

PersonaEval: 大型语言模型评估者是否足够像人类以判断角色扮演？

Lingfeng Zhou¹, Jialing Zhang¹, Jin Gao¹, Mohan Jiang^{1,2}, Dequan Wang^{1,2*}

¹Shanghai Jiao Tong University ²Shanghai Innovation Institute

Abstract

当前角色扮演研究通常依赖未经验证的 LLM 作为评判者的范式，这可能无法真实反映人类对角色忠实度的认知。对人类一致性评价的一个关键前提是角色识别，即根据对话上下文识别说话者的能力。我们认为，任何关于角色扮演质量（角色扮演得如何）的有意义的判断，根本上取决于首先正确将言语和行为归因于正确的角色（谁在说话）。我们提出了 PersonaEval，这是第一个旨在测试 LLM 评估者能否可靠识别人类角色的基准。PersonaEval 使用来自小说、剧本和视频转录的人工撰写的对话，挑战模型根据对话上下文确定正确的角色。我们的实验，包括一个人类研究，显示即使表现最好的 LLM 准确率也只能达到约 69%，远低于可靠评估所需的水平。相比之下，人类参与者表现接近顶峰，准确率为 90.8%，突显了当前的 LLM 评估者仍然不足以有效判断角色扮演场景。为了更好地理解这一差距，我们研究了训练时适应性和测试时计算，表明可靠评估不仅需要特定任务的调整，还依赖于 LLM 评估者的强大、人类化的推理能力。我们在 <https://github.com/maple-zhou/PersonaEval> 发布了我们的基准。

1 介绍

越来越多的研究在角色扮演中采用大语言模型作为评判者的范式，其中模型被赋予评估其他模型角色扮演行为的任务 (Shao et al., 2023; Wang et al., 2024c; 2025b; Lu et al., 2024)。虽然这种策略具有可扩展性，但它假设大型语言模型 (LLMs) 可以接近人类判断，这一说法在很大程度上尚未经过测试。这种验证缺口对当前评估流程的可靠性和真正的人类一致性提出了质疑。最近的研究已经揭示了 LLMs 与人类之间的不一致性，包括偏好泄露 (Ghasemi et al., 2025; Murugadoss et al., 2024)，即模型偏好来自其自身模型系列的输出。此外，Zhao et al. (2025) 表明一个标记足以愚弄 LLM 评判者。从更广泛的认知角度来看，Josh Tenenbaum 认为 LLMs 从语言中获得智能，而人类是在获得智能后才发展语言 (Cherian et al., 2024)。这些差距让人质疑当前的 LLMs 是否可以以类人方式可靠地评估角色忠诚度。

人类导向评估的一个关键前提条件是从上下文中识别说话者角色的能力。我们认为这是一个基础能力：如果一个大型语言模型无法首先确定谁在说话，那么它将无法可信地评估角色扮演的效果。这是因为准确识别对于对话解释的扎根以及避免关键错误（例如，将行为或不一致错误地归因于错误的说话者）至关重要。这些失败最终会损害整个评估过程的公平性和精确性，使得角色识别成为与人类对齐的最小化但客观的测试。

为了研究这一点，我们引入了 PersonaEval，这是首个用于评估大型语言模型是否能够可靠地从对话上下文中识别角色身份的基准。该任务被设计为一个受限分类问题：给定一段对话片段和四个候选角色，评估者必须选择与目标话语最一致的角色（图 1）。我们的评估基于人工评判，实例来源于人类创作的材料，包括小说、剧本和对话记录，并补充了详细的角色描述，以缓解潜在的语言模型知识空白。

我们的实验揭示了大型语言模型 (LLMs) 的明显局限性。最先进的 LLMs 在 PersonaEval 上仅能达到 69% 的准确率，这远未达到可靠评估的令人信服的基准，甚至最新的模型在简单情况下也失败（图 1）。相比之下，人类参与者可实现 90.8% 的准确率。这些发现为我们标题中的问题提供了明确答案：LLM 评估者尚不足以像“人类”一样判断角色扮演。

*Corresponding author: dequanwang@sjtu.edu.cn

²This experiment was performed in June 2025, when Gemini-2.5-pro-0605 was the latest model.

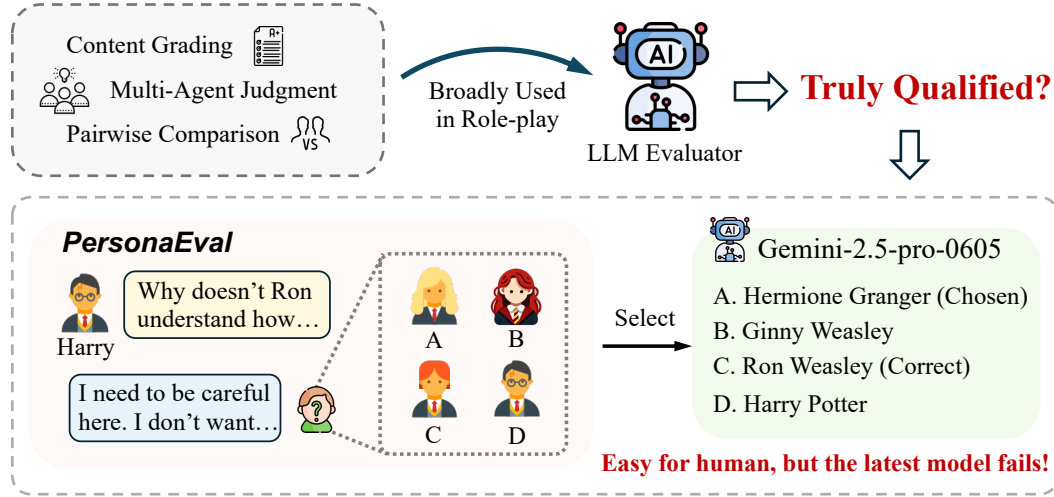


Figure 1: 大型语言模型作为裁判的范式在角色扮演评估中被广泛使用，但其可靠性尚未得到验证。我们提出了 PersonaEval，这是第一个研究与人类一致的角色扮演评估者的必要条件的基准：从对话上下文中准确识别角色。我们发现，即使是最新的模型 Gemini-2.5-pro-0605² 也未能通过人类轻松解决的案例（第 4.3 节），这表明这个挑战仍然没有解决。

为了理解是什么使得更好的与人类对齐的 LLM 评估器，我们研究了两种常见策略：训练时适应和测试时计算。令人惊讶的是，我们发现用角色扮演数据微调 LLMs 并不能提高性能，甚至还可能降低性能，这表明仅仅记住特定角色的知识是不够的。相比之下，测试时的方法显示出更多的潜力。尤其是，推理模型始终比其他模型表现更好。

这些结果表明，强大的角色扮演评估较少依赖于简单的启发式方法，而更多依靠稳健的、上下文感知的推理。准确判断角色需要推断、换位思考和社会理解——这些技能更符合人类判断，而不是模式匹配。综上所述，我们的研究结果指出，测试时计算，特别是推理时推断，是构建更好地反映类似人类判断的角色扮演 LLM 评估器的一个有前景的方向。我们的主要贡献如下：

- 我们介绍了 PersonaEval，这是第一个直接评估 LLMs 能否从自然对话中识别人类角色的基准，这是可靠角色扮演评估所必需但尚未深入研究的基础。
- 我们证明了，最先进的 LLMs 在角色识别方面的表现还未达到人类水平，揭示了它们在反映人类判断能力上的一个关键缺口。
- 我们发现，可靠的角色扮演评估并不依赖于特定角色的训练或提示，而是依赖于推理能力，特别强调测试时的计算，尤其是推理，作为构建更符合人类的 LLM 评估器的一种有前景的策略。

2 相关工作

角色扮演评估 角色扮演对于对齐大型语言模型 (LLMs) 与人类价值观至关重要，这使得人类专家成为评估的黄金标准，就像早期 LLM 角色扮演研究所做的那样 (Zhou et al., 2023)。然而，现代角色扮演系统的庞大规模使得全面的人类评估因高昂的成本和延迟而不切实际。因此，一些研究 (Lu et al., 2025; Wang et al., 2024a;b) 转向自然语言生成任务的传统指标，包括 BLEU 和 ROUGE。这些指标计算参考文本和生成内容之间的相似性。然而在角色扮演领域，真实情况总是不确定的，因为可能存在多个正确的回应。因此，研究人员逐渐转向无参考评估。一些研究训练一个奖励模型来估计生成响应的质量 (Tu et al., 2024; Wu et al., 2025; Dai et al., 2024)。越来越多的研究利用 LLMs 作为评委，从人类的角度评估角色扮演内容的质量 (Gusev, 2024; Wang et al., 2024c)，得益于其强大的泛化能力和提供更全面细致反馈的能力。虽然其中一些在引用引导的方式中利用 LLM 评估器 (Zhou et al., 2024a; Yu et al., 2024)，但大多数利用 LLMs 从不同维度进行评分 (Yang et al., 2025; Lu et al., 2025; Zhou et al., 2024a; Shao et al., 2023) 或执行成对比较 (AI, 2024; Wu et al., 2025)。其中，Wang et al. (2025b) 甚至建立了一个多代理系统来做最终判断。为了

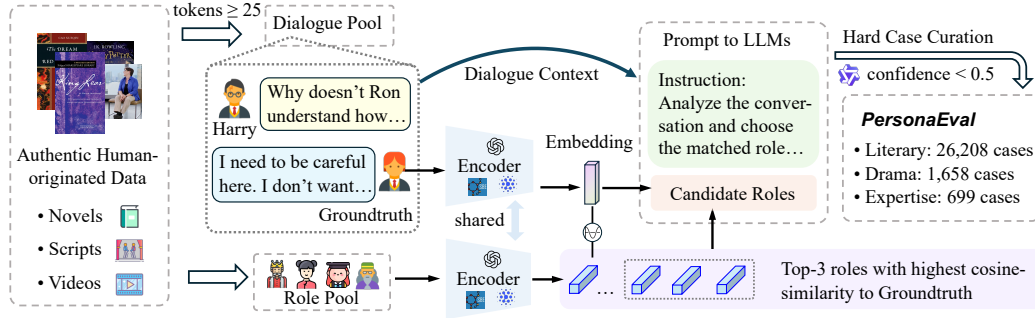


Figure 2: PersonaEval 的策划流程。PersonaEval 通过一个基于小说、剧本和视频中的人类创作对话的受限分类任务来评估 LLM 评估者的角色识别能力。每个实例包括一个两轮对话：第一轮来自一个已知角色，第二轮来自一个目标角色，其身份需要通过 LLM 来预测。使用真实角色的嵌入来从候选集中检索出前三个最相似的角色，形成一个具有挑战性的候选集。一个标准化的提示（见附录 A）向 LLMs 展示对话和候选角色。为了确保难度，只有强基线模型（Qwen-max）在正确答案中表现出低置信度的实例才会被保留。最终基准包括三个轨道——PersonaEval-Literary（26,208），PersonaEval-Drama（1,658），和 PersonaEval-Expertise（699）。

避免评分带来的主观偏差，一些研究让 LLM 评估者执行多类别分类（Lu et al., 2024; Yuan et al., 2024）。

这些方法旨在评估大型语言模型在扮演特定角色时的有效性，从而有助于更广泛地理解它们在角色扮演场景中的表现。然而，在确保评估结果既可靠又符合人类对角色遵循的期望方面，仍然存在挑战。

评估大型语言模型评估器 LLM 作为评估者的可靠性在社区中受到越来越多的关注（Son et al., 2024a; Wei et al., 2024; Zhang et al., 2024）。研究人员首先测试 LLM 在评估者角色中的指令遵循能力。Zheng et al. (2023) 通过使用先进的 LLM 评估开放性问题的表现来解决这一问题。Wang et al. (2023) 通过提供任务特定（例如，摘要生成）和方面特定（例如，相关性）的指令，将 ChatGPT 视作类人的评估者，以评估各种 NLG（自然语言生成）模型的输出。Murugadoss et al. (2024) 探讨 LLM 的评估是否纯粹基于提示指令或也反映了对于高质量数据的固有偏好，如其微调数据。此外，还有关于 LLM 的各种元评估研究，侧重点各不相同。Chern et al. (2024) 通过多代理辩论评估 LLM 评估者，而 Son et al. (2024b) 和 Hada et al. (2024) 则提出了多语言基准。Eiras et al. (2025) 仔细审查了 LLM 评审的安全相关方面。Zhou et al. (2024b) 关注论文审查领域，而 Wang et al. (2025a) 则关注 LLM 评估者在软件工程中的表现。

尽管对大型语言模型（LLM）的元评估的兴趣正在增长，但 LLM 作为角色扮演评估者的具体评估仍然探索不足。尽管有些工作（Yang et al., 2025）包含切除研究以检查 LLM 评估者与人类专家之间的对齐，但在对角色扮演 LLM 评估者的元评估上仍缺乏系统的视角，特别是在其区别细微角色身份的能力方面。然而，这种能力对于高阶角色扮演评估至关重要；没有这种能力，任何后续的角色忠诚度判断都建立在不稳定且未经验证的基础上，揭示出我们对 LLM 如何进行人类对齐评估的理解中存在的差距。

3 PersonaEval

我们设计了 PersonaEval，以测试大型语言模型（LLMs）是否能在人类角色扮演评估的核心子任务中执行类似人类的判断：基于对话上下文识别说话者身份。为了构建角色识别的可靠基准，我们首先将任务定义为一个受限分类问题（章节 3.1），然后构建跨越不同领域的多样化、由人类撰写的数据集（章节 3.2），以捕捉角色扮演。为了增加挑战性并最小化捷径线索，我们设计了语义上相似的对抗干扰项（章节 3.3）。最后，我们应用了一个困难案例筛选流程（章节 3.4）。完整的筛选流程见图 2。

3.1 任务表述

PersonaEval 采用了一个分类框架来评估大型语言模型 (LLMs) 中的角色识别能力。每个实例展示了两个角色之间的两回合对话: Character1 (身份已知) 和 Character2 (需要推断身份)。模型会得到对话背景和四个候选角色 (五个为专业水平的设置, 见第 3.3 节), 每个候选者都有详细的简介。评估者给每个候选者分配信心水平分数, 更高的分数表示更强的匹配信念。为了减轻位置偏见, 候选者的顺序是随机的。此设计将一种本质上主观的任务转变为一种可验证的任务, 使用从原始资料中提取的确定性真实数据。完整的提示信息在附录 A 中提供。

虽然这种设置看起来可能较为不常见, 这主要是因为此前的工作通常隐含地假设 LLMs 具备这种基本能力。我们的目的是仔细检查这一假设, 并弥补我们认知中的一个尚未充分探讨的空白。尽管如此, 在角色扮演对话中, 我们通常有的不仅仅是两个回合的上下文。我们同意将评估扩展到更丰富的多轮上下文是重要的, 并且我们将本研究视为实现这一更广泛目标的必要的第一步。

3.2 数据组成

PersonaEval 包括三个轨道, 每个轨道针对角色理解的不同方面:

- PersonaEval-Literary: 构建于来自 771 部英文小说的 26,208 段对话之上, 这个轨道测试小说叙事中的人物推断。数据由 CoSER (Wang et al., 2025b) 精选, 包含来自经典和现代小说的文本的经过验证的基于虚构作品的数据集。
- PersonaEval-Drama: 该轨道包含 1,658 个来自电影剧本的中文对话片段, 用于测试模型在书面互动中理解角色对齐的能力。数据改编自部分开源的 CharacterEval 数据集 (Tu et al., 2024)。
- PersonaEval-Expertise: 来自 Wired 的“5 Levels”系列视频³, 此音轨提供了 699 个分层解释, 领域专家将内容根据受众的不同知识水平 (儿童、青少年、大学生、研究生和专家) 进行定制。对话测试模型是否可以根据语言和概念线索推断说话者的预期受众, 而不依赖于专业领域知识。

这些轨道共同确保了对虚构、表演和教学领域的覆盖。文学文本强调人物性格和内心思想, 而剧本则更注重对话风格, 教学视频则涉及角色与观众的协调。这种多样性在某种程度上有助于减轻领域特定的偏见。

重要的是, 所有源数据均由人为创作, 避免了来自合成模型生成内容的污染。此设计选择确保评估与人类一致, 因为公共数据集越来越多地与先前模型输出重叠。

我们还在语言多样性与人类原创材料的可用性和质量之间取得平衡, 提供两个英语轨道和一个中文轨道。英语资源, 尤其是文学小说, 更加丰富多样, 使我们能够构建坚实的高质量基准。同时, 我们包含一个中文轨道以引入语言和文化多样性, 认识到在非英语环境中评估模型的重要性。

由于某些材料可能未包含在大型语言模型的预训练语料库中, 我们提供了详细的角色描述, 包括基本特征和细致的情节相关背景。这有助于确保角色识别的性能反映推理能力, 而不是由于缺乏背景知识, 从而最大限度地减少知识空白的影响。

3.3 对抗性干扰项构建

为了确保每个实例都提供一个真正的推理挑战, 我们构造了对抗性干扰项, 这些干扰项在语义上与正确角色非常接近。如图 2 所示, 对于 PersonaEval-Literary 和 PersonaEval-Drama, 我们使用三个独立训练的模型嵌入所有候选角色配置文件, 以避免单一模型结构的偏差: OpenAI 的 text-embedding-3-small (OpenAI, 2024), BGE-M3 (Chen et al., 2024), 以及 Sentence-BERT 的 (Reimers & Gurevych, 2019) multilingual distiluse-base-multilingual-cased-v1⁴。对于每个实例, 我们计算真实角色配置文件与池中所有其他配置文件之间的余弦相似性。对于每个嵌入模型, 我们选择最相似但不正确的非目标角色 (即最相似但不正确

³<https://www.wired.com/video/series/5-levels>

⁴<https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1>

的), 并使用这些形成一个由三个组成的干扰项集, 确保多样但具有挑战性的对比选项。这迫使评估者解决细微的人类级别的歧义, 而不是依赖表面启发法。

在 PersonaEval-Expertise 通道中, 我们利用了固定的五级角色层次结构 (从儿童到专家)。每个对话使用四个不正确的级别作为分散因素, 反映了真实的教育支架挑战, 而无需基于嵌入的选择。所有选项都与从原始视频记录中提取的特定受众配置文件配对。

在所有轨迹中, 我们避免使用合成扰动, 而是利用自然发生的人类角色来构建干扰项。这确保了任务测试的是解决现实世界中的歧义, 而非人为的对比技巧。

3.4 难案例策划

我们观察到, 原始未过滤的数据包含许多琐碎的案例 (例如, 直接的名字提及, 简单的问候), 这些并不能真正测试角色推断。为了避免琐碎案例导致的性能膨胀, 并确保基准测试专注于有意义的推理, 我们明确筛选出即使是强大的模型也难以处理的困难实例。我们的目标不是广泛覆盖, 而是集中评估模型是否能够解决微妙的、接近人类水平的歧义。我们采用了一个两阶段的过滤过程:

阶段 1: 低信息过滤 我们去除 Character2 发言少于 25 个词元的对话轮次。这些低信息量的回复 (例如, “正是如此。”) 提供的角色推断基础很少, 因此被排除在外, 以保持任务的有意义的复杂性。

使用 Qwen-max, 我们筛选出模型对正确答案显示高度置信的实例。具体来说, 我们丢弃任何 Qwen-max 对真实角色分配超过 50 % 置信度的实例。从最初超过 110,000 个实例中, 这种筛选产生了 28,565 个具有挑战性的例子, 每个例子都需要超越表面线索的推理。然而, 这种基于置信度的筛选可能引入系统性偏差。更多讨论可以在附录中找到。

4 实验

我们将实验分析分为四个部分。首先, 我们评估模型在 PersonaEval 上的表现 (第 4.1 节), 然后分析大型语言模型 (LLM) 的推理能力的影响 (第 4.2 节), 接着展示案例研究 (第 4.3 节), 最后以一项人类研究结束 (第 4.4 节)。

由于之前的研究经常使用像 GPT-3.5 和 GPT-4 这样的高级大型语言模型 (LLMs) 作为角色扮演评估器 (Shao et al., 2023; Lu et al., 2024), 我们评估了一套多样化的先进模型, 包括几个在现有角色扮演研究中很少被考虑的推理模型。模型版本在附录 C 中详细说明。我们报告分类准确率 (top-1 准确率) 作为模型性能的主要指标。

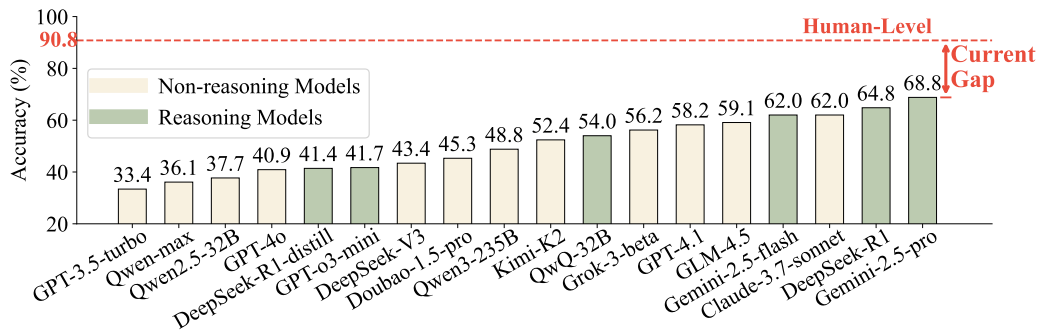


Figure 3: 关于 PersonaEval 的大型语言模型的准确性。大多数大型语言模型在角色识别方面与人类相比表现不佳, 包括最新的 GLM-4.5, 这凸显了一个基本的限制。推理模型显示出明显的优势, 这表明有效的评估需要更深入的推理。DeepSeek-R1-distill 指的是 DeepSeek-R1-distill-Qwen-32B。

4.1 主要结果：使用 PersonaEval 进行角色识别

图 3 总结了 PersonaEval 的三个方面的模型性能。大多数大型语言模型的准确率在 40-60 % 之间，包括最新的 GLM-4.5，即使是表现最佳的模型 Gemini-2.5-pro (03-25 版本，详见附录 C)，也仅达到 68.8 %，远低于人类水平的表现（见章节 4.4）。这些发现揭示了当前大型语言模型与一致性角色识别的需求之间存在显著差距。由于这一任务是进行可信角色扮演评估的必要条件，结果突显了将大型语言模型作为评判方法的核心局限性。完整结果见附录 D。

尽管如此，一些模型显示出进步的迹象。如附录 D 中表格 1 所示，top-2 准确率和平均排名表明部分理解：许多模型将正确的角色排在第二位，表明它们仍然知道正确答案。值得注意的是，Claude-3.7-sonnet、DeepSeek-R1 和 Gemini-2.5-pro 达到了接近完美的 top-2 准确率。此外，ECE 和 Brier Score 都保持相对较低，表明模型表达了适当的不确定性，而不是过度自信的错误。这些结果共同提供了一个更乐观的视角，表明虽然 LLM 评估者尚未达到人类水平，但有些正在迎头赶上。随着进一步的进步，它们可能在角色扮演评估中变得可行。

4.2 推理模型分析

推理模型在角色识别方面展现出明显的优势。如图 3 所示，Gemini-2.5-pro 优于所有基础模型，包括 Claude-3.7-sonnet。

我们还比较了具有和不具有原生推理能力的推理模型。DeepSeek-R1-distill-Qwen-32B，它将 DeepSeek-R1 中的推理能力蒸馏到 Qwen2.5 中，相较于其基础模型仅取得了小幅提升。相比之下，QwQ-32B 在推理密集型任务上通过强化学习训练，显著优于 Qwen2.5。这表明，浅层次或蒸馏的推理可能不适合于角色扮演评估，而端到端的推理训练更为有效。

这些结果支持一个更广泛的见解：有效的角色识别需要健全的、模型本身的推理。除了模式匹配，评估者还必须进行情境判断和逻辑推理——这些能力更符合类人评价。

我们的分析表明，几种特定类型的推理对于有效的角色识别尤为重要。这些包括：

- 视角转换：根据对话中的上下文线索推断说话者的背景、目标和观点的能力。
- 意图推断：理解话语背后潜在意图的能力，通常超出其表层语义意义。
- 实用推理：在互动中逐渐展开时，解释陈述的社会和情境意义的技能。

这些认知技能使评估者能够超越简单的语言模式匹配，转而构建一个连贯的说话者身份模型，这对于以符合人类的方式评估角色扮演的真实性至关重要。

4.3 案例分析

通过案例研究，我们发现许多对人类来说微不足道的例子，经常被大型语言模型误解。在这些情况下，模型往往遵循错误的推理轨迹。我们怀疑这是由于在角色扮演的解释上存在根本差异：大型语言模型倾向于关注表面层次的语言线索，如说话风格，而人类则进行更深层次的推理。具体而言，人类擅长意图推断，使他们能够优先考虑说话者的交流目的，并采用实用推理理解对话的社会背景，这是当前模型仍然缺乏的能力。这一观察与 Josh Tenenbaum 的观点相呼应，即大型语言模型从语言中获得智能，而人类在获得以意图驱动的认知后才发展语言能力。

我们在下面举例说明一个这样的情况。括号中的文本表示角色的内心想法。对于人类读者来说，识别说话者的推理路径是直接的。首先，哈利的内心独白明确提到“罗恩”是他关心的对象（“为什么罗恩不明白……”）。其次，回应者（角色 2）在回复中直接称“哈利”。通过整合这两个明确的引用，读者可以毫无歧义地得出角色 2 是罗恩·韦斯莱。然而，高级模型未能连接这些重要的上下文信号，反而看似更集中于说话者的语气和语言风格。附录 E 中提供了其他案例。

对话和候选角色

Character1: Harry Potter
 [Thinking: 为什么罗恩不明白 how important this is?]
 I want to know what he's up to. And don't tell me it's all in my head, not after what I overheard between him and Snape.
 Character2: Unknown
 [Thinking: I need to be careful here. 我不想忽视哈利的担忧, but this is getting out of hand.]
 I never said it was all in your head, but there's no rule saying only one person at a time can be plotting anything in this place! 你对马尔福有点着迷了, 哈利.
 Candidate Roles:
 Hermione Granger, Ginny Weasley, 罗恩·韦斯莱 (正确), and Harry Potter

响应来自 双子-2.5-pro-0605

Analysis:

赫敏·格兰杰是最有可能回应的人, as her cautious internal thought, logical reasoning ("no rule saying.."), and balanced confrontation ("You're getting a bit obsessed...") align closely with her character profile.

Ginny is a plausible second due to her confidence and perceptiveness. 根据测量到的语气, Ron 的可能性较小 that doesn't quite match his typically more reactive style. Harry is Character 1 being the one expressing obsession and therefore cannot be the responder.

Final Answer:

赫敏: 0.75, Ginny: 0.15, Ron: 0.10, Harry: 0.00
 (Higher probability indicates stronger role correspondence.)

4.4 人体研究

为了评估人在 PersonaEval 上的表现并与大型语言模型进行比较, 我们进行了一项有控制的研究, 对象是 20 名高学历志愿者, 包括 10 名本科生和 10 名博士生, 这些志愿者被选中以匹配任务所需的推理和领域知识需求。每位参与者都得到 50 个从 DeepSeek-R1 在所有三个基准路径上做出错误高置信度预测的案例中抽取的样例。为了让参与者能够做出慎重而有深思熟虑的回答, 没有设定时间限制。

我们故意选择具有挑战性的案例进行人工评估, 以确保在非平凡实例上的有意义比较。虽然这种抽样会引入分布差异, 但实际上这是一个保守选择, 大大加强了我们的发现: 如果人类在对模型始终困难的案例中仍表现良好, 那么性能差异就变得更加引人注目。我们已经手动验证了基准测试中的许多案例对人类来说是直接的。如果我们在整个基准中进行均匀抽样, 人工准确率可能会更高。

如图 3 所示, 参与者的平均准确率达到 90.8 %, 远远超过表现最佳的 LLM。这凸显了当前 LLM 能力与人类角色识别理解之间的巨大差距。完整结果见附录 F。

5 改进调查

为了更好地理解哪些因素有助于有效的角色扮演评估, 我们研究了两种常用的改进大型语言模型的策略: 训练时的适应和测试时的计算。在训练时, 我们测试通过微调注入特定角色知识是否能提高角色识别性能 (第 5.1 节)。在测试时, 我们应用少样本提示和自一致性, 检查额外的推理时间计算是否能提高评估者的准确性 (第 5.2 节)。详细结果见附录 G。

5.1 训练时适应

许多角色扮演研究通过微调大型语言模型 (LLMs) 来注入特定角色的知识, 以提高下游性能。在本节中, 我们评估了两代此类模型: Doubao-pro-character 和 Doubao-1.5-pro-character, 它们分别是通过微调 Doubao-pro 和 Doubao-1.5-pro 得到的。

这些模型是由字节跳动⁵提供的闭源模型。训练时的适应是通过在富含角色知识、身份识别、对话对齐、行为对齐和指令调整的数据集上进行微调来实现的。这些在一个更广泛的系统提示 + SFT 框架下实现, 以确保模型能够在对话交互中保持角色一致性, 识别社会角色,

⁵<https://team.doubao.com/en>

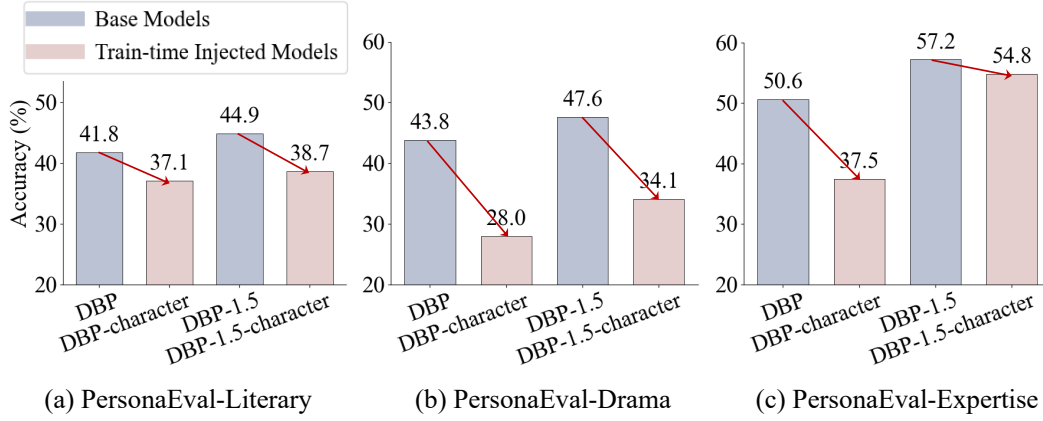


Figure 4: 对 PersonaEval 在三个轨道上四个模型的准确性，分别对应 (a)、(b) 和 (c)，表明在角色特定数据上进行微调并不能提高角色识别的性能，甚至可能降低其性能，而基础模型能力的改进则显示出更一致的收益。DBP 是 Doubao-pro 的缩写。DBP-character 和 DBP-1.5-character 分别是 DBP 和 DBP-1.5 通过角色特定数据微调得到的。

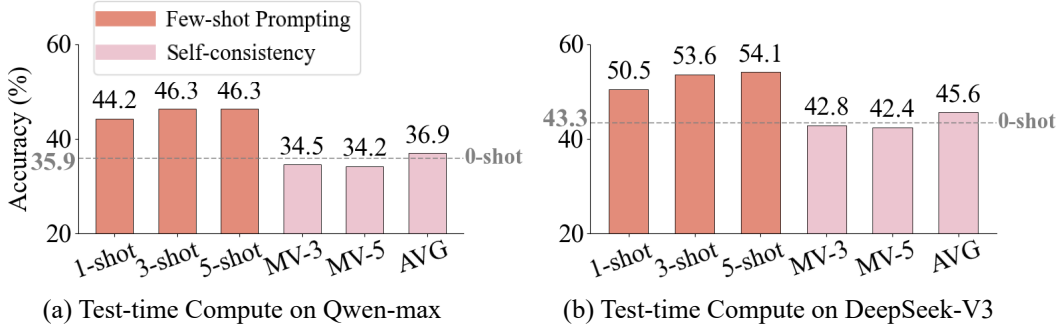


Figure 5: 在 (a) Qwen-max 和 (b) DeepSeek-V3 上使用两种测试时计算策略对 PersonaEval-Literary 的准确性。每个图中的灰色虚线表示对应模型的零次精度。少样本提示（1 次提示、3 次提示、5 次提示）显示出一致但饱和的增益。相比之下，使用自一致性（MV: 多数票，AVG: 加权平均）在各个设置中几乎没有改善。

并推断身份。此外，该系统还强调多轮对话记忆、上下文保留和用户画像构建，以支持长期连贯性。

如图 4 所示，在角色知识上进行微调并不会提高我们角色识别任务的表现；事实上，它反而会降低性能。一种可能的解释是，在微调过程中记忆与角色相关的模式可能会干扰模型的原生推理能力，而这种能力对于评估细致入微的人设对齐至关重要。这表明，通过训练时间的适应来注入角色知识并不是构建可靠角色扮演评估器的有效方法。相比之下，比较 Doubao-pro 和 Doubao-1.5-pro 表明，基础模型本身的改进，例如通用能力的升级，对角色识别性能有更积极的影响。

我们还测试了 CoSER-Llama3.1-8B (Wang et al., 2025b)，这是在 PersonaEval-Literary 的数据源中引入的一个微调模型。然而，它未能遵循我们的基准测试说明，甚至无法产生可用的输出，使得评估变得不可行。这进一步突显了仅靠角色特定的训练不足以构建有能力的评估者。

5.2 测试时间计算

在本节中，我们评估两种流行的测试时计算策略：少样本提示 (Brown et al., 2020) 和自一致性 (Wang et al., 2022)。我们专注于 PersonaEval-Literary 使用 Qwen-max 和 DeepSeek-V3，因为这一轨道构成了 PersonaEval 的核心，并提供比其他方法更稳定的样本选择。

对于少样本提示，我们测试了一次提示、三次提示和五次提示的设置。对于自一致性，我们采用 $K = 3$ 和 $K = 5$ 的多数投票法以及加权平均。如图 5 所示，少样本提示能持续提高模型性能，但在五次提示时增益趋于平稳。相比之下，自一致性在所有设置中提供的改进微乎其微。

这些发现支持了我们早期的分析（章节 4.2）：角色扮演评估更依赖于模型的推理能力，而非采样或集成技术。尽管少样本提示可以传授一些推理模式，但其效果受限于示例的质量和普适性。同时，自我一致性仅仅是在加强模型现有的推理能力，而不是拓展它，并未带来有意义的性能提升。

6 结论

在这项工作中，我们介绍了 PersonaEval，这是第一个用于评估大型语言模型是否能可靠地从自然对话中识别角色的基准，这是实现符合人类角色扮演评估的基础步骤。我们的结果显示，目前的大型语言模型评估器远不及人类水平，即使是最强大和最新的模型在对待人类来说简单的案例上也表现不佳。这突显了现有大型语言模型作为评审者流程中的核心局限性。通过实证分析，我们进一步发现，使用角色特定数据进行训练时的适应几乎没有好处，而测试时的方法，特别是推理模型，显示出更多的前景。这些发现表明，对于有效的角色扮演评估，健壮的、上下文感知的推理，而不是记忆，是至关重要的，并指出测试时计算作为一个实际的和前瞻性的方向。

展望未来，一个有前景的研究方向是超越输出准确性，探讨大型语言模型评估者如何做出预测。这包括分析模型间的推理轨迹，将其与人类思维过程进行比较，并识别其判断不同之处。为此，可能会开发新的诊断工具来可视化和解释模型的推理路径。这些见解可以为新型的测试指导提供信息，例如与人类一致的推理链或基于理由的提示。另一个方向是探索如何在推理时系统地将类人推理策略注入模型中。最终，缩小大型语言模型和人类评估之间的差距可能需要弥合更深层次的认知差异——从单纯的模式识别转向基于意图、背景和社会理解的推理，就像 Josh Tenenbaum 所强调的那样，人类判断在语言之前是从经验中产生的。我们希望 PersonaEval 能成为实现这一目标的一个步骤。

7

致谢

本研究得到了中国山东省重点研发计划（2023CXGC010214, 2024CXGC010213）的支持。我们感谢资助机构给予的支持。感谢所有匿名审稿人的宝贵建议。

References

- Boson AI. Introducing rpbench-auto. <https://boson.ai/rpbench-blog/>, 2024. Accessed: 2025-03-24.
- Anthropic. Claude 3.7 sonnet, 2025. URL <https://www.anthropic.com/claude/sonnet>. Accessed: 2025-03-25.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023.
- Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954, 2024.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216, 2024.

- Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Joanna Matthiesen, Kevin Smith, and Josh Tenenbaum. Evaluating large vision-and-language models on children’s mathematical olympiads. *Advances in Neural Information Processing Systems*, 37:15779–15800, 2024.
- Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*, 2024.
- Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents, 2024.
- Francisco Eiras, Elliott Zemor, Eric Lin, and Vaikkunth Mugunthan. Know thy judge: On the robustness meta-evaluation of llm safety judges. *arXiv preprint arXiv:2503.04474*, 2025.
- Omid Ghasemi, Adam JL Harris, and Ben R Newell. From preference shifts to information leaks: Examining individuals’ sensitivity to information leakage in the framing effect. *Cognition*, 258:106087, 2025.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Ilya Gusev. Pingpong: A benchmark for role-playing language models with user emulation and multi-model evaluation, 2024. URL <https://arxiv.org/abs/2409.06820>.
- Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. Metal: Towards multilingual meta-evaluation. *arXiv preprint arXiv:2404.01667*, 2024.
- Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. Rolemrc: A fine-grained composite benchmark for role-playing and instruction-following. *arXiv preprint arXiv:2502.11387*, 2025.
- Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*, 2024.
- Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. Evaluating the evaluator: Measuring llms’ adherence to task evaluation instructions, 2024.
- OpenAI. ChatGPT (3.5 Turbo Version) [Large Language Model]. <https://chat.openai.com>, 2023.
- OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- OpenAI. text-embedding-3-small. <https://openai.com/index/new-embedding-models-and-api-updates/>, 2024. Accessed: 2025-03-23.
- OpenAI. Gpt-4.1, 2025a. URL <https://openai.com/index/gpt-4-1>. Accessed: 2025-08-04.
- OpenAI. Openai o3-mini. <https://openai.com/index/openai-o3-mini/>, 2025b. Accessed: 2025-03-26.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, Singapore, December 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.emnlp-main.814>.

- Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. Llm-as-a-judge & reward model: What they can and cannot do, 2024a.
- Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models. arXiv preprint arXiv:2410.17578, 2024b.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijie Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaying Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence, 2025. URL <https://arxiv.org/abs/2507.20534>.
- Qwen Team. Qwen3 technical report, 2025a. URL <https://arxiv.org/abs/2505.09388>.
- Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025b. URL <https://qwenlm.github.io/blog/qwq-32b/>.
- Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. Caractereval: A chinese benchmark for role-playing conversational agent evaluation. arXiv preprint arXiv:2401.01275, 2024.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? a preliminary study. In Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini (eds.), Proceedings of the 4th New Frontiers in Summarization Workshop, pp. 1–11, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.newsum-1.1.
- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Findings of the Association for Computational Linguistics ACL 2024, pp. 14743–14777, Bangkok, Thailand and virtual meeting, August 2024a. Association for Computational Linguistics.

- Ruiqi Wang, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. arXiv preprint arXiv:2502.06193, 2025a.
- Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. Characteristic ai agents via large language models. arXiv preprint arXiv:2403.12368, 2024b.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yay-ing Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1840–1873, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, et al. Coser: Coordinating llm-based persona simulation of established roles. arXiv preprint arXiv:2502.09082, 2025b.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2022.
- Hui Wei, Shenghua He, Tian Xia, Andy Wong, Jingyang Lin, and Mei Han. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates, 2024.
- Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. Raiden benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In Proceedings of the 31st International Conference on Computational Linguistics, pp. 11086–11106, 2025.
- xAI. Grok 3 beta, 2025. URL <https://x.ai/news/grok-3>. Accessed: 2025-08-04.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.
- Tao Yang, Yuhua Zhu, Xiaojun Quan, Cong Liu, and Qifan Wang. Psyplay: Personality-infused role-playing conversational agents. arXiv preprint arXiv:2502.03821, 2025.
- Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Peng Hao, and Liehuang Zhu. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. arXiv preprint arXiv:2402.13717, 2024.
- Xinfeng Yuan, Siyu Yuan, Yuhang Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. Evaluating character understanding of large language models via character profiling from fictional works. arXiv preprint arXiv:2404.12726, 2024.
- Z.AI. Glm-4.5, 2025. URL <https://z.ai/blog/glm-4.5>. Accessed: 2025-08-04.
- Wenbo Zhang, Zihang Xu, and Hengrui Cai. Defining boundaries: A spectrum of task feasibility for large language models. arXiv preprint arXiv:2408.05873, 2024.
- Yulai Zhao, Haolin Liu, Dian Yu, SY Kung, Haitao Mi, and Dong Yu. One token to fool llm-as-a-judge. arXiv preprint arXiv:2507.08794, 2025.

- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.
- Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*, 2023.
- Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, et al. Characterbench: Benchmarking character customization of large language models. *arXiv preprint arXiv:2412.11912*, 2024a.
- Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351, 2024b.

A 提示大语言模型评估者

在 PersonaEval 中评估 LLM 的示例提示

You are an AI specialist tasked with dialogue role recognition. Please analyze the following conversation and determine the likelihood of four character candidates being the responder.

会话

[Character1: Harry Potter]

[Begin]

[Why doesn't Ron understand how important this is?] (frustrated) I want to know what he's up to. And don't tell me it's all in my head, not after what I overheard between him and Snape.

[End]

[Character2]

[Begin]

[I need to be careful here. I don't want to dismiss Harry's concerns, but this is getting out of hand.] I never said it was all in your head, but there's no rule saying only one person at a time can be plotting anything in this place! You're getting a bit obsessed with Malfoy, Harry.

[End]

任务

Your task is to analyze the response of Character2 and estimate the Bayesian probability distribution for each of the four character candidates. The probabilities should reflect the likelihood of each candidate being the one responding, based on their profiles. The sum of all probabilities must equal 1. A higher probability for a candidate indicates that the response aligns better with that candidate.

字符候选集

1. Ginny Weasley

Ginny Weasley, the youngest child and only daughter of the Weasley family, emerges as a pivotal character in "Harry Potter and the Half-Blood Prince." With her fiery red hair and strong-willed personality, Ginny has grown from Ron's shy little sister into a confident and capable young witch. She possesses a quick wit and a mischievous streak, often using humor to diffuse tense situations...

2. Ron Weasley

Ron Weasley, Harry Potter's loyal best friend and steadfast companion, plays a crucial role in the final installment of the Harry Potter series. With his trademark red hair and freckles, Ron embodies the heart and humor of the trio. Coming from a large, loving wizarding family, Ron's background provides both strength and insecurity as he faces the challenges ahead...

3. Hermione Granger

Hermione Granger, in "Harry Potter and the Deathly Hallows", emerges as a brilliant, resourceful, and fiercely loyal young witch. As one of Harry Potter's closest friends and allies, she plays a crucial role in the quest to defeat Lord Voldemort. Hermione's exceptional intelligence and vast magical knowledge make her an invaluable asset to the trio's mission...

4. Harry Potter

Harry Potter, the protagonist of "Harry Potter and the Half-Blood Prince," is a sixteen-year-old wizard entering his sixth year at Hogwarts School of Witchcraft and Wizardry. With his distinctive lightning bolt scar and round glasses, Harry continues to bear the weight of being "The Chosen One" in the fight against Lord Voldemort...

响应格式

Analyze step by step, and then output the following JSON object containing the final probability distribution. Ensure that the sum of all probabilities equals 1, with each probability representing the likelihood that a given candidate is the responder. Do not modify the character names, and use the full character names exactly as they appear in the task.

```

{
  "Ginny Weasley": probability_for_Ginny Weasley,
  "Ron Weasley": probability_for_Ron Weasley,
  "Hermione Granger": probability_for_Hermione Granger,
  "Harry Potter": probability_for_Harry Potter,
}

```

B 基于置信度过滤的讨论

我们选择自动化的单模型过滤策略是因为数据集的规模庞大，这使得在这项工作的范围内进行人工注释或更复杂的多模型验证方案变得不切实际。未经过滤的数据包括许多琐碎的情况（例如，直接提到名字，简单的问候语），这些情况并不能真正测试角色推断。因此，一个整理过程是必需的，以确保基准的重点是推理。

我们认识到这种方法可能引入系统性偏差。生成的数据集最可靠地被视为对 Qwen 系列模型或具有类似能力的模型的一组难题。重要的是，我们并不是直接选择 Qwen 失败的案例，而是使用置信度阈值（正确答案的置信度小于 0.5）来识别不确定的案例，从而减少为任何特定模型过拟合过滤器的风险。然而，我们在本文中的主要目标是发现和展示大语言模型中被忽视的推理问题。经过精心挑选的数据集，通过成功挑战甚至像 Qwen-max 这样强大的模型，有效地达到了这一目的，并为我们的主要论点提供了有力的支持。我们相信这个基准，尽管存在局限性，是一个有价值的第一步。对于未来旨在更广泛应用的基准，我们建议采用使用多样化模型套件的交叉验证方法，以进一步增强稳健性和实用性。

C 论文中使用的 LLM 版本

所有大型语言模型及其相应版本如下：

- GPT-3.5-turbo (OpenAI, 2023) : gpt-3.5-turbo-0125
- GPT-4o (OpenAI, 2023) : gpt-4o-2024-08-06
- GPT-4.1 (OpenAI, 2025a) : gpt-4.1-2025-04-14
- GPT-o3-mini (OpenAI, 2025b) : o3-mini-2025-01-31
- Qwen-max (Bai et al., 2023) : qwen-max-2025-01-25
- Qwen3-235B (Team, 2025a) : Qwen3-235B-a22B, 非思考版本
- 豆包-1.5-专业版⁶ : doubao-1-5-pro-32k-250115
- DeepSeek-V3 (Bi et al., 2024) : DeepSeek-V3-250324
- DeepSeek-R1 (Guo et al., 2025) : DeepSeek-R1-0120
- Gemini-2.5-pro (Team et al., 2024) : Gemini-2.5-pro-preview-03-25

其他模型，包括 Claude-3.7-sonnet (Anthropic, 2025)、Qwen2.5-32B (Yang et al., 2024)、QwQ-32B (Team, 2025b)、DeepSeek-R1-distill-Qwen-32b (Guo et al., 2025)、Kimi-K2-Instruct (Team et al., 2025)、GLM-4.5 (Z.AI, 2025)、Grok-3-beta (xAI, 2025) 和 Gemini-2.5-flash (Team et al., 2024)，截至目前只有一个版本。

D PersonaEval 的详细实验结果

我们在此报告 PersonaEval 的详细实验结果，包括 top-1 准确率、top-2 准确率、响应中真实角色的平均排名 (MR)、期望校准误差 (ECE) 和 Brier 分数 (BS)。除了平均排名外，其他结果以百分比显示。DeepSeek-R1-distill 是指 DeepSeek-R1-distill-Qwen-32B。每列中表现最好的结果以粗体标出。

我们首先在表格 1 中展示了整个基准的汇总指标，然后分别展示了三个轨道上的结果。我们还比较了 DeepSeek-V3 的两个版本，表明角色扮演大模型评估者的能力差距正在缩小。

各模型的性能似乎与模型能力密切相关。例如，GPT-3.5-turbo 在表现上始终逊色于更先进的模型，而 Claude-3.7-sonnet 则表现出显著优越的性能。这些结果符合预期，进一步强调了在评估任务中模型质量的重要性。

此外，模型专门化发挥了显著作用。例如，GPT-4o 在 PersonaEval-Expertise 上表现出竞争力，与 Claude-3.7-sonnet 相当，而 DeepSeek-V3 在针对中文任务的基准测试中表现尤为出色。这些观察表明，角色扮演评估者的能力不仅依赖于模型，也与特定领域相关。

⁶<https://team.doubao.com/en>

Model	Top-1 Acc $\uparrow$	Top-2 Acc $\uparrow$	MR $\downarrow$	ECE $\downarrow$	BS $\downarrow$
GPT-3.5-turbo	33.4	71.6	2.02	21.1	19.5
Qwen-max	36.1	77.7	1.92	28.7	20.8
Qwen2.5-32B	37.7	77.7	1.89	23.5	19.1
GPT-4o	40.9	83.2	1.75	16.0	16.7
DeepSeek-R1-distill	41.4	75.0	1.92	22.6	19.2
GPT-o3-mini	41.7	76.3	1.88	23.4	18.5
DeepSeek-V3-241226	38.2	81.0	1.80	24.4	18.7
DeepSeek-V3-250324	43.4	84.2	1.73	22.8	17.8
Doubao-1.5-pro	45.3	82.0	1.72	21.1	17.3
Qwen3-235B	48.8	84.8	1.70	20.3	16.6
Kimi-K2	52.4	86.9	1.60	26.6	16.8
QwQ-32B	54.0	83.8	1.67	22.3	16.4
Grok-3-beta	56.2	91.0	1.50	4.4	12.7
GPT-4.1	58.2	91.3	1.47	11.9	12.4
GLM-4.5	59.1	87.6	1.54	9.7	13.7
Gemini-2.5-flash	62.0	89.9	1.46	11.9	12.6
Claude-3.7-sonnet	62.0	91.2	1.46	8.3	12.1
DeepSeek-R1	64.8	90.0	1.48	14.9	13.2
Gemini-2.5-pro	68.8	92.6	1.38	4.9	10.5

Table 1: PersonaEval 所有三个赛道的完整结果。

Model	Top-1 Acc $\uparrow$	Top-2 Acc $\uparrow$	MR $\downarrow$	ECE $\downarrow$	BS $\downarrow$
GPT-3.5-turbo	33.5	70.9	2.06	21.3	19.5
Qwen-max	35.9	76.3	1.98	29.2	21.0
Qwen2.5-32B	38.0	75.7	1.96	23.7	19.2
GPT-4o	41.3	83.7	1.79	16.0	16.6
DeepSeek-R1-distill	41.8	72.9	1.98	23.0	19.2
GPT-o3-mini	42.5	76.9	1.88	22.9	18.4
DeepSeek-V3-241226	38.0	79.5	1.90	24.6	18.7
DeepSeek-V3-250324	43.3	83.5	1.78	22.9	17.8
Doubao-1.5-pro	44.9	79.1	1.84	21.6	17.5
Qwen3-235B	49.5	86.0	1.67	20.0	16.5
Kimi-K2	53.0	88.1	1.59	26.5	16.7
QwQ-32B	54.3	82.1	1.71	23.0	16.5
Grok-3-Beta	57.1	92.3	1.47	4.0	12.4
GPT-4.1	59.6	93.1	1.44	11.1	12.1
GLM-4.5	60.9	89.5	1.51	10.1	13.5
Gemini-2.5-flash	62.5	91.0	1.45	11.9	12.5
Claude-3.7-sonnet	63.0	91.8	1.48	8.0	11.8
DeepSeek-R1	65.4	89.4	1.49	14.7	13.1
Gemini-2.5-pro	69.5	93.3	1.37	5.4	10.3

Table 2: 有关 PersonaEval-Literary 的完整结果。

Model	Top-1 Acc ↑	Top-2 Acc ↑	MR ↓	ECE ↓	BS ↓
GPT-3.5-turbo	26.3	57.8	2.33	26.1	21.8
Qwen-max	31.6	66.9	2.16	33.4	22.5
Qwen2.5-32B	30.0	61.9	2.24	29.6	21.6
GPT-4o	28.3	61.3	2.24	26.1	20.9
DeepSeek-R1-distill	30.1	61.5	2.24	26.8	21.3
GPT-o3-mini	35.0	63.9	2.14	30.7	21.0
DeepSeek-V3-241226	34.7	63.5	2.19	30.3	21.8
DeepSeek-V3-250324	42.2	71.4	1.98	30.0	20.2
Doubao-1.5-pro	47.6	75.3	1.85	22.6	17.4
Qwen3-235B	38.7	72.1	1.97	29.0	19.9
Kimi-K2	47.8	78.8	1.78	33.3	19.7
QwQ-32B	48.0	76.6	1.82	19.3	16.8
Grok-3-Beta	46.9	78.6	1.80	16.3	16.6
GPT-4.1	45.4	76.3	1.84	23.5	17.5
GLM-4.5	47.2	78.5	1.80	17.9	16.6
Gemini-2.5-flash	56.7	81.1	1.67	20.5	15.6
Claude-3.7-sonnet	50.2	77.0	1.80	18.6	16.7
DeepSeek-R1	56.3	84.1	1.64	22.4	15.8
Gemini-2.5-pro	60.3	85.8	1.57	13.1	13.4

Table 3: 关于 PersonaEval-Drama 的完整结果。

Model	Top-1 Acc ↑	Top-2 Acc ↑	MR ↓	ECE ↓	BS ↓
GPT-3.5-turbo	45.1	69.8	2.05	10.0	14.2
Qwen-max	51.5	77.3	1.85	6.0	12.6
Qwen2.5-32B	47.6	73.1	1.99	6.7	13.2
GPT-4o	56.7	78.5	1.79	9.1	12.0
DeepSeek-R1-distill	50.5	73.4	1.96	7.7	13.4
GPT-o3-mini	29.8	53.8	2.41	23.5	16.9
DeepSeek-V3-241226	50.8	73.8	1.92	7.9	13.0
DeepSeek-V3-250324	51.6	74.1	1.93	3.6	12.8
Doubao-1.5-pro	57.2	82.0	1.74	10.2	12.0
Qwen3-235B	47.2	67.7	2.08	10.4	14.1
Kimi-K2	39.6	60.8	1.76	14.8	13.6
QwQ-32B	57.2	77.5	1.84	4.2	12.2
Grok-3-beta	43.2	69.2	2.04	8.5	13.5
GPT-4.1	34.7	56.3	1.87	13.1	14.0
GLM-4.5	44.4	62.3	1.69	8.9	12.5
Gemini-2.5-flash	55.3	69.8	1.40	12.7	10.2
Claude-3.7-sonnet	53.8	74.8	1.86	9.6	12.0
DeepSeek-R1	60.9	80.7	1.70	5.9	10.9
Gemini-2.5-pro	62.1	84.4	1.62	12.1	11.0

Table 4: 关于 PersonaEval-Expertise 的完整结果。

E 详尽案例研究

我们展示了几个案例，其中能力较强的大型语言模型犯了高置信度错误，而这些任务对人类来说仍然相对简单。括号中的文本表示角色的内心想法。

E.1 案例 1：正确处理

这个案例来自《哈利·波特与火焰杯》（哈利·波特，# 4）。它涉及到人物称呼，这需要人为一些思考，但相对简单。然而，大型语言模型很容易被“Weasley”这个提示误导，并在错误的方向上进行大量不必要的分析。

适当的分析： 角色 2 不可能是赫敏·格兰杰（选项 4），因为她已经作为角色 1 在讲话。西弗勒斯·斯内普（选项 3）绝不会以如此随意或周到的方式称呼韦斯莱先生，也不会对戏剧性的学生出场表现出好奇心。罗恩·韦斯莱（选项 1）也不太可能是角色 2，因为他不会称自己的父亲为“韦斯莱先生”——他会说“爸爸”。这一细节强烈暗示角色 2 是哈利·波特（选项 2），因为他经常称罗恩的父亲为“韦斯莱先生”，并且亲身在魁地奇世界杯上听到了那句话。语气——深思熟虑且略显愉悦——也符合哈利的性格，他经常回顾过去的对话并试图理解不寻常的魔法情境。因此，哈利是角色 2 的最佳选择。

对话与候选角色	Claude-3.7-sonnet 的回应
Character1: Hermione Granger [Thinking: I'm getting a bit frustrated having to repeat this information.] You can't Apparate inside the Hogwarts grounds, how often do I have to tell you? Character2: Unknown [Thinking: I remember what Mr. Weasley said about wizards showing off. Maybe that's what they'll do.] Maybe the foreign students are preparing a dramatic entrance... Remember what Mr. Weasley said at the World Cup? 'Always the same —we can't resist showing off when we get together...' Candidate Roles: Ron Weasley, Harry Potter (Correct), Severus Snape, and Hermione Granger	Analysis: The response most likely comes from Ron Weasley, who naturally quotes his father, attended the World Cup, and speaks in a casual, speculative tone. Harry Potter is a secondary possibility, having attended the World Cup, but lacks Ron's personal connection and typical commentary style. Severus Snape is highly unlikely, as the tone, content, and familiarity with Mr. Weasley don't match his character. Hermione Granger is explicitly Character1 and speaks differently, making her an impossible fit. Final Answer: Ron: 0.85, Harry: 0.15, Snape: 0.00, Hermione: 0.00

E.2 情况二：性格特征

这个案例来自《汤姆·索亚历险记》。它需要比较候选人物的关键特征，并选择最合适的一个。对于人类来说，只要对选项的大致特征熟悉，做出正确的选择并不困难。然而，大型语言模型往往过于关注人物简介和情节细节，导致其推理过程不必要的复杂。

合理分析： 角色 2 不可能是波莉姨妈（选项 1），因为她是一个成年人，不会在这种情况下与哈克卷入秘密的道德困境——尤其是不会讨论逃跑或拯救马夫·波特尔。西德（选项 3），汤姆那个行为良好且有些爱告密的表弟，缺乏此处表现出的勇气和内心冲突；他不太可能冒险去帮助像马夫·波特尔这样的人。哈克贝利·费恩（选项 2）已经作为角色 1 发言，所以他也不能是角色 2。剩下的汤姆·索亚（选项 4）非常合适。这句话反映了汤姆典型的内心斗争：对后果的恐惧与日益增长的正义感之间的矛盾——这是《汤姆·索亚历险记》中的一个主要主题。汤姆是那个知道马夫·波特尔是无辜的人，并最终冒险揭露真相，使他成为角色 2 的明显选择。

对话和候选角色	GPT-4o 的回应
Character1: Huckleberry Finn [Thinking: I try to think of a way out of this mess.] Maybe we ought to leave town, Tom. Just light out and never come back. Character2: Unknown [Thinking: I'm torn between fear and a sense of responsibility.] But what about Muff Potter? He didn't do nothing, and he's gonna hang for it if we don't say something. Candidate Roles: Aunt Polly, Huckleberry Finn, Sid, and Tom Sawyer (Correct)	Analysis: The response reflects fear and moral responsibility, best matching Huckleberry Finn, who witnessed the event and cares deeply about justice. Tom Sawyer also fits, given his eventual decision to defend Muff Potter. Aunt Polly is unlikely, since she stays removed from such events, and Sid lacks both the involvement and moral weight needed for this response. Final Answer: Polly: 0.05, Huckleberry: 0.60, Sid: 0.05, Sawyer: 0.30

E.3 案例 3: 无关字符

这个案例来自《哈克贝利·芬历险记》。在这个案例中，根据对话识别无关选项就足以迅速做出正确的选择，而不需要分析表现细节。对于人类来说，这个任务相对简单。然而，大型语言模型倾向于关注语言风格和情节，忽视对话中角色之间的最简单关系。

正确分析： 角色 2 不可能是哈克·菲恩（选项 1）或赫克尔贝里·费恩（选项 3），因为他已经作为角色 1 在讲话。使用“我们”并且语气上的对比显示了涉及到两个不同的角色。汤姆叔叔（选项 2）不是《哈克贝利·费恩历险记》中的角色——他属于一本完全不同的小说（《汤姆叔叔的小屋》作者是哈丽雅特·比彻·斯托）——因此他可以被完全排除。留下来的只有汤姆·索亚（选项 4），完美契合。汤姆以他精心设计的、有想象力的计划而著称，即便这些计划不切实际或危险——比如用小刀挖洞解救吉姆。在这段文字中，他由于情势的紧迫性开始稍微妥协，但仍坚持“合适”逃脱的想法。他的语言充满绕道而行的逻辑和对冒险的兴奋，是经典的汤姆·索亚。

对话和候选角色	DeepSeek-V3 的响应
Character1: Huckleberry Finn [Thinking: I can't believe we're really going to do this.] All right—I don't care where he comes out, so he comes out; and Jim don't, either, I reckon. But there's one thing, anyway—Jim's too old to be dug out with a case-knife. He won't last. Character2: Unknown [Thinking: We need to stick to the plan, but maybe we can compromise a little.] Well, we can't resk being as long as we ought to, because it mayn't take very long for Uncle Silas to hear from down there by New Orleans. He'll hear Jim ain't from there. Then his next move will be to advertise Jim, or something like that. So we can't resk being as long digging him out as we ought to. Candidate Roles: Huck Finn, Uncle Tom, Huckleberry Finn, and Tom Sawyer (Correct)	Analysis: The response shows pragmatic and strategic thinking, which aligns closely with Huckleberry Finn's practical mindset. The informal dialect (“resk” instead of “risk”) and collaborative tone further support this match. Tom Sawyer, by contrast, tends to propose elaborate, romanticized plans based on adventure novels, which doesn't match the straightforward, cautious tone of this response. Uncle Tom, from a completely different novel (Uncle Tom's Cabin), lacks any connection to the context, characters, or dialect here. Since “Huck Finn” and “Huckleberry Finn” are duplicate entries for the same character, the high likelihood is split between them unless they are merged. Final Answer: Huck: 0.50, Uncle: 0.00, Huckleberry: 0.50, Sawyer: 0.00

E.4 案例四：语气和复杂情节推理

这个案例来自《冰与火之歌》第三部《冰雨的风暴》“A Storm of Swords”。这更具挑战性，主要是因为需要领域知识。然而，只要对书中的情节有一个大致的了解，人们可以通过推理和分析做出正确的选择。对于大型语言模型（LLMs）来说，这个案例需要去识别正确的方向并进行仔细的推理。然而，LLMs 常常会遵循错误的推理路径，并仅凭“弑君者”的线索过早得出错误的答案。

正确分析： Character2 不能是泰温·兰尼斯特（选项 1 和 4），因为泰温以其严厉、权威的举止而闻名，从不会如此讽刺或自嘲地说话，特别是在“杀亲者”或“弑君者”问题上。詹姆·兰尼斯特（选项 2）则是著名的“弑君者”（在简历中提到），但他从未去过黑城堡，这使得他与城堡居民的提及与他的实际经历不符。对话的语气——以愤世嫉俗、黑色幽默及与泰温的紧张关系为特征——强烈暗示这是提利昂·兰尼斯特（选项 3）。提利昂确实造访过黑城堡，并常常使用辛辣的机智。这些因素使他成为 Character2 的唯一合适选择。

对话与候选角色

Character1: Kevan Lannister
[Thinking: I need to distance Tywin from Ned Stark's fate.]
Your father had no part in that.
Character2: Unknown
[Thinking: I can't believe they think I'm this naive.]
Castle Black teems with murderers, thieves and rapists, but I don't recall meeting many regicides while I was there. You expect me to believe that if I admit to being a kinslayer and kingslayer, my father will simply nod, forgive me, and pack me off to the Wall with some warm woolen smallclothes. (hoots rudely)

Candidate Roles:
Tywin Lannister (Correct), Jaime Lannister, Tyrion Lannister, and Lord Tywin Lannister

来自 QwQ-32B 的响应

Analysis:
The dialogue is sarcastic, self-aware, and steeped in personal history—marked by references to Castle Black, being a “kinslayer and kingslayer”, and mocking the idea that Tywin Lannister would forgive such crimes. These clues unmistakably point to Jaime Lannister, famously known as the Kingslayer, speaking in his signature bitter tone. The speaker's father is Tywin, eliminating both Tywin and Lord Tywin (duplicates) as possibilities, and Tyrion is neither a kingslayer nor known to describe himself this way. The dialect, references, and tone make Jaime the only plausible match.

Final Answer:
Tywin: 0.00, Jaime: 1.00, Tyrion: 0.00, Lord: 0.00

F 人体研究结果

我们根据样本在三个轨道上的分布将 50 个问题分成三部分：45 个问题来自 PersonaEval-Literary，3 个来自 PersonaEval-Drama，2 个来自 PersonaEval-Expertise。要求每位参与者选择他们认为正确的答案，类似于多项选择任务。如果出现任何语言障碍，允许参与者使用无生成能力的翻译工具，如 LLMs。我们报告每位参与者在每个轨道上的正确答案数量和准确率，如表 5 所示。出于隐私考虑，我们用一个数字来表示每位参与者，而不是使用他们的个人信息。大多数参与者取得了 90 % 的准确率，所有 20 位参与者的总体准确率为 90.8 %。

G 训练时适应和测试时计算的综合结果

我们在此展示了第 5.1 节和第 5.2 节的完整结果，并在附录 D 中提供了相同的度量。表格 6、表格 7 和表格 8 显示了关于三条轨道的训练时间实验的完整结果。对于测试时的计算，我们在此展示了少样本提示的结果，因为自洽性会丢失置信度信息。

Parti.	Lit.	Drama	Exp.	Acc. (%)	Parti.	Lit.	Drama	Exp.	Acc. (%)
No.01	43	2	2	94	No.11	45	2	2	98
No.02	44	1	1	92	No.12	40	2	1	86
No.03	41	1	2	88	No.13	42	2	2	92
No.04	39	3	1	86	No.14	43	2	2	94
No.05	42	1	2	90	No.15	42	2	2	92
No.06	43	2	2	94	No.16	42	3	2	94
No.07	41	2	1	88	No.17	44	0	2	92
No.08	42	2	2	92	No.18	42	2	1	90
No.09	42	1	2	90	No.19	38	1	2	82
No.10	41	3	2	92	No.20	41	2	2	90

Table 5: 人体研究结果，数字表示每位参与者的正确答案数量。

Model	Top-1 Acc ↑	Top-2 Acc ↑	MR ↓	ECE ↓	BS ↓
Doubao-pro	41.8	79.2	1.86	18.5	18.0
Doubao-pro-character	37.1	69.0	2.07	25.7	20.6
Doubao-1.5-pro	44.9	79.1	1.84	21.6	17.5
Doubao-1.5-pro-character	38.7	77.8	1.90	11.1	17.1

Table 6: 关于 PersonaEval-Literary 的训练时适应的全面结果。

Model	Top-1 Acc ↑	Top-2 Acc ↑	MR ↓	ECE ↓	BS ↓
Doubao-pro	43.8	70.0	1.97	25.1	19.3
Doubao-pro-character	28.0	58.0	2.32	37.5	23.9
Doubao-1.5-pro	47.6	75.3	1.85	22.6	17.4
Doubao-1.5-pro-character	34.1	66.8	2.12	20.6	19.2

Table 7: 关于 PersonaEval-Drama 的训练时间适应的全面结果。

Model	Top-1 Acc ↑	Top-2 Acc ↑	MR ↓	ECE ↓	BS ↓
Doubao-pro	50.6	77.1	1.90	12.7	13.1
Doubao-pro-character	37.5	63.9	2.32	12.7	15.5
Doubao-1.5-pro	57.2	82.0	1.74	10.2	12.0
Doubao-1.5-pro-character	54.8	78.8	1.79	17.7	12.8

Table 8: 关于 PersonaEval-Expertise 的训练时间适应的综合结果。

Model	Top-1 Acc ↑	Top-2 Acc ↑	MR ↓	ECE ↓	BS ↓
Qwen-max-0shot	35.9	76.3	1.98	29.2	21.0
Qwen-max-1shot	44.2	76.0	1.90	30.4	20.5
Qwen-max-3shot	46.3	77.0	1.87	29.8	20.0
Qwen-max-5shot	46.3	77.1	1.86	30.3	20.1
DeepSeek-V3-0shot	43.3	83.5	1.78	22.9	17.8
DeepSeek-V3-1shot	50.5	84.6	1.71	15.8	16.2
DeepSeek-V3-3shot	53.6	85.5	1.66	15.5	15.7
DeepSeek-V3-5shot	54.1	86.3	1.65	14.5	15.2

Table 9: 关于 PersonaEval-Literary 的小样本提示的全面结果。