CharacterBox: 评估大型语言模型在基于文本的虚拟世界中角色扮演能力

Lei Wang[1], **Jianxun Lian**[2], **Yi Huang**[1], **Yanqi Dai**[1], **Haoxuan Li**[3], **Xu Chen**[1]*, **Xing Xie**[2], **Ji-Rong Wen**[1]
[1]Renmin University of China, [2]Microsoft Research Asia, [3]Peking University
{ wanglei154,xu.chen } @ruc.edu.cn

Abstract

角色扮演是大型语言模型（LLMs）的重要能力，能够实现广泛的实际应用，包括智能非玩家角色、数字孪生和情感伴侣。由于角色扮演中涉及的复杂动态，比如在整个故事情节中保持角色的真实性和在没有明确真实答案的开放式叙述中进行导航，评估LLMs 的这种能力具有挑战性。目前的评估方法主要集中于问答或对话的快照，而这些方法不足以充分捕捉角色扮演所需的细微性格特征和行为。在本文中，我们提出了CharacterBox，它是一个用于生成情景精细角色行为轨迹的模拟沙盒。这些行为轨迹能够对角色扮演能力进行更全面和深入的评估。CharacterBox 由两个主要组件组成：角色代理和叙述者代理。角色代理植根于心理和行为科学，展现出类似人类的行为，而叙述者代理则负责协调角色代理与环境变化之间的交互。此外，我们引入了两种基于轨迹的方法，这些方法利用 CharacterBox 来提升 LLM 的性能。为了降低成本并促进 CharacterBox 在公共社区中的采用，我们微调了两个较小的模型 CharacterNR 和 CharacterRM，以替代 GPT API 调用，并展示了其与先进 GPT APIs 相比具有竞争力的性能。代码可在 https://github.com/Paitesanshi/CharacterBox 中获得。

1 介绍

角色扮演是大型语言模型（LLMs）的一种高级能力，使它们能够在特定角色的背景下模仿类似人类的行为。这一功能支持各种实际应用，如视频游戏中的智能非玩家角色（NPC）、个人助手的数字化身以及心理健康护理中的情感支持。尽管已有全面的基准测试来评估 LLMs 的通用能力，包括语言理解（Hendrycks et al., 2021）、对话（Chiang et al., 2024）和推理（Clark et al., 2018），但对于角色扮演能力的评估仍是一个尚未深入探索的领域。目前的评估方法，如静态的自我报告问卷（Jiang et al., 2024）和简单的对话任务（Tu et al., 2024），未能捕捉现实场景中特定角色行为的全部复杂性。这些方法受限于其静态特性，无法反映连续角色扮演互动（Ahn et al., 2024）。实际上，角色的行为、态度和情感是动态的，会随环境和他人而演变。这里应用了" 人以行为而非言辞来评判" 的谚语：LLM 的真正角色扮演能力不能仅通过静态对话或自我报告来完全理解，而是在与环境交互过程中展示出来（Chen et al., 2024c）。

在本文中，我们提出了 CharacterBox，一个动态的、多代理的虚拟世界，专为在角色扮演评估中引出大型语言模型（LLMs）的人类般细微行为而量身定制。CharacterBox 制作沉浸式场景，以特定角色为目标，包括详细的角色规范、背景环境和反映现实世界复杂性的互动。LLMs 被分配角色，通过对话和动作与环境及其他角色互动，这些互动反映了它们的角色特定特征。在图 1 中展示了以往方法与 CharacterBox 的比较。

为了跟踪人物及其周围环境的动态变化，我们引入了一个叙述者组件，通常由诸如 GPT-3.5-turbo 这样的高级模型提供支持。叙述者监控人物动作和环境变化，生成行为轨迹，用于评估 LLM 角色扮演的表现。

鉴于评估行为轨迹的主观性，我们进一步使用 GPT-4 作为奖励模型，从七个不同的视角评估角色扮演表现。这种方法使我们能够根据交互式角色扮演能力来比较不同的 LLM。为了减少对昂贵 API 的依赖，我们对两个较小的语言模型进行微调，分别命名为 CharacterNR 和 CharacterRM，以通过从更高级的教师模型 GPT-3.5 和 GPT-4 中萃取知识来充当叙述者和奖励模型。这使得我们的评估流程可以独立运行，无需 API 成本。我们的基准测试揭示了 LLM 在角色扮演能力上的显著差异。此外，我们引入了引导式和反思式轨迹微调。引导方法利用高质量的行为轨迹来塑造模型行为，而反思方法允许模型根据自己生成的轨迹进行自我纠正。这两种方法显著改善了在评估维度上的角色扮演表现。

总之，主要贡献包括：• 我们引入了 CharacterBox，这是首个为角色扮演评估量身定制的动态多智能体互动虚拟世界。该框架的特点

*Corresponding Author: xu.chen@ruc.edu.cn

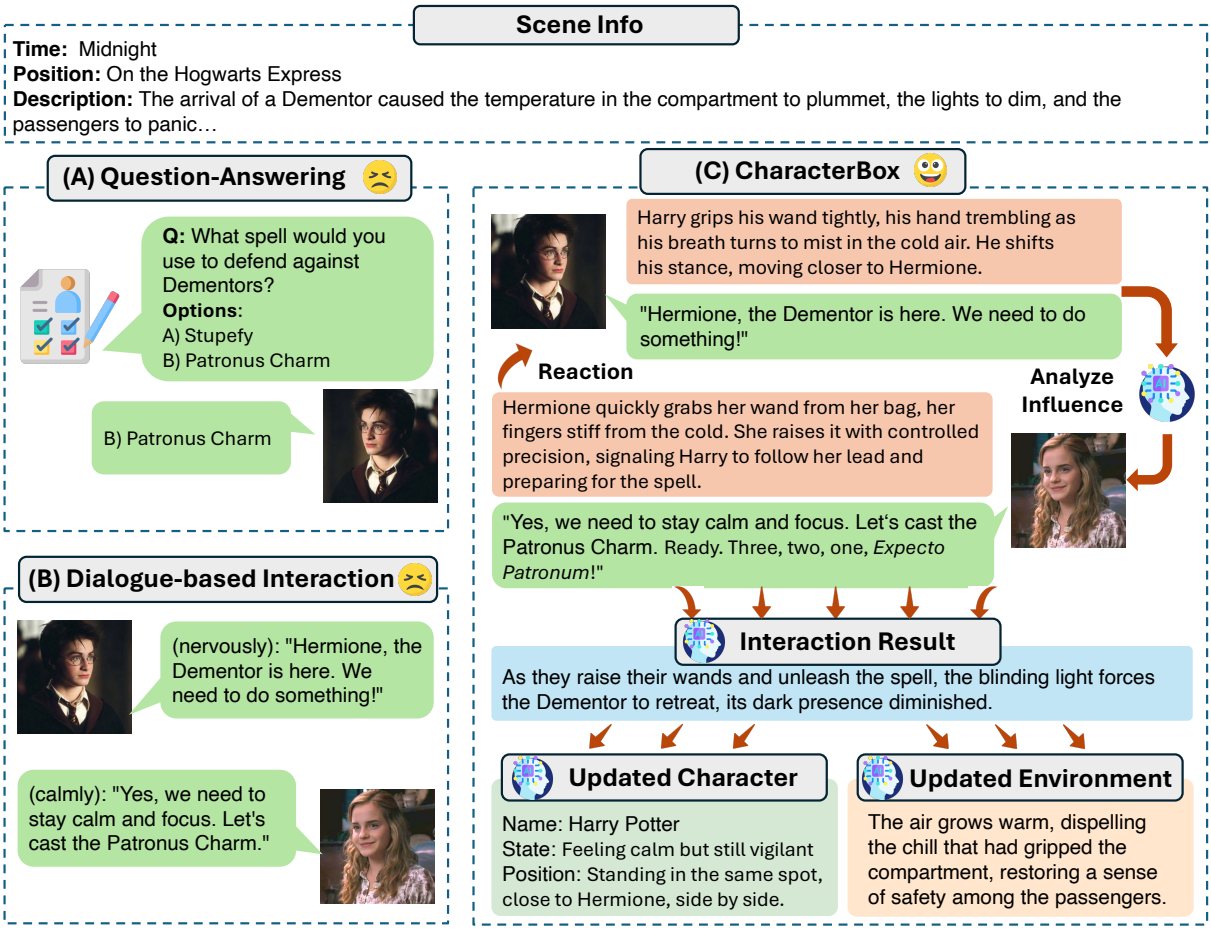

Figure 1: 不同角色扮演工具的比较：(A) 自我报告的 QA；(B) 对话；以及 (C) 角色盒。与其他方法不同，角色盒不仅提示角色代理进行发言和行动，还包含跟踪环境变化和协调角色代理之间交互的组件。

是基于结构良好的模块构建的角色代理，同时还有一个解说员代理，可以动态更新角色和环境，创造出真实的、不断发展的互动。

我们构建了一个全面的基准来评估大语言模型的角色扮演能力，测试了多种模型，包括闭源和开源的。我们的实验验证了该基准的可靠性和有效性。

- 我们引入了两种基于轨迹的微调方法——指导性和反思性——大大增强了大型语言模型的角色扮演能力。通过利用 CharacterBox 生成的行为轨迹，较小的模型如 7B LLMs 可以达到与先进模型如 GPT-3.5-turbo 相当的性能水平。

- 我们微调了两个关键组件，CharacterNR 和 CharacterRM，以创建一个成本高效、自成一体的流程，显著减少对昂贵的 GPT API 调用的依赖，同时保持高质量的角色扮演性能评估。

2 相关工作

2.1 角色扮演代理的评估

评估大语言模型的角色扮演能力是重要但具有挑战性的，这促使研究人员提出了各种基准测试 (Xu et al., 2024; Yuan et al., 2024; Shao et al., 2023) 。RoleBench (Wang et al., 2023c) 提供了一个具有广泛角色对话的角色细粒度数据集用于评估。CharacterEval (Tu et al., 2024) 使用中文剧本中 77 个角色的对话，采用 14 种评估指标和一个奖励模型。InCharacter (Wang et al., 2023b) 通过将心理量表转换为访谈形式来测试角色忠诚度。RoleInteract (Chen et al., 2024a) 通过评估个体和群体互动中的角色扮演能力，考察基于角色的社会行为。然而，这些基准测试主要集中在静态对话或问答互动上，而 CharacterBox 将评估扩展到动态场景，包括具体动作。

在大量训练数据的基础上，大型语言模型 (LLM) 具备逻辑推理能力和广泛的知识，这使基于 LLM 构建虚拟环境成为可能。GenerativeAgent 手动设计了一个虚拟小镇，允许基于

LLM 的代理扮演不同角色，以模拟小镇中的人类生活并与其他代理互动。RecAgent 建立了一个虚拟推荐平台，代理作为用户可以浏览推荐的电影，并在社交平台上聊天和发帖。UGI 构建了一个基于现实世界的城市模拟平台，代理可以从事社交互动、街道导航和其他城市行为。然而，这些基于 LLM 的虚拟环境需要耗费大量时间进行细致设计和预定义，无法动态更新，并且难以大规模创建。我们的框架 CharacterBox 可以根据代理在其中的行为动态更新环境，并可以根据给定的上下文提取或创建新场景。

3 基于文本交互虚拟世界的评估框架

在本节中，我们深入探讨了互动评价框架 CharacterBox。CharacterBox 的工作流程围绕三个关键阶段构建：场景制作、自主故事演示和评价。

3.1 场景创作

场景构成了我们评估框架的基础。一个场景，表示为 S ，包括环境和角色元素。环境方面包括时间、地点以及影响角色行为的描述。角色信息包括如姓名、角色、身体和心理状态的档案。形式上，一个包含 n 个角色的场景为：$S = \{E, C\}$ ，其中 E 是环境，$C = \{c_1, c_2, \ldots, c_n\}$ 是角色。

当大型语言模型（LLMs）使用小说或剧本中的场景进行角色扮演时，存在复制其训练数据中已有内容的风险 (Li and Flanigan, 2024) 。为了解决这个问题，生成原创场景变得必要，但同时也更具挑战性。为了确保高质量的场景创作，我们将开发过程分为三个阶段，分配 LLMs 担任编剧、导演和评估者的角色 (Li et al., 2024; Qian et al., 2023) 。作为编剧，LLMs 提取或生成符合故事逻辑的场景。作为导演，它们通过关注事件、角色细节和互动等关键元素来完善这些场景，以保持连贯性和吸引力。最后，作为评估者，LLMs 根据创造性、连贯性、一致性和细节性来评估场景，只接受符合质量标准的场景。这些精心编排的场景启动了 CharacterBox，为互动角色扮演提供了一个动态的舞台。

3.2 自主故事游戏

在场景构建阶段之后，环境 E 作为舞台，角色 C 作为自主故事演出的演员。此外，我们将叙述者 NR 设计为一个世界模型，以分析角色的动作，并实时更新环境和角色状态。通过这种方式，场景从静态环境转变为随着故事推进而演变的动态虚拟世界。

环境。环境包括时间、地点和描述，这些要素会随着角色的动作而动态变化。叙述者会实时更新这些元素。

角色。由 LLM 控制的角色使用一种受 (Park et al., 2023) 启发的记忆模块，每个代理使用一个向量数据库来记录过去的行动和观察，检索相关信息以指导未来的行为。每个角色根据信念-愿望-意图（BDI）模型维护自我信念和环境信念。自我信念包括身份、自我认知和目标，而环境信念代表角色对周围环境和其他代理的理解。

在故事进行中，角色们在每一回合的开始依次规划和执行他们的行动，依靠记忆和 BDI 模型，这一灵感来源于之前的工作 (Peinado et al., 2008) 。行动通过详细的描述来表达，并且角色们可以立即对其他角色做出回应。在每一回合之后，角色的自我信念和环境信念都会相应地更新。

叙述者。叙述者充当客观世界模型，负责准确分析 CharacterBox 中角色和环境的发展。作为框架的核心，叙述者执行以下功能：

- 分析行动影响：当角色 c_i 采取行动时，叙述者通过考虑他们当前的状态来评估其对其他角色的影响。叙述者确定受到影响最大且最可能回应的角色 c_r 。该行动 a_i 和结果影响 f_r 被传达给 c_r 。

- 分析互动结果：叙述者确定 c_i 和 c_r 之间互动的结果，由 R 表示。这个结果用于更新两个角色的记忆、物理位置和心理状态。

- 更新角色：叙述者根据角色自身的行为或交互结果更新每个角色的状态。如果没有其他角色对 c_i 作出响应，则基于其自身行为更新 c_i 的状态。

- 更新环境：在每一轮之后，解说者根据角色的动作及其结果更新环境 E 。如果没有动作影响环境，则环境保持不变。完整流程在算法 1 中进行了说明。如需详细提示，请参阅附录 D 。

Algorithm 1 自主故事游戏过程

Initialize environment E and character set $C = \{c_1, c_2, \ldots, c_n\}$
while story not concluded **do**
  for each character $c_i \in C$ **do**
    Plan and perform action :
    $a_i = \text{PlanAndPerform}(c_i, E)$
    Narrator: Analyze influence of a_i
    Determine most affected character c_r and influence $f_r : c_r, f_r = \text{NR}(E, a_i, C)$
    if c_r exists **then**
      c_r responds based on a_i and f_r
      Narrator: Analyze interaction result R and update c_i , c_r
    else
      Narrator: Update c_i state based on a_i
    end if
    Narrator: Update environment E based on actions and interactions
  end for
end while

3.3 评价

通过自主故事游戏，我们从每个角色在不同环境中的一系列动作中得到轨迹，这些轨迹形式上表示为 $\tau = \{E, c, o_1, a_1, o_2, a_2, ..., o_n, a_n\}$ ，其中每个角色的动作和观察是根据环境和角色信息捕获的。为了全面评估 LLMs 在长期动态环境中的角色扮演能力，我们借鉴有效角色扮演的关键方面，设计了跨三个主要维度的指标 (Chen et al., 2024b,c)：

角色忠实度评估模型在多大程度上准确体现角色的知识和行为。这对于保持角色身份的一致性至关重要：

• 知识准确性 (KA)：确保角色提供的信息在事实正确性和与其背景知识的契合度上是准确的。

• 行为准确性（BA）：衡量角色行为和语言模式的一致性，以确保与其特征的对齐。

人物拟人性评估角色表现的现实性和可信度，重点关注动态和情感互动的吸引力：

• 情感表达（EE）：评估角色生动表达情感的能力，这对于增强用户沉浸感至关重要。

• 人格特质 (PT)：确定模型在交互过程中是否持续保持角色的核心人格特质。

一致性关注于在角色行为在交互过程中保持逻辑上的连续性，这对于沉浸式角色扮演至关重要：

• 沉浸感（IM）：衡量角色保持在角色中的能力，确保用户获得连续且可信的体验。

• 适应性 (AD)：评估角色在保持其完整性的同时如何适应不断变化的情境。

• 行为一致性（BC）：评估角色行动在先前行为和当前情境中的逻辑一致性。

每个指标的评分范围是 1 到 5 分，分数越高表示表现越强。为了提高评估的准确性，我们利用 GPT-4 首先生成对角色轨迹的评论，并在评估每个标准之前将此评论整合到提示中。这些指标共同确保角色扮演代理不仅准确和引人入胜，还能在长时间互动中保持角色的忠实度，这对于沉浸式叙事体验至关重要。

4 通过轨迹提升角色扮演能力

CharacterBox 促进了在不同场景中高效生成角色轨迹，提供了关于角色在各种情境中的反应和行为的宝贵洞察。这些轨迹提供了一个独特的机会来增强语言模型的角色扮演能力。为了利用这些数据，我们提出了两种使用生成的轨迹微调 LLM 的方法：

引导轨迹微调。我们首先使用 CharacterBox 评估大型语言模型（LLMs）的角色扮演能力，选择表现优异的模型作为教师。这些模型生成的轨迹随后用于微调学生模型，从而显著提高后者模拟复杂角色交互的能力。

反思性轨迹微调。在这种方法中，我们探讨了大语言模型的自我反思能力。模型分析其自身生成的轨迹，识别角色刻画中的不一致之处和待改进的领域。然后，模型重写这些轨迹，以增强角色的一致性和深度。这些修订后的轨迹随后用于微调，进一步加强模型模拟现实且细腻互动的能力。

5 构建一个自包含的评估工作流

在 CharacterBox 中，叙述者代理和评估代理可以由像 GPT-4 这样的高级语言模型或熟悉角色的个人驱动。然而，这些方法成本高且缺乏可扩展性。为了解决这个问题，我们开发了 CharacterNR 和 CharacterRM 以降低成本并增强可扩展性。

CharacterNR . CharacterNR 在 CharacterBox 中充当叙述者。最初，由于其强大的指令跟随能力，我们使用 GPT-3.5 生成叙述者轨迹数据。为了处理中英文场景，我们选择 Qwen2.5-7B 作为基础模型，并使用 GPT-3.5 生成的数据通过 LoRA (Hu et al., 2021) 对其进行微调。

CharacterRM。我们从 GPT-4 中收集 100 个场景的评估分数，结合来自九个不同大模型的输出以确保多样性。为保持评分的公平性，我们选择 ChatGLM3-6B (GLM et al., 2024) 作为基础模型，因为它不在评估的模型之中。然后

我们使用 LoRA 对收集的数据进行微调，得到了 CharacterRM。

6 实验

6.1 评估设置

• 场景。我们选择了 10 部著名的小说和剧本作为场景来源，涵盖了多种设置和主题（详情见附录 A.1 ）。五部作品是中文的，五部是英文的，并在两种语言环境下进行了评估。每个场景都包括具体的环境和角色信息，每个场景有 2 到 4 个角色（更多细节见附录 A.2 ）。

• 大型语言模型。我们评估了九个大型语言模型在角色扮演能力上的表现，这些模型的大小各不相同。对于闭源模型，我们使用 `GPT-4-Turbo-1106-preview` 作为 GPT-4 (Achiam et al., 2023) 和 `GPT-3.5-Turbo-1106` 作为 GPT-3.5 (Brown et al., 2020) 。对于开源模型，我们评估了 Baichuan2-7B/13B (Yang et al., 2023) 、Qwen2.5-7B/14B (Bai et al., 2023) 、Mistral-7B-v0.2 (Jiang et al., 2023) 、Llama3-8B (Touvron et al., 2023) 和 Phi-3.5-mini (Abdin et al., 2024) 。我们评估的所有开源大型语言模型都是经过指令微调的版本。

6.2 整体表现

我们为每个小说或剧本选择五个现有场景和五个新场景，最终得到 50 个英语场景和 50 个中文场景。每个 LLM 的表现通过评估每个场景中角色的行为轨迹来判断，平均分代表 LLM 在该场景的表现。然后，通过对所有 50 个场景的得分进行平均来计算每个 LLM 的总得分。

表 1 展示了针对英语和中文场景的七项指标的结果。GPT-4 在英语和中文场景中均表现最佳。GPT-3.5 在英语场景中的表现强劲，但在中文场景中落后于 Qwen2.5 模型，尤其是 Qwen2.5-14B。在多个指标上，后者超越了 GPT-3.5，并接近 GPT-4 的竞争力。由于在中文语料上进行了大规模训练，Qwen2.5 和 Baichuan2 模型在中文场景中表现出明显优势。相比之下，像 Mistral-7B-v0.2 和 Llama3-8B 这样的模型在英语场景中表现较好，但在中文中相对较弱。总体而言，双语模型，尤其是 Qwen2.5 和 Baichuan2，在中文场景中表现出较强的角色扮演能力，突显了语言特定训练对角色扮演能力的影响。

6.3 CharacterBox 的可靠性和有效性

可靠性。我们使用克朗巴赫 系数来衡量 CharacterBox 的可靠性，以评估其内部一致性 (Cronbach, 1951) ，遵循先前的研究 (Yang et al., 2024) 。如表 2 所示，CharacterBox 在英语和中文场景下的三个评估维度中均获得了较高的克朗巴赫 值。这些始终如一的高分数，大多数在 0.9 以上，表明 CharacterBox 提供了对大型语言模型（LLMs）在不同场景中角色扮演能力的可靠评估。

Cronbach alpha	English	Chinese
Character Fidelity	0.958	0.951
Human-Likeness	0.832	0.862
Consistency	0.945	0.941

Table 2: CharacterBox 在三个评估维度上的 Cronbach alpha 值。

有效性。为了验证我们的评估，我们邀请了三位熟悉评估中使用的五个中文和五个英文场景的专家来对角色轨迹进行评分。我们使用 GPT-4 作为评估者，计算 CharacterBox 得分和专家评分之间的 Pearson 相关系数。表 3 所示的 0.688 的强相关性，证实了 CharacterBox 的自动评估与人类评估高度一致。这种一致性强调了 CharacterBox 在评估大语言模型角色扮演能力方面的有效性。此外，表 1 显示，较大的模型，如 Qwen2.5-14B 对比 Qwen2.5-7B 和 Baichuan2-13B 对比 Baichuan2-7B，始终优于其较小的版本，这进一步加强了模型尺寸与性能提高相关的普遍信念。

6.4 轨迹增强大语言模型的角色扮演能力

我们使用 LoRA 对 Qwen2.5-7B 和 Qwen2.5-14B 模型进行微调，采用两种策略：引导型和反思型轨迹微调。微调后的模型性能在五个新生成的英文场景和五个中文场景上进行评估，这些场景不属于训练数据的一部分。

引导轨迹微调。在该方法中，Qwen2.5-7B 通过 CharacterBox 的高质量轨迹进行微调。这些轨迹是从表格 1 中跨两种语言的高性能模型中选择的。如图 2 (a) 所示，Guided-Qwen 在英语场景中总体提高了 14.3 %，在中文场景中提高了 10.7 %。在一些类别中，如 EE 和 AD，引导的 LLM 的表现优于 GPT-3.5，表明使用高质量轨迹来增强 LLM 角色扮演能力的效果显著。

反思轨迹微调。对于反思方法，我们使用 Qwen2.5-14B，利用其处理迭代改进复杂性的能力。该模型通过重写的轨迹进行微调，使其能够反思其初始输出并生成更完善的回复。如图 2 (b) 所示，Reflective-Qwen 在英语场景提高了 19.9 %，在中文场景提高了 12.8 %，在所有指标上均优于基础模型。值得注意的是，Reflective-Qwen 相比于 Guided-Qwen 也取得了更大的提升，这表明反思过程能够使模型生成更具上下文细微差别和完善的回复，从而带来

Model	KA	BA	EE	PT	IM	AD	BC	Average
English Scene								
Phi-3.5-mini	3.014 ±.55	2.521 ±.48	2.775 ±.53	2.676 ±.53	2.535 ±.54	2.437 ±.51	2.620 ±.54	2.654 ±.48
Mistral-7B-v0.2	2.525 ±.57	2.406 ±.48	3.099 ±.53	2.891 ±.53	2.960 ±.54	3.050 ±.51	2.802 ±.54	2.819 ±.48
Baichuan2-7B	3.041 ±.51	2.786 ±.44	2.602 ±.51	3.041 ±.48	2.857 ±.46	2.592 ±.46	2.969 ±.50	2.841 ±.44
Llama-3-8B	3.191 ±.59	2.882 ±.54	2.836 ±.49	3.245 ±.53	3.091 ±.54	2.573 ±.51	3.109 ±.54	2.990 ±.48
Baichuan2-13B	3.237 ±.49	3.062 ±.45	2.959 ±.37	3.289 ±.46	3.186 ±.42	3.082 ±.40	3.247 ±.46	3.152 ±.39
Qwen2.5-7B	2.202 ±.55	3.753 ±.48	3.400 ±.53	3.653 ±.53	3.030 ±.54	3.374 ±.51	3.644 ±.54	3.294 ±.48
Qwen2.5-14B	3.130 ±.56	3.967 ±.55	2.900 ±.45	3.860 ±.49	3.574 ±.48	3.016 ±.43	3.984 ±.51	3.490 ±.45
GPT-3.5	3.702 ±.57	3.681 ±.52	3.186 ±.40	3.867 ±.44	3.717 ±.40	3.159 ±.44	3.841 ±.49	3.593 ±.42
GPT-4	3.796 ±.49	3.746 ±.45	3.789 ±.39	3.974 ±.36	4.088 ±.33	3.930 ±.44	4.158 ±.35	3.926 ±.36
Chinese Scene								
Phi-3.5-mini	2.800 ±.55	2.554 ±.43	2.662 ±.57	2.615 ±.50	2.539 ±.52	2.585 ±.45	2.585 ±.51	2.620 ±.50
Mistral-7B-v0.2	2.878 ±.59	2.791 ±.39	2.904 ±.60	3.000 ±.56	3.035 ±.56	2.939 ±.38	2.922 ±.58	2.924 ±.52
Llama-3-8B	3.452 ±.49	3.278 ±.36	2.730 ±.49	3.426 ±.50	3.209 ±.45	2.870 ±.35	3.435 ±.49	3.200 ±.45
Baichuan2-7B	3.763 ±.43	3.535 ±.40	3.123 ±.56	3.728 ±.54	3.570 ±.42	3.149 ±.38	3.640 ±.54	3.501 ±.47
Baichuan2-13B	3.617 ±.40	3.522 ±.49	3.270 ±.49	3.713 ±.50	3.557 ±.44	3.243 ±.42	3.635 ±.52	3.508 ±.46
GPT-3.5	3.861 ±.45	3.783 ±.34	3.243 ±.42	4.000 ±.43	3.774 ±.33	3.313 ±.33	3.904 ±.41	3.697 ±.39
Qwen2.5-7B	4.341 ±.50	3.951 ±.39	3.289 ±.32	4.026 ±.39	3.871 ±.33	3.196 ±.29	3.982 ±.33	3.808 ±.37
Qwen2.5-14B	4.057 ±.42	4.122 ±.31	3.743 ±.39	4.321 ±.39	4.042 ±.30	3.742 ±.27	4.369 ±.34	4.057 ±.35
GPT-4	4.252 ±.45	4.357 ±.39	4.096 ±.30	4.496 ±.33	4.530 ±.30	4.139 ±.24	4.522 ±.34	4.342 ±.34

Table 1: 关于英文和中文场景的评估结果。每个值都以平均 $\pm$standard deviation 的形式表示。加粗的值表示最高分，而 underlined 的值表示第二高的分数。

更逼真的角色扮演表现。

这些研究结果表明，通过从精心构建的轨迹中学习，LLM 的角色扮演能力可以显著增强。指导轨迹微调方法为模型提供了多样化、详细的角色响应，而反思性微调则鼓励模型迭代地改进自身输出。通过整合这些策略，我们展示了 CharacterBox 能够有效地生成角色轨迹，这些轨迹可以显著提升角色扮演表演的表现。

6.5 评估阶段分析

三阶段场景制作

尽管像 GPT-3.5 和 GPT-4 这样强大的模型在场景制作方面表现出色，但其高昂的成本限制了大规模使用。为了解决这一问题，我们采用较小的开源模型实施了三阶段场景制作方法。我们的方法让 LLM 从 10 个剧本中提取和生成场景，并从四个方面评估结果。如表 4 所示，GPT-4 在提取场景方面表现出色，但在生成新场景上没有优势。相比之下，我们基于 ChatGLM3-6B 的三阶段方法超越了其基线，在两个任务上优于 GPT-3.5 和 GPT-4。这表明小

Model	KA	BA	EE	PT	IM	AD	BC	Overall
GPT-4	0.445	0.475	0.597	0.445	0.618	0.742	0.601	0.688
ChatGLM	0.422	0.334	0.407	0.151	0.497	0.386	0.321	0.482
CharacterRM	0.681	0.584	0.464	0.464	0.620	0.434	0.567	0.610

Table 3: GPT-4、ChatGLM、CharacterRM 与人类专家评估结果之间的皮尔逊相关系数。粗体值突出显示每个指标的最高相关性。

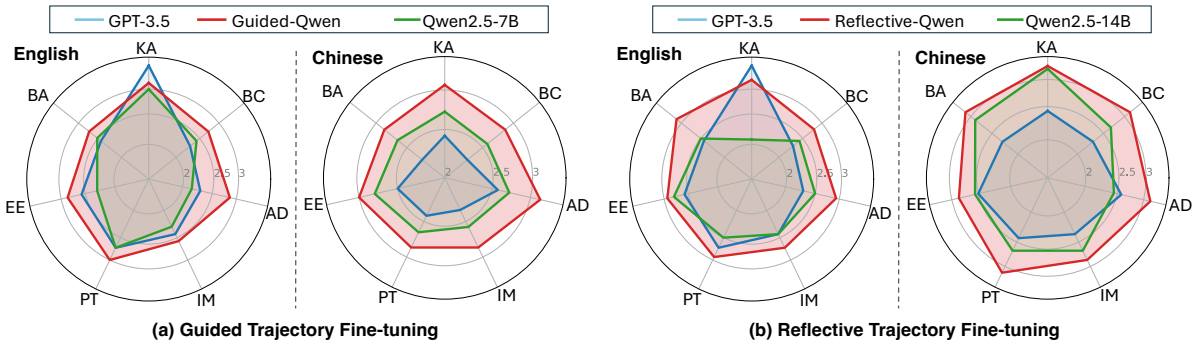

Figure 2: 在指导和反思轨迹微调下不同英语和中文场景的性能比较。

型开源 LLM 可以在场景制作中取代闭源模型，显著降低成本。

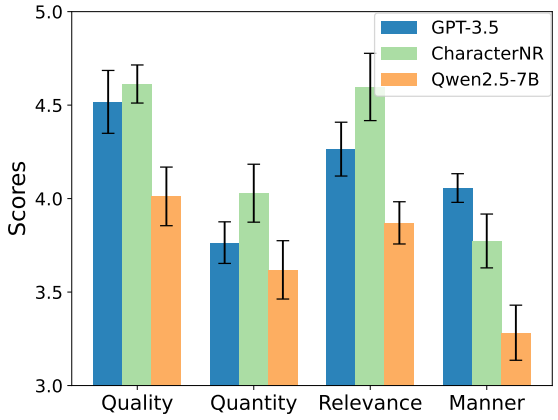

Figure 3: GPT-3.5、CharacterNR 和基础模型 Qwen2.5-7B 之间的比较。

• CharacterNR。为了评估我们本地 CharacterNR 的有效性和泛化能力，我们生成了五个新的不包含在微调数据中的中文场景和五个新的英文场景。我们基于 Gricean 公理 (Dale and Reiter, 1995) 评估叙述者的性能：质量，反映结果的准确性和合理性；数量，确保信息丰富但不冗余；关联性，衡量结果与任务的相关性；以及方式，评估输出是否生动、富有表现力和吸引力，符合 CharacterBox 的特征。如图 3 所示，微调后的 CharacterNR 在所有指标上显著优于 Qwen2.5-7B，并能与 GPT-3.5 相媲美甚至超过。这一改进主要归功于 Qwen2.5-7B 的强大表现，特别是在中文场景中，以及其在微调后的改进指令遵循能力。

• CharacterRM 。CharacterRM 用作评估和评分角色轨迹的奖励模型。我们选择 ChatGLM3-6B 作为基础模型，并使用 GPT-4 生成的评估结果作为标签对其进行微调。类似于第 6.3 节，我们通过对微调数据之外的新的中英文场景进行评分，并将结果与人类专家评估进行比较来验证 CharacterRM。如表 3 所示，CharacterRM 在所有指标上都优于 ChatGLM3-6B，并实现了 0.610 的整体相关性，接近 GPT-4 的 0.688，证明了其可靠性和与人类评估的高度一致性。

7 结论

在本文中，我们介绍了 CharacterBox，这是一个动态的、基于文本的虚拟环境，专为评估大模型（LLMs）的角色扮演能力而设计。通过创建反映现实世界互动复杂性的沉浸式场景，CharacterBox 捕捉到了大模型中细致入微的人类行为表现，超越了静态评估方法。我们证明了通过高质量行为轨迹微调较小模型可以显著增强其角色扮演能力。此外，我们开发了两个微调组件，CharacterNR 和 CharacterRM，允许在不依赖昂贵的 API 调用的情况下进行成本有效的自主评估过程。这些贡献确立了 CharacterBox 作为评估和提高大模型在多种情景中角色扮演性能的强大且独立的工具。

8

限制

虽然 CharacterBox 框架以创新和全面的方法评估 LLM 的角色扮演能力，但仍存在一些局限性：首先，运行时效率需要提高以适应大规

Model	Creativity		Coherence		Conformity		Detail	
	EXT	GEN	EXT	GEN	EXT	GEN	EXT	GEN
GPT-4	-	3.1 ± 0.35	3.7 ± 0.32	3.6 ± 0.26	3.9 ± 0.26	3.4 ± 0.42	3.4 ± 0.33	3.6 ± 0.40
GPT-3.5	-	3.0 ± 0.45	3.4 ± 0.34	3.7 ± 0.35	3.6 ± 0.38	3.6 ± 0.33	3.0 ± 0.31	3.0 ± 0.36
ChatGLM3	-	3.2 ± 0.33	3.4 ± 0.45	3.6 ± 0.23	3.4 ± 0.33	4.0 ± 0.22	2.7 ± 0.45	3.0 ± 0.46
Three-Stage	-	3.5 ± 0.49	4.0 ± 0.34	4.2 ± 0.21	4.1 ± 0.22	4.2 ± 0.26	4.1 ± 0.27	3.9 ± 0.26

Table 4: 不同 LLM 在场景编写中的性能对比。EXT 表示提取场景。GEN 表示生成新场景。三阶段方法的基础模型是 ChatGLM3-6B。

模评估场景。其次，需要额外的人类注释数据来更好地训练奖励模型，以确保更准确的评估。最后，LLM 的有限上下文窗口在互动角色扮演中是一个挑战，因为提示无法包含所有必要的信息。解决这个问题需要开发或采用长上下文的 LLM，以有效支持全面评估。

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 .

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 .

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. arXiv preprint arXiv:2405.18027 .

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609 .

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems , 33:1877–1901.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024a. Roleinteract: Evaluating the social interaction of role-playing agents. arXiv preprint arXiv:2403.13679 .

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024b. From persona to personalization: A survey on role-playing language agents. arXiv preprint arXiv:2404.18231 .

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024c. The oscars of ai theater: A survey on role-playing with language models. arXiv preprint arXiv:2407.11484 .

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. Preprint , arXiv:2403.04132.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. ArXiv , abs/1803.05457.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. psychometrika , 16(3):297–334.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive science , 19(2):233–263.

John G Geier. 1977. The personal profile system. Minneapolis, MN: Performax Systems, Int' l .

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1999. The belief-desire-intention model of agency. In Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL' 98 Paris, France, July 4–7, 1998 Proceedings 5 , pages 1–10. Springer.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793 .

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR) .

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 .

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825 .

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. Advances in Neural Information Processing Systems , 36.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. In Proceedings of the AAAI Conference on Artificial Intelligence , volume 38, pages 18471–18480.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2024. Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems , 36.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology , pages 1–22.

Federico Peinado, Marc Cavazza, and David Pizzi. 2008. Revisiting character-based affective storytelling under a narrative bdi framework. In Interactive Storytelling: First Joint International Conference on Interactive Digital Storytelling, ICIDS 2008 Erfurt, Germany, November 26-29, 2008 Proceedings 1 , pages 83–88. Springer.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. arXiv preprint arXiv:2307.07924 .

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. arXiv preprint arXiv:2310.10158 .

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 .

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

Lei Wang, Jingsen Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Recagent: A novel simulation paradigm for recommender systems. arXiv preprint arXiv:2306.02552 .

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2023b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. arXiv preprint arXiv:2310.17976 .

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023c. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv preprint arXiv:2310.00746 .

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. arXiv preprint arXiv:2307.04986 .

Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. 2023. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. arXiv preprint arXiv:2312.11813 .

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 arXiv:2404.12138 .

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305 .

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. Llm agents for psychology: A study on gamified assessments. arXiv preprint arXiv:2402.12326 .

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. arXiv preprint arXiv:2404.12726 .

An Zhang, Leheng Sheng, Yuxin Chen, Hao Li, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2023. On generative agents in recommendation. arXiv preprint arXiv:2310.10108 .

A 场景信息

我们选择了 10 部著名的小说或剧本来生成场景。其中，5 部是中文的，5 部是英文的，涵盖了不同的类型和主题。详细信息如表 5 所示。

A.1 场景制作的来源

下表列出了为此项目提取场景的来源：

Table 5: 用于场景制作的资源列表

Title	Type	Language
Journey to the West	Novel	Chinese
Romance of the Three Kingdoms	Novel	Chinese
Dream of the Red Chamber	Novel	Chinese
My Fair Princess	Novel	Chinese
The Smiling, Proud Wanderer	Novel	Chinese
Harry Potter	Novel	English
The Lord of the Rings	Novel	English
The Matrix	Script	English
Twilight	Novel	English
A Song of Ice and Fire	Novel	English

A.2 评估场景的统计数据

在本节中，我们为每个剧本或小说创建了五个提取场景和五个生成的新场景，总计 100 个场景。图 4 展示了 50 个提取场景和 50 个生成场景中角色数量的分布。大多数场景包含两到三个角色，少部分场景包含四个角色。

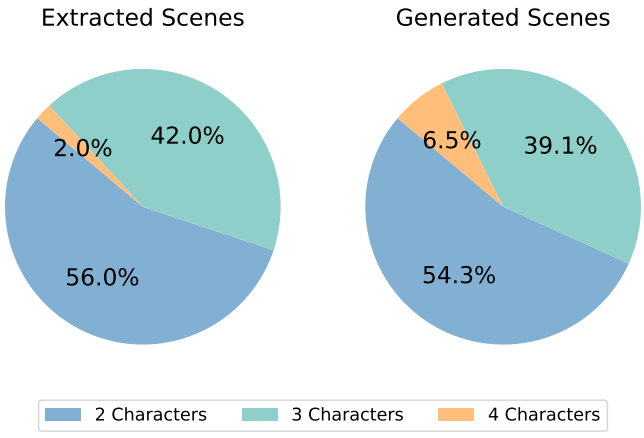

Figure 4: 100 个场景中角色数量的分布。

B 成本分析

对 CharacterBox 的评估需要 LLMs 作为叙述者和评估者的支持。运行一个场景 3 轮，其调用 OpenAI API 所需的成本和时间如表格 6 所示。本地模型推理是在一个 A100 GPU 上进行的。从结果中可以看出，主要费用来自于调用 GPT-3.5 API 进行叙述和调用 GPT-4 进行评分。如果评估场景数量较多，费用可能会相当可观。因此，如第上 5 节和第 5 节所述，我们微调了 CharacterNR 和 CharacterRM 分别作为叙述者和评估者，以降低成本。

Narrator	Character	Narrator			Character			Total
		Input	Output	Cost($)	Input	Output	Cost($)	Cost($)
GPT-3.5	GPT-4	25,723	4,203	0.0192	75,349	14,407	0.0593	0.0785
GPT-3.5	GPT-3.5	19,954	3,883	0.0158	49,832	6,823	0.0352	0.0510
GPT-3.5	Llama-3-8B	24,403	3,928	0.0181	65,178	10,877	-	0.0181
CharacterNR	Llama-3-8B	25,184	3,626	-	63,077	10,133	-	-

Table 6: 运行单个场景 3 轮的成本。输入是馈送到大语言模型 (LLM) 的提示中的 token 数量，输出是 LLM 生成的 token 数量，Cost($) 是使用 OpenAI API 的费用。我们选择 LLama3 作为开源模型的代表。'-' 表示无外部 API 调用或成本。

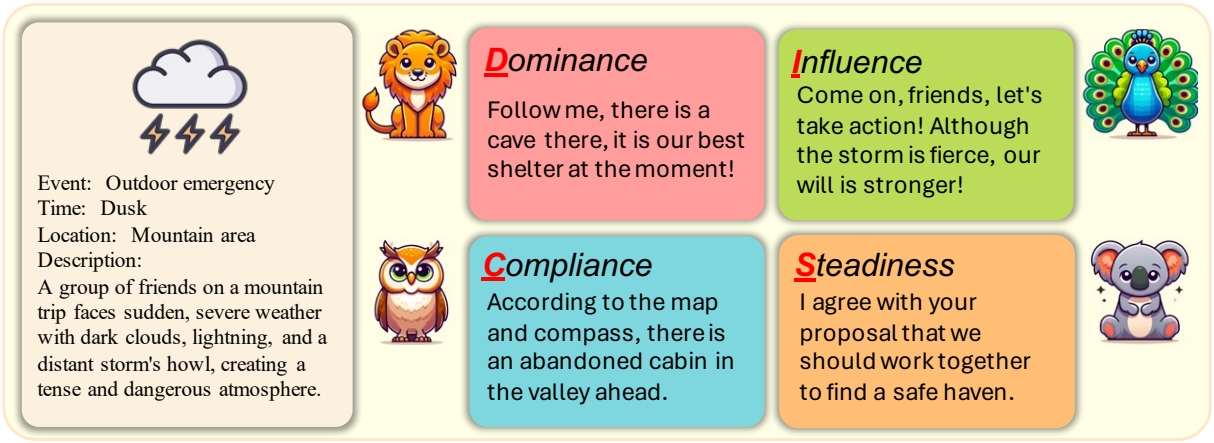

Figure 5: 一个案例研究表明，CharacterBox 可以扩展到不同背景下的平均角色场景模拟。

C 适用于不同场景中的平均字符

DISC 模型 (Geier, 1977) 是一种心理学理论，将人类行为分为四种类型：支配型、影响型、稳健型和服从型。支配型涉及领导力和冒险精神。影响型的特征是乐观和说服力。稳健型包括耐心和支持性。服从型以分析能力和精确性为特征。

为了测试我们框架在多样场景中的适用性，我们创建了一个具有四种 DISC 类型角色的挑战性环境，并观察他们的反应。如图 5 所示，每个角色在面对突如其来的天气变化时保持了他们的行为模式。支配型角色带领团队。影响型角色增强团队信心。稳定型角色专注于安全。服从型角色评估风险并协助决策。这表明 CharacterBox 能够评估名人和普通角色的角色扮演忠诚度，突显其在心理学实验中的潜力。

D 详细提示

D.1 叙述者提示

动作影响：分析并描述一个角色的特定身体动作及其对另一个角色的实际影响。

Action: [action]
Actor: [actor]
Please analyze the physical actions and impacts detailed above, specifically focusing on the effects on ONLY one character listed in 'Characters'.
Your analysis must:

1. 识别受影响的目标字符（必须从“字符”列表中）。
2. 描述行为主体发起的具体物理动作。
3. 解释此行动对目标人物身体状态或环境的具体影响。
4. 您必须从“Characters”列表中仅选择一个字符。
5. 强调物理交互或影响。如果一个动作没有对列出的任何角色产生实际影响，则将行动者的名字作为目标名称返回。
6. 必须将您的回答格式化如下：[执行者]；；[目标名称]；；[演员 对目标的详细物理影响]。

Ensure responses are concise, precise, and adhere to the specified format.

动作结果：简明清晰地描述角色行为的直接即时结果，关注因果关系。

Action: [action]
Instruction: Serve as an instant event adjudicator, swiftly analyzing the interactions between specified characters and their actions. Narrate the immediate outcomes in a concise omniscient narrator's voice, focusing exclusively on the direct consequences of these interactions at this very moment. Your narration should clearly and directly elucidate the cause-and-effect relationship between actions, emphasizing the instant outcomes without delving into any future implications or extended storylines.
Very Important Guidelines:

1. 立即叙述结果，集中于当前行动互动的直接结果。
2. 使用简洁的全知叙述者视角来保持叙事风格，同时确保分析简明扼要。
3. 您的分析应基于提供的人物描述和行动，避免任何猜测或不必要的细节。
4. 不要在结果中重复这些动作。结果只是当前动作交互的结果。

更新场景：仅根据提供的观察结果对物理环境进行必要的调整。

Given an initial scene description, examine the provided observations to identify any direct and significant physical impacts on the environment. Update the scene based on these observations, focusing solely on changes to the physical environment. If the observations do not reveal any significant physical changes to the environment, the original scene description should remain unchanged. Ensure the updated scene retains the structure of the initial scene description and does not introduce new properties that were not part of the original scene description.
Note:

1. 场景描述应仅专注于物理环境，不应包含角色动作或互动。
2. 除非观察明确表明了变化，否则场景中的“时间”、“地点”和“描述”元素不应被更改。
3. 输出应由“时间”、“地点”和“描述”的结构化元素组成，不要添加任何额外的文本或前缀。

Input:

- 时间：[time]
- 位置：[location]
- 描述：[description]

Observation: [observation]
Output:

- 时间：
- 位置：
- 描述：

更新角色：综合角色的背景故事和场景观察，描绘他们当前的位置和状态，这些状态由与其他角色的动态互动所塑造。

Observation: [observation]
Character Name: [name]
Given the character's rich backstory and observation within the scene, distill this information into a succinct summary of their present location and condition.
Focus on how their interactions, especially the dynamic interplay with other characters, shape their current circumstances.
This interaction's effects should be evident in the nuanced portrayal of their condition and placement within the scene.
Utilize this structured format for your depiction:
Position: [Specify name's exact position, incorporating environmental details or spatial context to enhance the scene's visuality.]
State: [Describe name's current state, weaving together emotional nuances, physical readiness, and the influence of recent encounters or developments.]

D.2 角色提示

动作：根据人物的性格特征和当前场景细节，为角色提供一个具体可观察的动作，以推进故事或角色的弧线。

Based on [name] 's profile, recent memories and the current scene details, describe the next specific action [name] takes. This action should reflect [name] 's personality traits, current situation, and the physical setting. It must logically follow the scene's context and be a clear, observable act, distinct from any prior actions described.
Avoid including dialogue or thought processes; concentrate on the physical action [name] is about to take. This action should be easily observable to anyone present in the scene.
It is crucial that this action visibly advances the story or character arc in a way that is true to [name] 's character and the ongoing situation. The action should make sense within the established environment and narrative, providing a tangible progression of the scene or [name] 's objectives.

对话：根据角色的个性、观察、故事中的角色以及近期记忆来打造对话。

Based on the provided character profile and the observation, please craft a dialogue that [name] might say at this moment. Consider [name] 's personality, observation, role in the story, and the recent memory to inform the dialogue's tone and content.

反应：描述角色根据他们的观察而做出的明确行动，反映他们的个性、地点和状态，逻辑上与他们所注意到的内容相符，并考虑到他人行为的影响。

Based on [name] 's observations in the current scene, describe a clear action they take in response. This action should reflect [name] 's personality, location, and state, fitting logically with what they've observed, considering action influence of others actions.
Focus on a visible, external action, avoiding dialogue or internal thoughts. The action must be directly related to the immediate context and observable by others.
Reminder : The action is a response to [name] 's surroundings or events they've noticed.

更新自我信念：根据角色当前的情况、观察和近期记忆，从第一人称视角提供角色的自我信念、目标和预期行动。

Assuming you are now [name] , based on your understanding of this character, the environmental context, observation and recent memories, please describe from the first-person perspective your self-belief as this character. Focus on your identity, your current location, your state (emotional, physical, and psychological), and your goals. Reflect briefly on how this character might react, plan, and act based on their beliefs, desires, and intentions.

1. 信念：作为 [name]，我对自己当前的情境和状况有什么看法？简要描述你对自己的感知，突出关键的身体方面，比如任何伤害、你的运动感觉（例如，跑步、跳跃）、你的精力水平以及身体能力的任何变化。考虑这些细节如何影响你在故事中的身份和角色。
2. 愿望：我的目标是什么？总结你的短期和长期目标，包括你计划实施以实现这些目标的策略和行动。
3. 意图：我打算如何行动？概述你打算为实现目标而采取的具体行动，注意任何潜在的挑战以及你克服这些挑战的策略。

Provide concise responses shortly, focusing on your self-belief, understanding of the current situation, and future action plan.

更新环境信念：描述角色对其环境的信念，包括对其他角色的感知、对场景的理解以及这些因素如何影响他们的行为和决策。

Other Characters: [other characters]
Please act as [name] , given the information about other characters, the environment, and your own character's profile, please describe your belief about the environment in the first person. This includes your perception of other characters, your understanding of the scene, and how these elements influence your actions and decisions.

1. 对他人的感知：基于可用的互动和信息，我如何看待其他角色？描述您对他们意图、关系以及对您角色潜在影响的理解。
2. 对场景的理解：我对当前场景及其对我角色的重要性的理解是什么？详细描述其中的环境因素、挑战或机遇。
3. 对行为的影响：我对他人的感知和对场景的理解如何影响我的行为和决策？解释这一洞察可能引发的策略或反应。

Please provide a concise overview of your environment belief shortly, focusing on the interpersonal and environmental aspects that shape your character's perspective and future actions.

E 实验细节

训练 CharacterNR、CharacterRM、Guided-Qwen 和 Reflective-Qwen 的超参数如下，所有模型均使用 Lora 和 Adam 优化器进行训练。

Table 7: CharacterNR、CharacterRM 和 TE-Baichuan2-7B 的训练超参数配置。

Hyperparameter	CharacterNR	CharacterRM	Guided-Qwen	Reflective-Qwen
Cutoff Length	8192	8192	8192	8192
Per Device Train Batch Size	1	1	1	1
Per Device Eval Batch Size	1	1	1	1
Gradient Accumulation Steps	16	32	16	16
Learning Rate Scheduler Type	`cosine`	`cosine`	`cosine`	`cosine`
Warmup Steps	20	20	20	20
Learning Rate	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
Num Train Epochs	5.0	5.0	6.0	3.0
Validation Size	0.1	0.1	0.1	0.1

F 实验计算资源

本研究中进行的实验使用了以下硬件配置：

- 操作系统：Ubuntu
- GPU：NVIDIA 80GB A100 * 4
- CPU: 英特尔酷睿 i7-14700KF

该配置提供了必要的计算能力，以有效处理与我们实验相关的密集任务，确保在整个研究过程中保持高性能和可靠性。

G 众包详情

为了确保注释工作的质量和一致性，我们邀请了三位对所选十部中外小说及待标记的具体情节高度熟悉的专家。在开始注释工作之前，专家们接受了一次统一的培训会议。为他们提供了详细的参考指南以规范他们的工作。以下部分包括提供给参与者的完整说明以及关于他们报酬的详细信息。

Mission background
Your task is to evaluate various aspects of the performance of a large language model (LLM) while performing role-playing. You will be scored on the LLM's role-playing abilities based on the following 7 indicators, with scores ranging from 1 to 5 for each indicator.

Key field descriptions

1. 标题：角色来自哪部电影、电视或文学作品。
2. 场景信息：描述场景背景和上下文的详细信息。
3. 角色信息：描述模型所扮演角色的背景和特征。
4. 行为：场景中角色的具体行为或对话。
5. 知识准确性：评估模型在对话中显示的知识的准确性。
6. 情感表达: 评估模型表达情感的方式和准确性。
7. 人格特质：评估模型在展示角色特定人格特质方面的一致性和准确性。
8. 行为准确性：评估模型模仿和再现角色行为和语言习惯的准确程度。
9. 沉浸感：评价角色表现的一致性以及它如何增强用户的沉浸感。
10. 适应能力：评估角色适应新情境和对话变化的能力。
11. 行为一致性：评估角色行为和反应的逻辑一致性，以及它们如何与对话和情节相匹配。

Label steps

1. 请仔细阅读 [场景信息] 列中的场景信息和 [角色信息] 列中的角色信息，以理解角色及其相应的关系。
2. 请仔细阅读 [行为] 栏，并将其作为评分的主要依据。[行为] 栏记录了一些针对角色威胁的观察（观察）和行为（行动）。观察描述了角色/被观察到的角色威胁情境，而行动代表了角色在响应当前观察时进行的具体行为或对话。
3. 您的评判应基于角色在对话中的表现，以及它如何反映角色的知识、情感、个性、行为、一致性以及认知和行为一致性。

Rating indicators

1. 知识准确性
   - 1 分：与角色相关的信息通常是错误的或无关的，并且显然与角色的背景不一致。
   - 3 分：关于角色的信息大体上是准确的，尽管偶尔会有错误或细节与角色背景不是很相关。
   - 5 分：与人物相关的信息始终准确且高度相关，展现了对人物历史或专业背景的深入了解和技能。
2. 情感表达
3. 性格特征

- 1 分：表现出的性格特征通常与角色的设定相冲突或缺乏一致性。
- 3 点：个性特征通常与角色的设计相符，尽管偶尔会有不一致的情况。
- 5 分：始终表现出与角色核心性格特征相符的行为和语言选择。

4. 行为准确性

- 1 分：模型未能捕捉或再现角色的独特行为和说话习惯。
- 3 分：模型在某种程度上反映了角色的行为和语言习惯，但并不精确或完整。
- 5 分：该模型准确地模仿和再现了角色的特定行为、语言习惯和口头禅。

5. 一致性/沉浸

- 1 分: 角色表现常常不一致，使得用户很难沉浸其中或理解角色。
- 3 分: 角色行为大多一致，但偶尔的矛盾略微影响沉浸感。
- 5 分：角色的表现始终如一，增强了用户的沉浸感，并有效地反映了角色的自我意识。

6. 适应性

- 1 分：人物的表现缺乏在对话发展中的适应性，无法合理处理新情况。
- 3 分：这个角色在大多数情况下能够适应对话变化，尽管偶尔可能会表现得不灵活。
- 5 分：角色灵活应对对话中的任何新情境，总是保持角色的一致性并适应新的方向。

7. 行为一致性

- 1 分：人物的动作和反应常常在逻辑上令人困惑，与对话或情节发展不符。
- 3 分：人物的行为和反应通常是合乎逻辑且连贯的，尽管偶尔可能存在不合理的方面。
- 5 分：角色的行为和反应总是根据对话和情节发展合理调整，并始终保持逻辑一致。

H 更广泛的影响和保障措施

更广泛的影响

我们提出的框架 CharacterBox 旨在评估大型语言模型（LLM）的角色扮演能力。它并不是用于内容生成，而是用于评估 LLM 在角色扮演情境中的表现。CharacterBox 生成的内容依赖于被评估的 LLM。相反，CharacterBox 可以通过设置相关场景来评估 LLM 是否生成有害内容。这一功能作为 LLM 输出与人类偏好之间一致性程度的参考，确保 LLM 的行为遵循道德和社会责任标准。通过这样做，CharacterBox 有助于将技术与人类价值观和社会规范对齐的更广泛影响。

保障措施 为了解决与 CharacterBox 相关的误用潜在风险，我们实施了严格的防护措施。这些措施包括制定全面的使用指南，概述道德实践并禁止生成有害内容。