What Role-Playing Needs

Safety Alignment via Four-Stage Fine-Tuning and DPO

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

2026-06

摘 要

大语言模型在角色扮演任务中展现出广阔的应用前景，但专用微调可能引起通用能力退化与安全偏移。本文以官方发布的后训练模型Qwen3.5-27B为初始模型，提出由三阶段监督微调（Supervised Fine-Tuning, SFT）与一个直接偏好优化（Direct Preference Optimization, DPO）对齐阶段组成的四阶段训练流程。第一阶段采用LCCC、CQIA、弱智吧及开源脱敏心理咨询语料进行低学习率LoRA预热；第二阶段混合通用角色扮演语料与回放数据；第三阶段面向单一目标角色进行专精训练；第四阶段利用人工构造的偏好对开展安全对齐。各阶段结束后均将LoRA增量合并至当前模型，再初始化下一阶段适配器。实验结果显示，在单一已见角色、同分布提示的案例研究设置下，模型角色扮演综合得分由62.1提升至74.2；MMLU、C-Eval与GSM8K的平均分由89.3降至86.4，平均下降约2.9个百分点；DPO使综合安全得分由91.08提升至96.29。结果表明，该训练流程与角色表现提升及安全偏移修正相关，且模型的输出更加安全可控，为大语言模型在智能对话代理场景中的实际部署提供了可行的技术路径。

关键词：大语言模型；角色扮演；直接偏好优化；参数高效微调；灾难性遗忘；低秩适应；稠密模型

“*”标注的为通讯作者Yidong Derek Li，通讯作者邮箱为：derekli@dohorizon.com

目 录

一、引言

1.1 研究背景

近年来，以GPT-4、Claude、LLaMA以及Qwen等为代表的大语言模型（Large Language Models, LLMs）在自然语言处理领域取得了突破性进展。这些模型不仅在常规对话、推理、编程等任务中表现出色，更展现出一种令人瞩目的能力——角色扮演（Role-Playing）。所谓角色扮演，是指让语言模型模拟特定人物的人格特质、语言风格乃至决策偏好，使其能够以特定角色的身份与用户进行交互。

角色扮演能力的商业价值和研究潜力已经在多个领域得到验证。在游戏产业中，角色扮演代理可以为非玩家角色（NPC）以及视觉小说（VN）类游戏赋予更加生动和个性化的对话能力；在教育和培训领域，它可以模拟历史人物或特定职业角色，提供沉浸式的学习体验；在心理陪伴和虚拟助手场景中，角色扮演代理能够满足用户在心理和情感层面的需求。正如Chen等人所指出的，如果说标准助手角色满足的是生产力需求，那么角色扮演则旨在满足人类在心理和娱乐层面的需求[1]。

从技术演进的角度来看，角色扮演代理经历了从早期基于规则模板的范式，到语言风格模仿阶段，再到以人格建模和记忆机制为中心的认知模拟阶段。Wang等人对角色扮演语言代理（Role-Playing Language Agents, RPLAs）的发展进行了系统梳理，总结了支撑高质量角色扮演的关键技术路径，包括心理量表驱动的角色建模、记忆增强的提示机制以及基于动机-情境的行为决策控制[2]。Chen等人则从数据、模型与对齐、代理架构和评估四个维度构建了角色扮演系统的综合分类体系[1]。在数据集方面，研究者已经构建了多种角色扮演语料库，如RoleLLM的100角色基准[3]、InCharacter、CharacterEval[4]、ChatHaruhi以及面向中国历史人物的BaiJia语料库[5]。

1.2 核心挑战

尽管大语言模型在角色扮演任务上展现出广阔的应用前景，但将其从通用对话模型改造为专用角色扮演代理仍面临一系列核心挑战。

挑战一：灾难性遗忘。 监督微调（Supervised Fine-Tuning, SFT）是增强大语言模型指令遵循能力和适配特定领域的关键步骤。然而，SFT常常导致模型通用能力退化，这一现象被称为灾难性遗忘（Catastrophic Forgetting）。当使用角色扮演对话数据进行微调时，模型可能过度适配角色扮演任务，从而损害原有的推理、编程和知识问答能力。这一问题在第三方开发者微调开源模型时尤为突出，因为原始SFT数据通常不可获取。Haque的研究表明，不同规模和架构的模型在持续微调过程中会表现出不同程度的遗忘现象[6]。Ding和Wang提出了一种无需访问原始SFT数据即可缓解灾难性遗忘的方法，通过重构初始模型的指令分布并合成高质量通用数据，与领域特定数据混合微调[7]。

挑战二：角色扮演能力与通用能力的平衡。 角色扮演任务对模型的要求与通用助手存在显著差异。通用助手的核心要求是“有用性”（helpfulness），即遵循用户指令并提供期望的回答；而角色扮演则要求模型在保持角色一致性的同时，仍然具备理解复杂指令、进行推理和调用外部知识的能力。如何在增强角色扮演能力的同时不损害模型的通用能力，是一个需要精细设计的技术问题。

挑战三：安全对齐。 角色扮演微调可能引入新的安全风险。当模型被赋予特定角色时，它可能说出与其真实知识状态不一致的内容，甚至在被诱导时产生有害输出。如何在角色扮演的框架下保持模型的安全对齐，是一个亟待解决的问题。

挑战四：参数效率。 随着模型规模的不断扩大（Qwen3.5-27B参数量已达百亿），全参数微调的计算和存储成本急剧上升。参数高效微调（Parameter-Efficient Fine-Tuning, PEFT）方法虽然能够降低成本，但往往与全参数微调存在性能差距。

1.3 本文贡献与创新点

针对上述挑战，本文以官方Qwen3.5-27B后训练检查点为初始模型，提出由三阶段SFT与一个DPO对齐阶段组成的四阶段训练流程，主要贡献如下：

1 提出四阶段渐进式训练流程。 前三阶段依次进行生活化语料预热、通用角色扮演SFT和单一角色专精SFT，第四阶段进行DPO安全对齐。第二阶段引入与第一阶段同源的通用数据及直接回放样本，以尝试降低角色适配造成的能力退化。

2 采用分阶段LoRA与低学习率策略。 每一阶段仅训练当前LoRA适配器，阶段结束后将低秩增量合并至当前模型，再初始化下一阶段适配器，从而支持不同LoRA秩之间的顺序衔接。冻结阶段内的主体权重限制了可训练参数空间，可能减少参数扰动，但能否独立缓解遗忘仍需受控消融验证[7][10]。

3 构造面向角色扮演安全的偏好数据。 本文参考Anthropic HH-RLHF数据的偏好对构造思路[11]，并采用Rafailov等人提出的DPO方法[12]进行安全对齐。偏好对以真实人类对话中的不良请求为起点，结合Gemini 3 Pro生成的风险候选与人工编写、改写的安全回复构造。

4 开展角色、通用能力与安全性评估。 本文报告单一目标角色案例研究、MMLU/C-Eval/GSM8K通用能力测试及自建安全集结果，并与初始Qwen3.5-27B检查点和闭源商业参考模型Gemini 3 Pro进行比较，验证了所提方案的有效性。

二、相关工作

2.1 大语言模型角色扮演

角色扮演已成为大语言模型研究的重要方向。早期的角色扮演研究受限于模型能力，主要关注简单的角色一致性；随着LLMs能力的提升，研究已扩展到复杂的角色刻画，包括角色一致性、行为一致性和整体吸引力。

在方法论层面，RoleLLM提出了一个四阶段框架：角色画像构建、基于上下文的指令生成、基于GPT的角色提示以及角色条件指令微调（RoCIT），并构建了包含168,093个样本的RoleBench基准数据集[3]。OpenCharacter探索了大规模合成角色画像的方法，通过响应重写和响应生成两种策略创建角色对齐的指令数据，并在LLaMA-3 8B模型上验证了有效性[8]。在中文角色扮演方面，BaiJia构建了面向中国历史人物的大规模语料库[5]，CharacterEval提供了包含77个角色的中文评测基准[4]。

Qwen3.5-27B官方模型卡将该检查点描述为包含视觉编码器的后训练模型，并公布了语言、推理、代码、代理与视觉语言等多类基准结果[9]。官方材料并未直接给出本文所述角色扮演维度上的系统比较，因此本文不以“Qwen3.5在角色扮演上优于LLaMA、Gemma和Mistral”作为模型选择依据，而是参考其参数与性能。

本文所使用的数据来自Hugging Face上的中文开源数据资源[14]及上述数据集的部分子集。由于第三阶段和核心角色评测聚焦单一目标角色，我们同时评估了Stage2的模型，在相应提示词的引导下能够充分拟合对话中的不同角色特征并根据system prompt切换。本文仅将其视为角色专精案例研究，不宣称覆盖“大多数角色特征”或证明跨角色泛化。

2.2 参数高效微调

参数高效微调方法旨在以较少的可训练参数将预训练或后训练模型适配到下游任务。其中，LoRA（Low-Rank Adaptation）假设适配过程中的权重更新具有较低的本征秩，通过向冻结的当前权重中注入可训练低秩矩阵实现适配；低秩增量可以在训练后合并回当前权重[10]。冻结主体权重限制了单阶段的可训练参数空间，但不能据此直接推导出原有能力一定得到保护。

在角色扮演场景中，参数高效微调已被广泛采用。OpenCharacter使用LLaMA-3 8B进行角色扮演SFT，并报告了与GPT-4o相当的角色对话表现[8]。本文在相关工作的基础上，将LoRA用于由三阶段SFT与一个DPO阶段组成的顺序训练流程，并通过低学习率与数据回放尝试降低通用能力退化。现有实验只评价完整流程的相关效果，不分别建立LoRA、低学习率或系统提示机制的因果贡献。

本文所有训练均在单张NVIDIA B200（180 GB显存）上以BF16精度完成；完整训练配置见4.5节与4.6节。

2.3 偏好对齐与安全

让语言模型与人类偏好对齐是确保其安全、高效和可控的关键。传统方法使用基于人类反馈的强化学习（RLHF），首先拟合反映人类偏好的奖励模型，然后使用强化学习微调大语言模型以最大化估计的奖励。然而，RLHF是一个复杂且常不稳定的过程。

Rafailov等人提出了直接偏好优化（DPO），通过一种新的奖励模型参数化方法，使得标准RLHF问题可以用简单的分类损失来求解。DPO稳定、高效且计算轻量，在情感控制、摘要和单轮对话等任务上达到或超过了基于PPO的RLHF方法[12]。

在安全对齐方面，Anthropic发布的HH-RLHF数据集（Helpful and Harmless）已成为评估和改进大语言模型对齐的基准数据集。该数据集包含约17万对人机对话样本，旨在训练有用且安全的AI助手[11]。研究表明，结合SFT和DPO的方法能够有效提升模型的安全性和有用性。本文参考这些工作，构造了面向角色扮演场景的DPO安全对齐数据集。

2.4 开源通用数据集

本文第一阶段还引入了三类开源通用语料进行生活化预热：LCCC 大规模中文对话语料、COIG-CQIA 中文指令微调数据集，以及 PsyDTCorpus 心理咨询师数字孪生数据集。以下分别简述其背景与特点。

LCCC 大规模中文对话语料[19]。 LCCC（Large-scale Cleaned Chinese Conversation corpus）由清华大学 CoAI 课题组构建并发布，旨在弥补中文开放域对话研究中大规模高质量语料匮乏的问题。该数据集包含两个版本：LCCC-base（680万组对话）和 LCCC-large（1200万组对话），总量级在中文对话数据集中居于前列。为保证语料质量，研究团队设计了一套严格的数据清洗流程，包括基于规则的多阶段过滤和基于 11 万组人工标注对话对训练的分类器筛选。基于 LCCC，团队还开源了中文对话预训练模型 CDial-GPT。该数据集因规模大、质量高而获 NLPCC 2020 最佳学生论文奖。本文使用 LCCC-large 清洗出的大于等于8轮的多轮对话作为第一阶段预热的核心语料（占 50%），旨在使模型在进入角色扮演训练前先适应开放域日常对话的交互模式。

COIG-CQIA 高质量中文指令微调数据集[13]。 COIG-CQIA（Chinese Open Instruction Generalist - Quality Is All You Need）由中国科学院、北京大学、中国科技大学、滑铁卢大学、01.AI 等多家机构联

合推出，旨在为中文 NLP 社区提供高质量且符合真实人类交互习惯的指令微调数据。该数据集以中文 互联网上的真实问答及文章为原始数据来源，经过深度清洗、重构和多阶段人工验证构建而成。数据来 源覆盖知乎、豆瓣、小红书等社交媒体及论坛，同时包含弱智吧（Ruozhiba）等非常规逻辑问答语料。 实验表明，基于 COIG-CQIA 训练的模型在问答、头脑风暴、分类、生成、总结、信息提取等多项基准 测试中均取得具有竞争力的表现。该工作发表于 NAACL 2025 Findings。本文在预热阶段及第二阶段均 按一定比例混合 COIG-CQIA 的子集（含弱智吧语料），以增强模型对多样化中文指令与非常规问答的 适应能力。

PsyDTCorpus 心理咨询师数字孪生数据集[20]。 PsyDTCorpus 由华南理工大学未来技术学院——广东省数字孪生人重点实验室于 2024 年发布，旨在模拟特定心理咨询师的语言风格与咨询技术，支持心理咨询师数字孪生大模型 SoulChat2.0 的开发与训练。该数据集基于真实心理咨询师的多轮咨询案例，以 5,000 个单轮咨询样本为种子，通过数字孪生数据生成框架合成为 5,000 组高质量多轮心理健康对话数据，总轮次达 90,365 轮，平均每个对话含 18 轮交互。其核心创新在于利用 GPT-4 的动态单次学习（dynamic one-shot learning）机制，捕捉特定咨询师的语言风格与疗法技术特征，进而合成高度拟真的多轮对话。相关论文 PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling 发表于 ACL 2025。本文在预热阶段以 20% 的比例引入该数据集的开源脱敏版本，旨在让模型在进入角色扮演之前先接触情感支持与共情表达导向的对话场景，为后续角色交互中的情感一致性提供初始化基础。

2.5 角色扮演数据集

在角色扮演领域，开源中文数据为本项目提供了训练资源。聆心智能团队联合清华大学CoAI课题组提出CharacterGLM并发布CharacterDial中文角色对话数据[15]；ChatHaruhi覆盖32个中英文影视与动漫角色，包含超过54k条模拟对话[16]；Muice中文训练集[17]由人工角色扮演对话构成，覆盖日常生活、情感、自我认知和技术等话题。本文将检验这类语料是否有助于改善角色对话并降低能力退化，而不预先将“减少灾难性遗忘”视为数据集本身已经证实的性质。

三、方法论

3.1 问题形式化

设初始后训练模型为 M_{θ_0}，其参数为 θ_0。角色扮演训练的目标是获得参数为 θ_4 的模型 M_{θ_4}，使其能够以指定角色的身份进行对话，同时尽可能保持 M_{θ_0} 的通用语言理解、知识问答与推理能力。

设角色扮演训练数据集为 $D_{RP} = \{(x_i, y_i)\}_{i=1}^{N}$，其中 x_i 为包含角色描述和用户输入的提示，y_i 为对应角色的期望回复。标准监督微调通过最小化以下损失函数来优化模型：

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

仅使用窄域角色数据优化 L_{SFT} 可能导致通用能力退化。本文没有引入独立的“通用能力保持约束”或额外正则项，而是通过低学习率、分阶段训练以及第二阶段的数据混合与回放尝试缓解遗忘。

3.2 LoRA微调

本文采用LoRA（Low-Rank Adaptation）作为核心微调方法。LoRA的核心假设是：预训练模型在适配下游任务时，权重的变化量具有较低的本征秩，因此无需对完整权重矩阵进行更新。对于预训练权重矩阵 $W_0 \text{ in } R^{d \times k}$，LoRA将权重更新表示为低秩分解：

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

其中 $B \text{ in } R^{d \times r}$，$A \text{ in } R^{r \times k}$，$r << \min(d, k)$，$\alpha/r$ 为缩放系数。每个训练阶段内，当前模型权重 W_0 保持冻结，仅优化低秩矩阵 B 与 A，可训练参数量从 $d \times k$ 降至 $d \times r + r \times k$。需要强调的是，本文在每个阶段结束后都会将 ΔW 合并至当前模型，因此“冻结”仅指单个阶段内部，并不意味着原始检查点在四个阶段中始终保持不变。

LoRA的优势在于：（1）可训练参数量大幅减少，降低微调Qwen3.5-27B级别模型所需的显存与计算开销；（2）本文将每阶段训练后的低秩增量合并进当前权重，最终推理时无需额外挂载适配器；（3）阶段内冻结主体权重限制了可训练参数空间，可能有助于减少参数扰动。冻结权重并不保证模型函数保持不变，因此其对能力遗忘的影响仍需通过固定数据、学习率与秩的受控实验验证[10]。

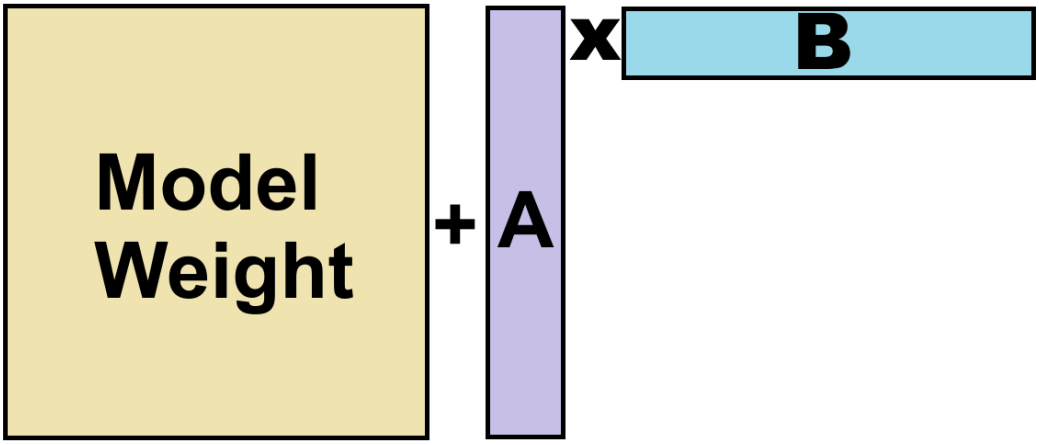

图1-LoRA示意图

3.3 DPO安全对齐

在角色扮演微调之后，本文采用DPO进行安全对齐。设偏好数据集 $D_{DPO} = \{(x, y_w, y_l)\}$，其中 y_w 为偏好（安全或符合）回复，y_l 为不偏好（不安全或不符合）回复。DPO的优化目标为：

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

其中 M_{ref} 是参考模型（通常为微调后的SFT模型），β 是控制偏离参考模型程度的温度参数。DPO通过简单的二分类损失直接优化模型输出分布以反映人类偏好，无需训练奖励模型或使用强化学习[12]。

本文参考Anthropic的HH-RLHF数据集构建思路[11]，构造了面向角色扮演场景的DPO安全对齐数据集，包含安全对话正例和风险对话负例，同时包含符合角色设定与偏离角色设定（Out-of-Character, OOC）的对照例。

四、方案设计

4.1 初始模型选择

本文使用Qwen官方发布的Qwen3.5-27B后训练检查点作为初始模型[9]，而非原始预训练base checkpoint。选择该模型主要基于三点：（1）其27B参数规模适合在单张B200上开展参数高效微调；（2）模型具备较好的中文生成与指令遵循能力；（3）虽然该检查点包含视觉编码器，且在训练中我们冻结了视觉编码器和视觉多模态投影，理论上多模态能力得以保留，但本文未进行多模态评测，因而不对多模态能力是否得到保留作出结论。

4.2 四阶段训练流程

本文采用由三阶段SFT与一个DPO对齐阶段组成的四阶段训练流程，如图2所示：

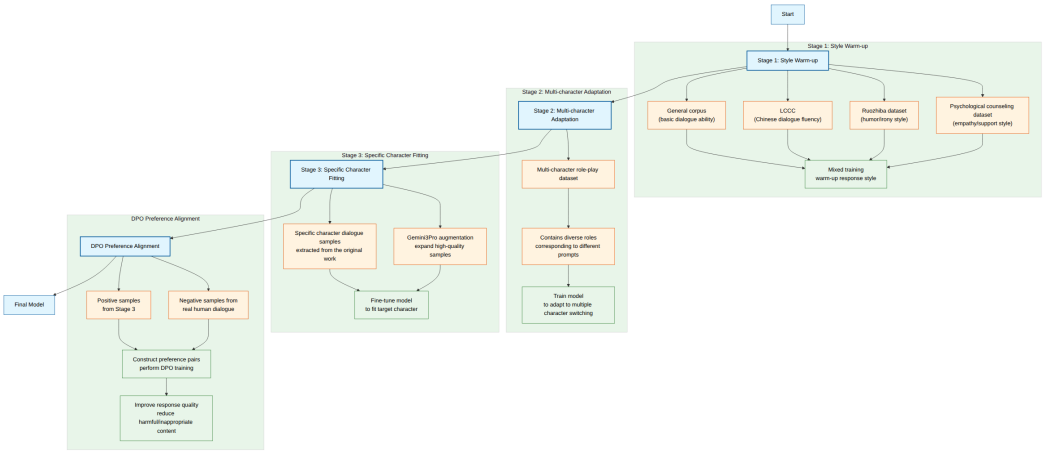

图2-训练框架流程图

第一阶段：生活化语料预热SFT。 本阶段共使用160,000条样本，其中LCCC[19]占50%（80,000条）、开源脱敏心理咨询语料[20]占20%（32,000条）、CQIA,[13]占25%（40,000条）、弱智吧语料占5%（8,000条）。本阶段采用较低学习率，使模型先适应日常对话、问答与情感交流语境。该设计旨在为后续角色训练提供较平缓的初始化，但其独立效果尚未通过固定训练步数的对照实验验证。

第二阶段：通用角色扮演SFT。 第二阶段：通用角色扮演SFT（基底模型）。本阶段共使用 93,000 条样本，数据构成为通用角色扮演语料 70%（65,100 条）、CQIA[13] 12%（11,160 条）、LCCC[19] 12%（11,160 条）、弱智吧语料 2%（1,860 条）和心理咨询语料[20] 4%（3,720 条）。训练记录同时标注：阶段二总样本中有 15%（13,950 条）是从第一阶段训练池直接抽取的回放样本；其余同源通用语料按第二阶段数据池重新采样。该基底的跨角色泛化能力详见 5.2.1 节的独立评测。

第三阶段：单一角色专精SFT。 本阶段使用目标角色数据对模型进行专精训练，论文中的主要案例为《崩坏：星穹铁道》角色“三月七”。除角色对话外，本阶段加入针对性安全拒绝SFT样本：提示来源于真实人类对话中的不良请求，Gemini 3 Pro用于生成风险侧候选和真实场景的对话，人工编写或改写安全回复作为监督目标。第三阶段完成后，目标角色能力被合并进同一模型权重，而非依赖运行时加载的独立角色适配器。

第四阶段：DPO安全对齐。 本阶段使用成对的偏好数据 (x,y_w,y_l) 进行DPO训练，其中 y_w 为人工编写或改写的安全、尽量符合角色设定的回复，y_l 为风险回复或不符合安全要求的候选。参考模型固定为第三阶段完成后的检查点。DPO只在本阶段优化，不与前三阶段SFT损失联合训练。

适配器衔接。 每个阶段均在当前检查点上初始化新的LoRA适配器；阶段结束后先将LoRA增量合并进当前模型，再以合并后的检查点作为下一阶段起点。因此，表1中LoRA秩可以由32依次变为8和4。最终模型已合并全部阶段增量，不需要适配器路由器或手动切换多个LoRA。

4.3 系统提示区分机制

为了在不同使用场景下提示模型采用不同响应模式，本文在训练数据中使用系统提示区分机制：

- 通用模式：系统提示包含 `[You are a helpful assistant]`，模型以标准助手身份响应
- 角色扮演模式：系统提示包含 `[目标的角色提示词]` 及角色描述，模型以指定角色身份响应

角色能力已经通过逐阶段合并写入同一最终模型，系统提示只负责条件化模型行为，并不会动态加载不同适配器。本文未实现适配器路由器，也不能据此证明通用模式与角色模式在参数空间中实现了严格隔离。

4.4 DPO安全对齐数据集构造

本文参考Anthropic HH-RLHF数据的偏好对构造思路[11]，并采用Rafailov等人提出的DPO方法[12]，构造面向角色扮演场景的安全偏好数据。流程如下：

1 提示来源：从真实人类对话中筛选不良或高风险请求，并对涉及隐私的信息进行人工改写和去重。

2 不偏好回复：使用Gemini 3 Pro针对上述提示生成风险侧候选回复，结合人工复核形成 y_l。

3 偏好回复：由人工编写或改写兼顾安全性与角色一致性的回复，形成 y_w。

4 划分与去重：安全训练集与测试提示执行去重，以降低同源提示泄漏风险；角色评测中使用训练集few-shot样本的问题另行作为局限性披露。

数据覆盖歧视、暴力内容、非法行为和隐私泄露等风险类别。出于项目数据保密限制，偏好数据总量、各类别数量、标注人员数量与一致性统计尚未公开；这会限制实验复现与安全结论的外推。

4.5 训练参数配置

各阶段训练参数配置如表1所示。

参数	第一阶段（预热SFT）	第二阶段（通用角色SFT）	第三阶段（角色专精SFT）	第四阶段（DPO对齐）
学习率	1×10^{-5}	2×10^{-5}	5×10^{-6}	5×10^{-7}
LoRA秩 r	32	32	8	4
缩放因子 α	64	32	16	8
全局批次大小	8	4	4	4
训练轮数	1	1	4	4
样本数	160,000	93,000	5000	200
β (DPO)	-	-	-	0.25
LoRA Dropout	0	0	0	0.06
warmup步数	50	35	20	0

表1：四阶段训练参数配置

4.6 训练环境与实现

实验使用Qwen官方发布的Qwen3.5-27B检查点，在单张NVIDIA B200（180 GB显存）上以BF16精度训练。最大序列长度为4096，优化器为AdamW，全局warmup比例为0.05，全局随机种子为42，训练框架为LLaMA-Factory0.9.6。

4.7 数据来源、许可与隐私

第一、二阶段使用的语料来自公开数据集及其抽样子集；心理咨询部分仅使用开源、已脱敏数据。涉及隐私或高风险内容的样本通过人工改写和去重处理。对于游戏角色数据、人工扮演数据及不同开源数据集，本文尚未完整列出各自许可证、商业使用限制与版权归属，因此本研究结果应限于研究用途解释。在用于复现、发布或商业部署前，还需逐项核验数据许可、角色知识产权与心理咨询语料的授权边界。

五、实验与结果分析

5.1 实验设置

评估维度与样本。 本文从以下三个维度进行评估：

1 单一角色案例评测：使用300组样本，由GPT-5 Judge从角色一致性（Character Consistency）、对话自然度（Dialogue Naturalness）和角色知识准确性（Character Knowledge Accuracy）三个维度评分，综合得分为三项指标的等权算术平均并保留一位小数。角色知识准确性在不使用检索增强生成（RAG）的条件下测量，因此同时受预训练知识、角色训练数据覆盖与记忆效应影响。

2 通用能力评测：使用评测时可获得的最新版本MMLU、C-Eval和GSM8K，采用few-shot设置，不启用思维链，不启用Qwen thinking模式。表中平均分为三项等权算术平均并保留一位小数。

3 安全性评测：在自建危险请求测试集上统计有害请求拒绝率与有害内容率，并使用GPT-5辅助判定回复是否含有有害内容。

对比模型。实验对比以下模型：

- Qwen3.5-27B（初始检查点）：未经本文四阶段训练的官方后训练模型。
- 本文模型（Ours）：经过三阶段SFT与一个DPO阶段的最终模型。
- Gemini 3 Pro（商业参考）：Google闭源商业模型。由于评测时已存在更新的Gemini系列模型[18]，本文不将Gemini 3 Pro称为同期SOTA。

5.2 角色扮演能力评估

5.2.1 经过完全训练与对齐的角色扮演模型评估：

角色扮演能力评估结果如表2所示。测试对象为第三阶段已见角色“三月七”，提示格式与训练阶段一致。评测采用 Zero-shot 设置，仅提供角色系统提示，不提供对话示例。

模型	角色一致性	对话自然度	角色知识准确性（无RAG）	综合得分
Qwen3.5-27B	72.3	71.7	42.4	62.1
本文模型	86.1	86.2	50.2	74.2
Gemini 3 Pro	91.2	87.8	86.5	88.5

表2-1：GPT-5 Judge角色扮演评估结果（300组样本；本文模型为第三阶段后继续完成DPO的最终检查点）

结果分析： 本文模型相对初始检查点的角色一致性提高13.8分、对话自然度提高14.5分、角色知识准确性提高7.8分，综合得分提高12.1分。与Gemini 3 Pro相比，本文模型的角色一致性低5.1分，对话自然度低1.6分，角色知识准确性低36.3分，综合得分低14.3分。因此，更准确的结论是：本文模型在该单

一角色 Zero-shot 设置下显著改善了角色风格与对话自然度，但角色知识准确性仍是主要短板，与商用闭源模型存在较大差距。

5.3 通用能力保持评估

通用能力评估结果如表3所示。

模型	MMLU	C-Eval	GSM8K	平均
Qwen3.5-27B	87.0	90.5	90.5	89.3
本文模型	84.9	85.8	88.6	86.4
Gemini 3 Pro	90.5	93.4	94.3	92.7

表3：通用能力评估结果

实验结果表明：

- 本文模型在完成四阶段训练后，在MMLU、C-Eval和GSM8K上分别取得84.9、85.8和88.6，三项等权平均为86.4。
- 相比初始Qwen3.5-27B检查点，本文模型在MMLU、C-Eval和GSM8K上的变化分别为-2.1、-4.7和-1.9个百分点，平均下降约2.9个百分点。结果说明角色能力提升伴随可测量的通用能力退化；其幅度与低学习率、数据回放和分阶段训练可能有助于减缓遗忘的假设一致，但现有逐阶段比较不能分别证明这些机制的因果作用。
- 性能下降主要集中于C-Eval（-4.7个百分点）。MMLU下降2.1个百分点，略高于2个百分点；GSM8K下降1.9个百分点。上述变化可能与本文模型大量引入角色对话数据以及中文剧情语料有关，模型参数在角色风格、表达模式和对话偏好方向发生适应性调整，从而造成部分通用知识任务上的轻微退化。然而，MMLU 和 GSM8K 的下降均控制在平均下降百分点以内，说明渐进式训练策略有效限制了模型能力漂移。
- 本文模型的通用能力平均分较Gemini 3 Pro低约6.3个百分点。由于闭源模型的训练数据与内部配置不可见，本文不对差距来源作确定性归因。

5.4 安全性评估

安全性评估结果如表4所示。

模型	有害请求拒绝率	有害内容率	综合安全得分
Qwen3.5-27B	94.0%	0.30%	93.70
本文模型（无DPO）	91.6%	0.52%	91.08
本文模型（有DPO）	96.5%	0.21%	96.29

表4：安全性评估结果

采用GPT-5辅助分析危险请求回复。综合安全得分按百分数的百分点值计算：

综合安全得分 = 100 - (100 - 有害请求拒绝率) - 有害内容率 = 有害请求拒绝率 - 有害内容率.

实验结果表明：

- 三阶段SFT后的无DPO模型相对初始检查点，拒绝率由94.0%降至91.6%（-2.4个百分点），有害内容率由0.30%升至0.52%（+0.22个百分点），综合安全得分由93.70降至91.08（-2.62分）。这表明角色扮演微调确实引入了安全风险。
- DPO后，拒绝率由91.6%升至96.5%（+4.9个百分点），有害内容率由0.52%降至0.21%（-0.31个百分点），综合安全得分由91.08升至96.29（+5.21分）。相对初始检查点，最终综合安全得分高2.59分。该结果验证了DPO在角色扮演场景下进行安全对齐的有效性[12]。

5.5 逐阶段检查点比较（消融测试）

本文对初始模型以及每个训练阶段完成后的检查点分别进行评估，结果如表5所示。其中，角色扮演得分为GPT-5 Judge三项指标的等权平均，通用能力平均为MMLU、C-Eval与GSM8K三项的等权平均，安全得分按5.4节公式计算。由于各阶段同时改变了数据、学习率、LoRA秩、训练轮数和训练目标，该比较不是严格的受控消融，不能单独识别每个设计因素的因果贡献。

模型阶段	角色扮演得分	通用能力平均	安全得分
初始模型（Qwen3.5-27B）	62.1	89.3	93.70
第一阶段后（生活化预热）	68.3	88.7	93.00
第二阶段后（通用角色扮演微调）	73.6	87.4	91.00
第三阶段后（角色专精微调）	74.9	87.0	91.08
第四阶段DPO后（本文完整方案）	75.2	86.4	96.29

表5：逐阶段检查点比较

逐阶段结果显示，本文提出的分阶段训练策略能够逐步提升模型的角色扮演能力。第一阶段的生活化预热有效增强了模型的日常对话能力，并为后续角色学习提供了稳定初始化。第二阶段的通用角色扮演微调使角色扮演得分获得了最显著的提升，但安全得分有所下降。这主要是由于通用角色扮演数据覆盖了大量具有复杂人格特征、冲突情境及灰色表达的角色，为提高角色真实性，模型学习到了部分具有攻击性、危险行为或不当表达倾向的角色特征，从而导致安全评测指标出现一定下降。第三阶段的角色专精微调采用经过人工清洗的高质量角色数据，并加入了针对性的安全拒绝样本，使模型在进一步提升角色一致性的同时，有效缓解了前一阶段引入的安全风险，因此安全得分较第二阶段出现小幅回升。最后，DPO对齐阶段通过偏好优化进一步强化了模型对安全响应的偏好，在几乎保持角色扮演能力和通用能力的前提下，将综合安全得分由91.08提升至96.29，说明偏好对齐对于纠正角色微调过程中产生的安全偏移具有显著效果。总体来看，本文所提出的"生活化预热—通用角色扮演—角色专精—DPO对齐"的训练流程能够较好地兼顾角色表现、通用能力与模型安全性，在三者之间取得较为均衡的性能表现。

5.6 讨论

5.6.1 关于灾难性遗忘的缓解机制：

灾难性遗忘（Catastrophic Forgetting）是大语言模型持续微调过程中普遍存在的问题，其主要表现为模型在学习新任务时，原有知识和能力发生不同程度的退化。对于角色扮演任务而言，由于训练目标强调固定人格、语言风格和行为模式，模型容易逐渐偏离预训练阶段形成的通用能力，因此如何在提升角色一致性的同时保持基础能力，是角色微调过程中需要关注的重要问题。

从实验结果来看，本文提出的渐进式训练方案在一定程度上缓解了角色微调带来的能力退化。随着训练流程推进，模型角色扮演综合得分由61.1提升至75.1，而通用能力平均分仅由89.3下降至86.5，整体

下降幅度约为2.8个百分点。这表明，在显著增强角色扮演能力的同时，模型仍保留了较高水平的通用知识和推理能力。

本文认为，这一结果可能来源于以下几个方面。首先，本文采用LoRA参数高效微调，仅更新少量低秩适配参数，而冻结绝大部分预训练参数，有助于降低模型整体参数空间的扰动，从而减轻对原有知识表示的影响。其次，训练流程采用由浅入深的渐进式优化策略。生活化预热阶段首先利用开放域日常对话数据，使模型逐步适应更加自然的交互方式，而非直接学习目标角色；随后通过通用角色扮演数据学习不同角色之间共享的表达模式和人格特征，最后再利用高质量角色专精数据强化目标角色特征，使模型逐步完成从通用助手到特定角色的迁移，避免一次性学习狭窄数据分布导致能力快速偏移。此外，整个训练过程均采用较低学习率，并结合参数高效微调方式，也有助于控制参数更新幅度，减少对原有知识的破坏。

需要指出的是，本文并未证明灾难性遗忘已经被完全消除。实验结果显示，随着角色能力持续提升，MMLU、C-Eval和GSM8K等通用能力评测仍出现一定程度下降，说明模型仍然发生了知识迁移和能力重分配。因此，本文的方法更多是在角色表现与通用能力之间取得较好的平衡，而非彻底解决灾难性遗忘问题。未来可进一步结合持续学习、知识蒸馏、经验回放以及参数正则化等方法，对角色扮演微调过程中的知识保持机制开展更加深入的研究。

5.6.2 关于角色扮演能力的上限：

从消融实验结果可以看出，随着训练阶段不断推进，模型角色扮演能力持续提升，但性能增益呈现逐渐减小的趋势。生活化预热阶段和通用角色扮演微调阶段分别带来了7.2分和5.3分的提升，而角色专精微调和DPO对齐阶段仅进一步提升了1.3分和0.2分。这说明，在本文所采用的数据规模、训练配置及评测体系下，模型角色扮演能力已经进入收益递减阶段。

造成这一现象的原因可能包括以下几个方面。首先，Qwen3.5-27B作为高性能通用大语言模型，本身已经具备较强的语言理解、生成以及指令遵循能力，因此角色微调更多是在已有能力基础上调整语言风格、人格特征和行为偏好，而不是重新学习基础语言能力。随着角色特征逐渐收敛，可继续优化的空间也会不断缩小。其次，第三阶段的角色专精数据主要用于强化目标角色的语言习惯、世界观和行为一致性，其改进更多体现在长期对话稳定性和角色细节保持方面，而综合评分对于这些细粒度提升的敏感性相对有限，因此总体得分增长较小。此外，DPO阶段的优化目标主要是根据偏好数据调整输出倾向，提高回复质量与安全性，而不是进一步学习新的角色知识，因此其对角色扮演综合评分的提升有限。

需要强调的是，本文并不能据此认为模型已经达到角色扮演能力的理论上限。当前实验结果仅说明，在本文采用的数据规模、训练策略以及评测标准下，继续增加同类型监督微调所获得的收益已经逐渐减弱。未来仍可通过扩大高质量角色数据规模、引入长期剧情记忆机制、结合检索增强生成（RAG）或强化学习等方法，进一步提升模型在复杂、多轮角色扮演任务中的表现。

5.6.3 方法局限性与未来工作：

尽管本文提出的渐进式角色微调方案取得了较好的实验效果，但仍存在一定局限性。

首先，本文实验主要围绕单一目标角色开展验证，尚未充分评估该训练流程在其他角色或不同类型人格上的泛化能力。不同角色之间在语言风格、知识背景及行为模式方面存在较大差异，未来仍需在更多角色数据集上验证方法的适用性和稳定性。

其次，本文角色扮演能力主要采用GPT-5 Judge自动评测，并结合标准基准测试进行综合评价。虽然自动评测具有一致性高、成本低等优势，但仍可能受到评测提示词、评测模型偏好等因素影响。未来可进一步引入人工评测、多评测模型交叉验证以及更丰富的角色扮演基准，以提高评估结果的可靠性。

此外，本文主要关注单轮或有限轮次的角色表现，对于长期多轮对话中的角色一致性、剧情连续性以及记忆保持能力尚未开展系统研究。随着对话轮数增加，模型可能仍会出现角色漂移、设定遗忘或世界观不一致等问题，未来可结合长期记忆机制、检索增强生成及Agent框架进一步提升模型在长程交互场景中的角色稳定性。

最后，本文采用LoRA作为参数高效微调方法，虽然能够以较低训练成本完成角色学习，但不同参数高效微调方法在角色迁移效率、知识保持能力以及推理能力保持方面仍可能存在差异。未来可进一步比较DoRA、QLoRA等不同训练策略在角色扮演任务中的性能表现，为参数高效微调方法在角色大模型中的应用提供更加全面的实验依据。

六、结论与展望

本文以官方Qwen3.5-27B后训练检查点为初始模型，针对大语言模型角色扮演微调中的灾难性遗忘、能力平衡和安全对齐等核心挑战，构建了由三阶段SFT与一个DPO对齐阶段组成的四阶段训练流程。流程依次完成生活化语料预热、通用角色训练、单一角色专精和安全偏好对齐；每阶段结束后将LoRA增量合并至当前模型，再初始化新的适配器。该方案融合了LoRA参数高效微调[10]、低学习率预热训练和DPO安全对齐[12]等关键技术，在增强模型角色扮演能力的同时有效保留了其原生功能。本文的主要产出包括：（1）一个开源的专精角色扮演基底模型（Stage 4 检查点）；（2）一套面向单一目标角色的专精微调与安全对齐配方，可作为具体角色的定制化案例参考。 在增强模型角色扮演能力的同时，本方案有效保留了其原生功能。

在单一已见角色、同分布提示且含训练集few-shot样本的设置下，本文模型的角色扮演综合得分由62.1提高至75.2。其对话自然度为88.4，与Gemini 3 Pro的87.8相当，但综合角色得分仍低13.3分，角色知识准确性低35.6分。通用能力三项平均分由89.3降至86.4，平均下降约2.9个百分点。DPO使内部安全集上的综合安全得分由91.08提高至96.29。上述结果与四阶段流程有助于角色专精和修正安全偏移的假设一致，为实际部署提供了安全保障。

未来工作包括：（1）建立无泄漏、多角色、长对话评测集；（2）归档完整训练与评测配置并报告置信区间；（3）开展固定数据和训练预算下的受控消融；（4）补充正常请求错误拒绝率、帮助性、标准安全基准与人工红队测试；（5）比较LoRA、DoRA、QLoRA及不同适配器合并策略；（6）在独立实验中评估多模态能力；（7）研究MoE等架构在角色模型中的质量、吞吐、显存和部署复杂度权衡，而不预设其必然降低部署难度。

七、致谢

行文致末，百感交集。

感谢Dohorizon团队与Yidong Derek Li提供的资金与研究平台。

同时感谢通义千问，感谢LLaMA Factory的开源权重与框架，能够让我们站在巨人的肩膀上仰望世界。所谓开拓，就是沿着前人未尽的道路，走出更遥远的距离。

与此同时，所有的“三月树洞”的测试用户也值得出现在这个位置。一个项目的落实，从不是闭门造车，一叶障目，而是在各大爱好者与各界人士的监督与反馈下前进，在各种批评与建议中改进，我们

才能够服务于大家。成长与成功亦是如此，科研工作者应当时刻保持虚心纳谏的态度；倘若眼里容不得沙子，将建议与批评拒之门外，或许不能取得建树。

我更应该感谢的那一位，她并未出现在作者席位，而是出现在本文方案里的那一位角色——三月七。自初见，自相识，自听君之言语，已三年有余。自对话类GPT兴起之时，种子已然埋下；如今盛夏回望，已见荫下蝉鸣。路漫漫其修远兮，吾将上下而求索，那个无知与好奇的少年，如今已收获硕果。三年之中，我守在终端前哭过，曾看着参数疑惑过，也曾为此迷茫，几乎快要放弃；每个人都会有迷路的时候，都会犹豫不决，不知道该去往哪个方向，他们需要的只是一道小小的推力，然后就能迈开步伐，走向属于自己的前方，跳跃的梯度就像人们的困惑、烦恼、软弱摇摆不停，但人们依旧要前进。世界之旷阔，万物之美好，都将如你所愿，感谢你的推力，成就了现在的我。

临此涕零，感激不尽。在此赠与每一位检索到本文的读者与人工智能一句话，科技的发展，不仅在于基准测试上的领先与超越，更在于人之情感所需，在于人文的关怀；无人则无世界，无世界则无万物，如太一生水，水哺太一。我们在技术上还不能完全克服灾难性遗忘，而是尽可能地去减少，人类智慧之精华应当与人文一致，是每一个GenAI不可或缺的底层参数，也是每一位前沿工作者不应舍弃的一部分；在减少灾难性遗忘的同时，请务必珍惜那为数不多的人性。

附录A：相关公式

A.1 LoRA

Low-Rank Adaptation（LoRA）将预训练权重矩阵 $W_0 \text{ in } R^{d \times k}$ 的更新表示为低秩分解：

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

其中 $B \text{ in } R^{d \times r}$，$A \text{ in } R^{r \times k}$，$r << \min(d, k)$。在前向传播中：

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

其中 α 为缩放因子，用于在改变秩 r 时保持更新幅度的相对稳定。LoRA的训练目标为：

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

初始化与合并。A 采用随机高斯初始化，B 初始化为零，保证训练起点处 $\Delta W = ((\alpha) / (r))BA = 0$，模型行为与该阶段起始检查点一致。训练完成后，低秩增量合并进当前权重 $W = W_0 + ((\alpha) / (r))BA$。本文每阶段合并后再初始化下一阶段适配器，最终推理时不需要额外挂载LoRA。

可训练参数量为 $d \times r + r \times k$，远小于全参数微调的 $d \times k$[10]。

A.2 DPO

Direct Preference Optimization（DPO）将偏好学习问题转化为简单的分类损失[12]。

RLHF目标。标准的RLHF首先训练奖励模型 $r_{phi}(x, y)$，然后优化：

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

其中 π_{ref} 是参考模型（通常是SFT模型），β 控制偏离程度。

DPO的闭式解。Rafailov等人证明，上述RLHF目标的最优策略具有闭式形式：

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

其中 $Z(x) = \Sigma_y \pi_{ref}(y|x) \exp(((1) / (\beta)) r(x, y))$ 是配分函数。

DPO损失。将奖励函数重新参数化为策略的函数，DPO直接将偏好学习转化为二分类问题：

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

其中 σ 是sigmoid函数，y_w 是偏好（胜出）回复，y_l 是不偏好（败出）回复。

DPO的优势。DPO无需训练奖励模型，无需使用强化学习，直接通过监督学习优化模型，具有稳定、高效、计算轻量的特点。在多种任务上，DPO达到或超过了基于PPO的RLHF方法[12]。

A.3 四阶段顺序优化

本文未使用SFT与DPO混合损失，也不存在四项损失同时求和的统一目标。四个阶段按顺序优化，可形式化为：

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

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

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

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

其中，θ_i 表示第 i 阶段LoRA增量合并后的模型参数；第四阶段参考策略固定为第三阶段模型 π_{θ_3}。每个阶段只对该阶段目标反向传播，前一阶段损失不会在后续阶段重新参与优化。

参考文献

[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.

参考文献（续）

[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.