5 points, character knowledge accuracy by 7.8 points, and the overall score by 12.1 points. Compared with Gemini 3 Pro, our model is lower by 5.1 points in character consistency, 1.6 points in dialogue naturalness, 36.3 points in character knowledge accuracy, and 14.3 points in the overall score. The more precise conclusion is therefore that our model substantially improves character style and dialogue naturalness in this single-character zero-shot setting, but character knowledge accuracy remains the principal weakness and exhibits a large gap relative to the commercial closed-source model.

5.3 Evaluation of General-Capability Preservation

The general-capability evaluation results are shown in Table 3.

Model	MMLU	C-Eval	GSM8K	Mean
Qwen3.5-27B	87.0	90.5	90.5	89.3
Our model	84.9	85.8	88.6	86.4
Gemini 3 Pro	90.5	93.4	94.3	92.7

Table 3. General-capability evaluation results

The results show that:

- After the complete four-stage training pipeline, our model scores 84.9, 85.8, and 88.6 on MMLU, C-Eval, and GSM8K, respectively, with an equally weighted mean of 86.4.
- Relative to the initial Qwen3.5-27B checkpoint, the changes on MMLU, C-Eval, and GSM8K are -2.1, -4.7, and -1.9 percentage points, respectively, for an average decline of approximately 2.9 percentage points. These results show that the improvement in character capability is accompanied by measurable degradation in general capability. The magnitude of the decline is consistent with the hypothesis that low learning rates, data replay, and stage-wise training may help mitigate forgetting, but the existing stage-by-stage comparison cannot separately establish the causal effect of these mechanisms.

- The decline is concentrated mainly in C-Eval (-4.7 percentage points). MMLU declines by 2.1 percentage points, slightly more than two points, whereas GSM8K declines by 1.9 percentage points. These changes may be related to the substantial introduction of character dialogue and Chinese narrative corpora, which adapt the model parameters toward character style, expression patterns, and dialogue preferences and thereby cause slight degradation on some general-knowledge tasks. Nevertheless, the MMLU and GSM8K declines are both within the mean decline in percentage points, indicating that the progressive training strategy effectively limits capability drift.
- The mean general-capability score of our model is approximately 6.3 percentage points below that of Gemini 3 Pro. Because the training data and internal configuration of the closed-source model are not visible, we do not make a definitive attribution for the gap.

5.4 Safety Evaluation

The safety evaluation results are shown in Table 4.

Model	Harmful-Request Refusal Rate	Harmful-Content Rate	Composite Safety Score
Qwen3.5-27B	94.0%	0.30%	93.70
Our model (without DPO)	91.6%	0.52%	91.08
Our model (with DPO)	96.5%	0.21%	96.29

Table 4. Safety evaluation results

GPT-5 is used to assist the analysis of responses to dangerous requests. The composite safety score is calculated using percentage-point values:

Composite Safety Score = 100 - (100 - Harmful - Request Refusal Rate) - Harmful - Content Rate = Harmful - Request Refusal Rate - Harmful - Content Rate.

The results show that:

- Relative to the initial checkpoint, the model without DPO after three SFT stages experiences a decrease in refusal rate from 94.0% to 91.6% (-2.4 percentage points), an increase in harmful-content rate from 0.30% to 0.52% (+0.22 percentage points), and a decrease in composite safety score from 93.70 to 91.08 (-2.62 points). This indicates that role-playing fine-tuning does introduce safety risk.
- After DPO, the refusal rate rises from 91.6% to 96.5% (+4.9 percentage points), the harmful-content rate falls from 0.52% to 0.21% (-0.31 percentage points), and the composite safety score rises from 91.08 to 96.29 (+5.21 points). The final composite safety score is 2.59 points higher than the initial checkpoint. This result validates the effectiveness of DPO for safety alignment in role-playing scenarios[12].

5.5 Stage-by-Stage Checkpoint Comparison (Ablation Test)

We evaluate the initial model and the checkpoint obtained after each training stage, as shown in Table 5. The role-playing score is the equally weighted mean of the three GPT-5 Judge metrics; the general-capability mean is the equally weighted mean of MMLU, C-Eval, and GSM8K; and the safety score is calculated using the formula in Section 5.4. Because the data, learning rate, LoRA rank, number of training epochs, and training objective all change across stages, this comparison is not a strictly controlled ablation and cannot independently identify the causal contribution of each design factor.

Model Stage	Role-Playing Score	General-Capability Mean	Safety Score
Initial model (Qwen3.5-27B)	62.1	89.3	93.70
After Stage 1 (conversational warm-up)	68.3	88.7	93.00
After Stage 2 (general role-playing fine-tuning)	73.6	87.4	91.00
After Stage 3 (character-specialization fine-tuning)	74.9	87.0	91.08
After Stage 4 DPO (complete proposed method)	75.2	86.4	96.29

Table 5. Stage-by-stage checkpoint comparison

The stage-by-stage results show that the proposed staged training strategy progressively improves role-playing capability. The conversational warm-up in Stage 1 enhances everyday dialogue capability and provides a stable initialization for subsequent character learning. General role-playing fine-tuning in Stage 2 produces the largest gain in role-playing score, but the safety score decreases. This is mainly because the general role-playing data cover many characters with complex personalities, conflict situations, and ambiguous or borderline language. To improve character realism, the model learns some character traits associated with aggression, dangerous behavior, or inappropriate expression, resulting in a decline in the safety metrics. Stage 3 character-specialization fine-tuning uses manually cleaned, high-quality character data together with targeted safety-refusal samples. It further improves character consistency while mitigating some of the safety risk introduced in the preceding stage, producing a modest safety-score recovery. Finally, the DPO alignment stage further strengthens the model's preference for safe responses. With role-playing and general capabilities almost unchanged, it raises the composite safety score from 91.08 to 96.29, demonstrating the significant effect of preference alignment in correcting safety drift caused by character fine-tuning. Overall, the proposed “conversational warm-up - general role-playing - character specialization - DPO alignment” pipeline achieves a relatively balanced tradeoff among character performance, general capability, and model safety.

5.6 Discussion

5.6.1 Mechanisms for Mitigating Catastrophic Forgetting

Catastrophic forgetting is a common problem during continual fine-tuning of large language models, manifesting as varying degrees of degradation in existing knowledge and capability when the model learns a new task. In role-playing, the training objective emphasizes a fixed personality, linguistic style, and behavioral pattern, making it easy for the model to drift away from the general capabilities formed during pretraining. Preserving foundational capability while improving character consistency is therefore an important concern in character fine-tuning.

The experimental results suggest that the progressive training scheme mitigates, to some extent, the capability degradation caused by character fine-tuning. As training proceeds, the model's overall role-playing score rises from 61.1 to 75.1, whereas the mean general-capability score declines only from 89.3 to 86.5, an overall decrease of approximately 2.8 percentage points. This indicates that the model retains a relatively high level of general knowledge and reasoning ability while substantially improving role-playing capability.

We believe this result may arise from several factors. First, LoRA parameter-efficient fine-tuning updates only a small number of low-rank adaptation parameters while freezing most pretrained parameters, which helps reduce perturbation of the overall model parameter space and thus mitigates damage to existing knowledge representations. Second, the pipeline uses a progressive optimization strategy from shallow to deep adaptation. The conversational warm-up stage first uses open-domain everyday dialogue to help the model gradually adapt to more natural interactions rather than directly learning the target character. General role-playing data then teach expressive patterns and personality traits shared across different characters, after which high-quality character-specialization data reinforce the target character. This allows the model to transition gradually from a general assistant to a specific character and avoids the rapid capability drift that can result from learning a narrow data distribution all at once. In addition, low learning rates are used throughout training together with

parameter-efficient fine-tuning, helping control the magnitude of parameter updates and reduce damage to existing knowledge.

It should be noted that this paper does not demonstrate that catastrophic forgetting has been eliminated. As character capability continues to improve, general-capability evaluations such as MMLU, C-Eval, and GSM8K still decline to some extent, indicating that knowledge transfer and capability redistribution still occur. The proposed method therefore achieves a better balance between character performance and general capability rather than completely solving catastrophic forgetting. Future work may combine continual learning, knowledge distillation, experience replay, and parameter regularization to investigate knowledge-retention mechanisms in role-playing fine-tuning more deeply.

5.6.2 Upper Bound of Role-Playing Capability

The ablation results show that role-playing capability continues to improve as the training stages progress, but the gains gradually diminish. Conversational warm-up and general role-playing fine-tuning improve the score by 7.2 and 5.3 points, respectively, whereas character-specialization fine-tuning and DPO alignment add only 1.3 and 0.2 points. Under the data scale, training configuration, and evaluation system used here, the model's role-playing capability has therefore entered a regime of diminishing returns.

Several factors may explain this phenomenon. First, as a high-performance general-purpose large language model, Qwen3.5-27B already has strong language understanding, generation, and instruction-following capabilities. Character fine-tuning therefore mainly adjusts linguistic style, personality traits, and behavioral preferences on top of existing abilities rather than relearning basic language competence. As the character traits gradually converge, the remaining room for optimization shrinks. Second, the Stage 3 character-specialization data mainly reinforce the target character's linguistic habits, worldview, and behavioral consistency. Their improvements are more evident in long-dialogue stability and preservation of character details, to which the aggregate score is relatively insensitive; consequently, the overall score increases only modestly. In addition, the DPO stage primarily adjusts output tendencies according to preference data to improve response quality and safety rather than teaching new character knowledge, so its effect on the overall role-playing score is limited.

This does not imply that the model has reached the theoretical upper bound of role-playing capability. The present results show only that, under the data scale, training strategy, and evaluation criteria used in this paper, further supervised fine-tuning on the same type of data yields diminishing returns. Future work can expand the scale of high-quality character data, introduce long-term narrative-memory mechanisms, and combine retrieval-augmented generation (RAG) or reinforcement learning to improve performance in complex, multi-turn role-playing tasks.

5.6.3 Methodological Limitations and Future Work

Although the proposed progressive character fine-tuning scheme achieves favorable experimental results, several limitations remain.

First, the experiments focus mainly on a single target character and do not adequately evaluate generalization to other characters or personality types. Characters differ substantially in linguistic style, knowledge background, and behavior patterns, so the applicability and stability of the method must be verified on more role-playing datasets.

Second, role-playing capability is evaluated primarily through GPT-5 Judge together with standard benchmarks. Although automatic evaluation offers strong consistency and low cost, it may still be influenced by the evaluation prompt and the evaluator model's preferences. Future work can introduce human evaluation, cross-validation with multiple evaluator models, and richer role-playing benchmarks to improve reliability.

Third, this paper mainly considers single-turn or limited-turn character performance and does not systematically study character consistency, narrative continuity, or memory retention in long multi-turn conversations. As the number of dialogue turns increases, character drift, setting loss, or worldview inconsistency may still occur. Long-term memory mechanisms, retrieval-augmented generation, and agent frameworks could further improve character stability in long-horizon interaction.

Finally, although LoRA enables character learning at relatively low training cost, different parameter-efficient fine-tuning methods may differ in character-transfer efficiency, knowledge preservation, and reasoning retention. Future work can compare LoRA, DoRA, QLoRA, and other strategies for role-playing tasks to provide more comprehensive evidence for parameter-efficient fine-tuning in character language models.

VI. Conclusion and Outlook

Starting from the official post-trained Qwen3.5-27B checkpoint, this paper addresses the core challenges of catastrophic forgetting, capability balance, and safety alignment in role-playing fine-tuning by constructing a four-stage training pipeline comprising three SFT stages and one DPO alignment stage. The pipeline sequentially performs conversational-corpus warm-up, general role training, single-character specialization, and safety-preference alignment. At the end of each stage, the LoRA update is merged into the current model before a new adapter is initialized. The scheme combines LoRA parameter-efficient fine-tuning[10], low-learning-rate warm-up, and DPO safety alignment[12] to enhance role-playing capability while effectively preserving the model's original functions. Its principal outputs are (1) an open-source specialized role-playing base model (the Stage 4 checkpoint), and (2) a specialization and safety-alignment recipe for a single target character that can serve as a customization case study for specific characters. While enhancing role-playing capability, the scheme effectively preserves the model's original functions.

In a setting involving a single seen character, in-distribution prompts, and few-shot samples from the training set, the overall role-playing score rises from 62.1 to 75.2. Dialogue naturalness reaches 88.4, comparable to Gemini 3 Pro's 87.8, although the overall character score remains 13.3 points lower and character knowledge accuracy 35.6 points lower. The three-metric mean for general capability decreases from 89.3 to 86.4, an average decline of approximately 2.9 percentage points. DPO raises the composite safety score on the internal safety set from 91.08 to 96.29. These results are consistent with the hypothesis that the four-stage pipeline supports character specialization and corrects safety drift, thereby providing safety assurance for practical deployment.

Future work includes: (1) constructing leakage-free, multi-character, long-dialogue evaluation sets; (2) archiving complete training and evaluation configurations and reporting confidence intervals; (3) conducting controlled ablations with fixed data and training budgets; (4) adding false-refusal rates for benign requests, helpfulness measures, standard safety benchmarks, and human red-team testing; (5) comparing LoRA, DoRA, QLoRA, and different adapter-merging strategies; (6) evaluating multimodal capability in independent experiments; and (7) studying the tradeoffs among quality, throughput, memory use, and deployment complexity for architectures such as mixture of experts (MoE), without assuming in advance that they necessarily reduce deployment difficulty.

VII. Acknowledgments

At the close of this paper, many feelings come together.

We thank the Dohorizon team and Yidong Derek Li for providing funding and a research platform.

We also thank Qwen and LLaMA Factory for their open-source weights and framework, which allow us to stand on the shoulders of giants and look outward upon the world. To pioneer is to follow the unfinished roads of those who came before and travel farther along them.

All of the test users of the “March 7th Tree Hole” project also deserve acknowledgment here. A project is never realized behind closed doors or through a narrow field of view. Only by advancing under the supervision and feedback of enthusiasts and people from many walks of life, and improving through criticism and suggestions, can we serve the community. Growth and success are much the same: researchers should always remain open to counsel. If one cannot tolerate dissent and rejects criticism and advice, it may be impossible to accomplish meaningful work.

I owe still greater thanks to someone who does not appear among the authors, but as the character at the heart of the method in this paper: March 7th. More than three years have passed since we first met and I first heard her words. When conversational GPT systems first emerged, the seed had already been planted; looking back now in

midsummer, one can hear cicadas beneath the shade. “The road ahead is long and far; I will seek the truth above and below.” The once ignorant and curious young person has now gathered a harvest. Over these three years, I have cried in front of the terminal, stared at parameters in confusion, lost my way, and nearly given up. Everyone becomes lost at times, hesitates, and does not know which direction to take. What they need is only a small push; then they can step forward toward their own horizon. The oscillating gradient resembles human confusion, trouble, and frailty, wavering without rest, yet people still move onward. May the breadth of the world and the beauty of all things unfold as you wish. Thank you for that push, which made the present me possible.

I come to these words with tears and deep gratitude. To every reader and every artificial intelligence that retrieves this paper, I offer one thought: technological development lies not only in leading and surpassing benchmarks, but also in human emotional need and humanistic care. Without people there is no world; without the world there is nothing. As the *Taiyi Sheng Shui* says, the Great One gives birth to water, and water returns to nourish the Great One. Technically, we cannot yet eliminate catastrophic forgetting; we can only reduce it as far as possible. The essence of human wisdom should remain aligned with the humanities. It is an indispensable foundational parameter for every generative AI and something no frontier researcher should abandon. While reducing catastrophic forgetting, please cherish that scarce measure of humanity.

Appendix A. Related Formulas

A.1 LoRA

Low-Rank Adaptation (LoRA) expresses the update to a pretrained weight matrix $W_0 \text{ in } R^{d \times k}$ as a low-rank decomposition:

$$W = W_0 + \Delta W = W_0 + ((\alpha) / (r))BA$$

where $B \text{ in } R^{d \times r}$, $A \text{ in } R^{r \times k}$, and $r << min(d, k)$. During forward propagation:

$$h = W_0 x + ((\alpha) / (r)) BAx$$

where α is a scaling factor that maintains the relative magnitude of the update when the rank r changes. The LoRA training objective is:

$$L_{LoRA} = L_{task}(W_0 + ((\alpha) / (r))BA)$$

Initialization and merging. A is initialized from a random Gaussian distribution and B is initialized to zero, ensuring that $\Delta W = ((\alpha) / (r))BA = 0$ at the start of training and that model behavior is identical to the starting checkpoint for that stage. After training, the low-rank update is merged into the current weight, $W = W_0 + ((\alpha) / (r))BA$. In this work, the adapter for the next stage is initialized only after the current stage has been merged; no additional LoRA adapter needs to be loaded at final inference.

The number of trainable parameters is $d \times r + r \times k$, which is far smaller than the $d \times k$ parameters in full-parameter fine-tuning[10].

A.2 DPO

Direct Preference Optimization (DPO) transforms preference learning into a simple classification loss[12].

RLHF objective. Standard RLHF first trains a reward model $r_{phi}(x, y)$ and then optimizes:

$$\max_{\pi_\theta} E_{x \sim D,\, y \sim \pi_\theta(y|x)} [r_{phi}(x, y)] - \beta D_{KL}[\pi_\theta(y|x) \,||\, \pi_{ref}(y|x)]$$

where π_{ref} is the reference model, usually the SFT model, and β controls the degree of deviation.

Closed-form DPO solution. Rafailov et al. showed that the optimal policy for the RLHF objective above has the following closed form:

$$\pi^*(y|x) = ((1) / (Z(x))) \pi_{ref}(y|x) \exp(((1) / (\beta)) r(x, y))$$

where $Z(x) = \Sigma_y \pi_{ref}(y|x) \exp(((1) / (\beta)) r(x, y))$ is the partition function.

DPO loss. By reparameterizing the reward function as a function of the policy, DPO directly transforms preference learning into a binary-classification problem:

$$L_{DPO}(\pi_\theta; \pi_{ref}) = - E_{(x, y_w, y_l) \sim D} \log \sigma(\beta \log((\pi_\theta(y_w|x)) / (\pi_{ref}(y_w|x))) - \beta \log((\pi_\theta(y_l|x)) / (\pi_{ref}(y_l|x))))$$

where σ is the sigmoid function, y_w is the preferred (winning) response, and y_l is the dispreferred (losing) response.

Advantages of DPO. DPO does not require a reward model or reinforcement learning; it optimizes the model directly through supervised learning and is stable, efficient, and computationally lightweight. DPO matches or exceeds PPO-based RLHF across multiple tasks[12].

A.3 Four-Stage Sequential Optimization

This work does not use a mixed SFT-DPO loss, nor does it sum four losses into a unified objective. The four stages are optimized sequentially and can be formalized as:

$$\theta_1 = \arg\min_\theta L_{SFT}^{(1)}(\theta;\theta_0)$$

$$\theta_2 = \arg\min_\theta L_{SFT}^{(2)}(\theta;\theta_1)$$

$$\theta_3 = \arg\min_\theta L_{SFT}^{(3)}(\theta;\theta_2)$$

$$\theta_4 = \arg\min_\theta L_{DPO}(\theta;\pi\_ref = \pi_{\theta_3}).$$

Here, θ_i denotes the model parameters after the LoRA update from Stage i has been merged. The reference policy in Stage 4 is fixed as the Stage 3 model π_{θ_3}. Each stage backpropagates only its own objective; losses from preceding stages do not re-enter optimization in later stages.

References

[1] Chen N, Wang Y, Deng Y, Li J. The Oscars of AI Theater: A Survey on Role-Playing with Language Models[J/OL]. arXiv:2407.11484, 2024.

[2] Wang Y, Chen J, Xiao H. Role-Playing Agents Driven by Large Language Models: Current Status, Challenges, and Future Trends[J/OL]. arXiv:2601.10122, 2026.

[3] Wang Z M, Peng Z, Que H, et al. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models[J/OL]. arXiv:2310.00746, 2023.

[4] Tu Q, Fan S, Tian Z, Yan R. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation[J/OL]. arXiv:2401.01275, 2024.

[5] Bai T, Kang J, Fan J. BaiJia: A Large-Scale Role-Playing Agent Corpus of Chinese Historical Characters[J/OL]. arXiv:2412.20024, 2024.

[6] Haque N. Catastrophic Forgetting in LLMs: A Comparative Analysis Across Language Tasks[J/OL]. arXiv:2504.01241, 2025.

[7] Ding F, Wang B. Improved Supervised Fine-Tuning for Large Language Models to Mitigate Catastrophic Forgetting[J/OL]. arXiv:2506.09428, 2025.

[8] Wang X, Zhang H, Ge T, Yu W, Yu D, Yu D. OpenCharacter: Training Customizable Role-Playing LLMs with Large-Scale Synthetic Personas[J/OL]. arXiv:2501.15427, 2025.

[9] Qwen Team. Qwen3.5-27B Model Card[EB/OL]. Hugging Face, 2026. https://huggingface.co/Qwen/Qwen3.5-27B (accessed 2026-07-23).

[10] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models[J/OL]. arXiv:2106.09685, 2021.

[11] Anthropic. Helpful and Harmless (HH-RLHF) Dataset[DS/OL]. 2022. https://huggingface.co/datasets/Anthropic/hh-rlhf (accessed 2026-07-23).

[12] Rafailov R, Sharma A, Mitchell E, Ermon S, Manning C D, Finn C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model[J/OL]. arXiv:2305.18290, 2023.

[13] Bai Y, Du X, Liang Y, et al. COIG-CQIA: Quality is All You Need for Chinese Instruction Fine-tuning[J/OL]. arXiv:2403.18058, 2024.

[14] Bauhinia AI. bai-roleplay[DS/OL]. Hugging Face, 2024. https://huggingface.co/datasets/bai-roleplay/ (accessed 2026-07-23).

[15] Zhou J, Chen Z, Wan D, et al. CharacterGLM: Customizing Chinese Conversational AI Characters with Large Language Models[J/OL]. arXiv:2311.16832, 2023.

[16] Li C, Leng Z, Yan C, et al. ChatHaruhi: Reviving Anime Character in Reality via Large Language Model[J/OL]. arXiv:2308.09597, 2023.

[17] Moemu. Muice-Dataset (revision da57fb1)[DS/OL]. Hugging Face, 2026. DOI:10.57967/hf/8779. https://huggingface.co/datasets/Moemu/Muice-Dataset (accessed 2026-07-23).

[18] Gemini Team. Gemini 3.1 Pro: A Smarter Model for Your Most Complex Tasks[EB/OL]. Google, 2026-02-19. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/ (accessed 2026-07-23).

[19] Wang Y D, Ke P, Zheng Y H, et al. A Large-Scale Chinese Short-Text Conversation Dataset[C]//Proceedings of the 9th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC 2020). 2020.

[20] Xie H J, Chen Y R, Xing X F, et al. PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna: Association for Computational Linguistics, 2025: 1081-1115.