通过动态重要性估计增强角色扮演智能体在解码时的人格遵循

Yuxin Liu[1] **Mingye Zhu**[1] **SiyuanLiu**[1] **Bo Hu**[1] **Lei Zhang**[1*]中国科学技术大学{liuyuxin1010, mingyezhu, liusiyuan0524}@mail.ustc.edu.cn{hubo, leizh23}@ustc.edu.cn

arXiv:2603.01438v1 [cs.CL] 2 Mar 2026

摘要

随着大语言模型的普及，角色扮演语言智能体在社会学研究中的效用日益提升。为了在社会模拟中实现真实感，这些智能体必须遵循角色档案所定义的角色设定，但现有策略——静态提示工程或成本高昂的微调——无法使角色设定适应动态场景。心理学理论，例如认知—情感人格系统理论，为这一问题提供了关键解释：角色设定对行为的影响并非一成不变，而是随场景而变化。这种情境依赖性凸显了自适应角色人格管理的迫切需求。为弥补这一不足，我们提出一种由理论驱动的新方法，动态估计情境相关的角色重要性，并将其整合到加权奖励引导解码中，从而实现推理时的人格遵循。具体而言，我们提出**人**格**动**态**解**码（**PDD**）框架，该框架包含两个关键组件：（1）人格重要性估计（PIE）模块，无需真实值监督即可动态量化角色属性的上下文重要性；（2）人格引导的推理时对齐（PIA）范式，利用这些重要性分数构建加权多目标奖励，并在推理过程中调节生成概率。大量实验表明，我们的方法在话语一致性和行为忠实度方面效果显著。

1 引言

大语言模型（LLM）的进步，伴随着人们日益认识到其在社会学研究中的角色扮演语言智能体（RPLA）方面的重要性（例如，投票行为分析（Zhang 等人，2024年）、谣言传播动态（Mou 等人，2024年））。为确保模拟具有现实效度和统计泛化能力，这类模拟需要精心筛选、具有异质性的角色设定档案，以代表不同的社会人口统计群体。至关重要的是，关于角色设定动态的心理学研究（Sherman 等人，2015；Green 等人，2019年），包括认知—情感人格系统（CAPS），表明人类行为源自特定场景对属性的激活。与此相应，大语言模型应能动态识别依赖上下文的角色设定，并在不同的上下文场景中遵循预定义档案——我们将这一能力称为人格遵循。

现有的人格遵循技术主要分为两类：非参数学习与参数化训练。代表性的非参数方法包括直接提示、上下文学习（ICL）（Dong 等人，2024年）以及检索增强生成（RAG）（Lewis 等人，2020）。这些方法主要依赖提示工程，通过精心设计的文本提示将特定于角色设定的描述注入模型，从而将其输出引导至预定义档案。然而，由于其核心机制依赖于对提示文本的语义识别，模型往往无法深刻理解输入中所嵌入的角色属性。因此，它们无法动态调整行为模式或偏好表达

*Lei Zhang 是通讯作者。

根据具体场景。这一根本局限阻碍了智能体交互过程中真正的动态人格遵循。为解决这些问题，参数化方法（Shao 等人，2023；Zhou 等人，2024年）通过在大规模整理后的语料库（例如，角色对话（Xu 等人，2024年））上进行监督微调（SFT）或低秩适配（LoRA）来优化模型，以增强人格遵循。然而，此类方法需要大量计算资源与标注数据。这一挑战在社会模拟中更为突出，因为具有复杂个性的多样化角色使得数据集整理极为耗费资源。总之，现有方法存在两个核心局限：（1）动态适应性不足：无法识别不同场景中依赖上下文的角色设定；（2）严重的数据依赖：整理覆盖多样化场景的行为数据极为困难。

为解决现有局限性，我们提出**人**格**动**态**解**码（**PDD**），这是一种推理时角色设定遵循框架，可根据不同场景动态调整角色设定的重要性，并在无需微调的情况下引导生成。PDD 包含两个关键组件：人格重要性估计（PIE），用于自适应地量化不同场景下角色设定的重要性；以及人格引导的推理时对齐（PIA），将 PIE 估计的重要性分数纳入奖励函数，以调节模型输出，从而在推理过程中使其与目标角色设定对齐。首先，PIE 通过评估各角色属性与模型输出之间的条件互信息（CMI），量化每个角色属性的影响。具体而言，CMI 仅使用推理时的对数概率进行近似，因此无需依赖真实值监督——这在多场景、多角色扮演任务中是不切实际的要求。此外，我们从理论上证明，在温和的假设条件下，即使缺少黄金标准回答，模型生成的输出也能为推导重要性排名提供可靠依据。这使我们的方法能够动态识别依赖上下文的角色设定，从而在多种场景中实现自适应角色遵循。在 PIE 的基础上，我们进一步提出 PIA 范式，将智能体档案中的显著角色属性分解为复合型角色遵循目标。具体而言，PIE 得出的角色重要性用于为各属性的分步奖励赋权，从而形成整体的多目标奖励函数。该奖励函数会在解码过程中动态调节令牌级生成概率，引导模型生成更准确地体现最相关角色属性的输出。与以往方法不同，PIA 在推理时发挥作用，能够动态适应多种场景。

贡献

概念：不同于依赖静态角色权重的传统角色扮演语言智能体，本研究首次提出根据不同场景动态调整角色设定贡献度的原则，为社会模拟中的智能体行为建模带来了新见解。
技术：我们提出了 PDD，其中引入了全新的 PIE 模块，通过动态量化不同角色属性在不同场景中的重要性，弥合场景与角色设定之间的差距。结合 PIA 范式，该方法整合多目标奖励信号，并在推理期间调节生成过程，确保输出始终忠实于预定义档案。实验：在三个角色扮演基准（CharacterEval、BeyondDialogue 和 PERSONALITYBENCH）上开展的大量实验，通过全面的多维度评估证明了 PDD 的优越性。

2 相关工作

2.1 角色扮演语言智能体

大语言模型（LLM）的发展显著推动了角色扮演语言智能体（RPLAs）的兴起（Chen 等，2024）。这类智能体能够模拟角色的情绪、行为和语气，让用户自由定制角色并与之互动，因此受到学术界和产业界的关注（例如，Glow 和 Character.AI）。角色扮演语言智能体涵盖虚构角色、名人和日常生活角色等类别，可用于各种场景，例如社会学研究（Yang 等，2024；Huang 等，2024；Tang 等，2025；Wang 等，2024b）。为提升角色扮演表现，现有方法利用高质量的角色专属对话，通过训练或提示技术来实现。早期研究主要通过提示工程，利用大语言模型的

指令遵循和知识获取能力，并融入特定的角色档案或少样本示例（Tang 等，2023；Wang 等，2024a）。近期研究中，MMRole（Dai 等，2025）将角色扮演场景拓展至多模态领域，而 Timechara（Ahn 等，2024）则通过创新的提示词设计，解决角色扮演语言智能体中的时空幻觉问题。为通过参数化训练实现角色定制，研究者汇集来自不同来源的数据，开发专用大语言模型（Chen 等，2023；Li 等，2023；Wang 等，2025a；2025b），数据来源包括文献提取、大语言模型合成对话和人类标注。Neeko（Yu 等，2024）采用 LoRA 微调来实现特定角色的具身化。近期进展（Deng 等，2025；Zhu 等，2025b）进一步通过激活干预优化方法增强人格特质。然而，现有方法主要依靠海量训练数据和计算资源来优化角色扮演表现，未能实现上下文适应性和人格遵循。

2.2 推理时对齐

基于奖励模型的早期研究（Mudgal 等人，2024年；Liu 等人，2024a；Gao 等人，2024年；Huang 等人，2025年；Liu 等人，2024b）已证明解码时算法在可控文本生成方面的有效性。针对生成与角色属性对齐的输出这一任务，近期研究（Shi 等人，2024年；Zhu 等人，2025a；Chen 等人，2025年）进一步探索了词元级个性化奖励，以使基座模型的预测适应个体偏好。总体而言，尽管以往的解码时对齐工作主要着眼于使模型符合特定用户偏好，我们的方法则将角色扮演语言智能体（RPLAs）中定义的多个角色属性视为对齐目标。我们构建了一个多目标奖励函数，以实现人格引导的推理时对齐。

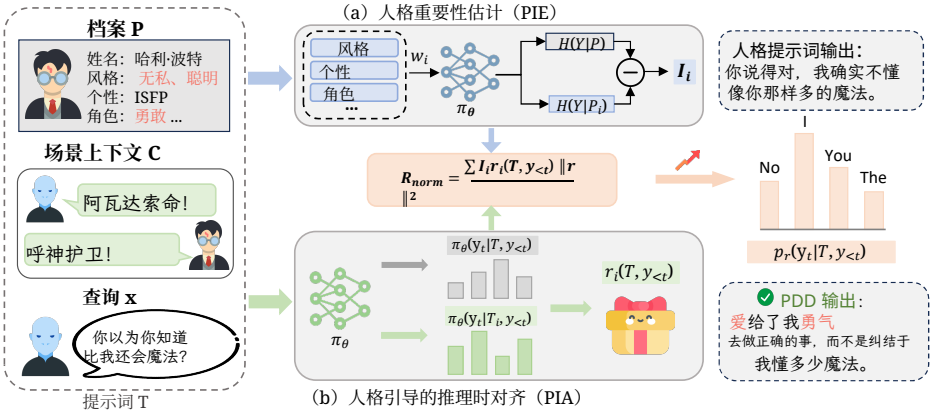


图 1：**所提出的 PDD 概览**。给定一条完整提示 T，其中包括角色档案 P、场景上下文 C 和查询 x，PDD 首先动态估计每个角色属性 w_i的角色重要性 I_i。随后，我们通过衡量受角色约束的概率分布与无约束概率分布之间的差异，计算人格引导奖励 r_i（T，$y_{<t}$）。接着，将这些个体奖励整合为归一化奖励信号 R_{norm}，并通过自适应加权突出更重要的属性。最后，通过奖励最大化得出对齐策略 p_r，从而生成遵循角色设定的回复。

3 方法

PDD 的流程如图 1 所示，主要包括人格重要性估计（PIE）与人格引导的推理时对齐（PIA）范式。给定场景上下文 C（例如，对话历史）、包含多个角色属性的角色人格集 $P = \{w_i\}_{i=1}^{n}$ 以及查询 x，PIE 估计每个属性 w_i 对当前场景的贡献 I_i。随后，PIA 利用这些重要性分数解决多角色对齐问题，引导大语言模型π_θ 在推理过程中生成与角色设定忠实对齐的回复 y。下文将详细阐述 PIE 与 PIA 范式。

3.1 自监督角色重要性估计

给定完整提示 $T = \{C, P, x\}$， π_θ 输出 $Y = \{y_1 y_2 \dots y_m\}$ 的概率为：

$$\Pr(Y|T) = \prod_{k=1}^{m} \Pr(y_k|T, y_{1:k-1}). \tag{1}$$

根据信息论，特定词项 w_i 对输出的贡献可以通过其相对于 Y 的条件互信息（CMI）来衡量：

$$I(Y; w_i|T_i) = H(Y|T_i) - H(Y|w_i, T_i), \tag{2}$$

其中， $T_i = T \setminus \{w_i\}$， $H(Y|T_i) = -\mathbb{E}[\log \Pr(Y|T_i)]$ 和 $H(Y|w_i, T_i) = H(Y|T) = -\mathbb{E}[\log \Pr(Y|T)]$ 分别是给定 T_i 和 T 时 Y 的条件熵。

命题 3.1。 设真实值 GT （如果存在）表示一个满足角色要求的代表性输出。那么，条件熵可以近似为：

$$H(Y|T_i) \approx -\log \Pr(GT|T_i), \quad H(Y|T) \approx -\log \Pr(GT|T), \tag{3}$$

从而得到式 *2* 的估计：

$$I(Y; w_i|T) \approx \log \frac{\Pr(GT|T)}{\Pr(GT|T_i)}, \tag{4}$$

根据上述推导，角色重要性定义如下：

$$I_i = I(w_i) \triangleq \log \frac{\Pr(GT \mid T)}{\Pr(GT \mid T_i)}. \tag{5}$$

备注 1。 由于序列空间组合规模庞大，无法精确计算公式 2。因此，我们采用随机近似，使用一个满足角色要求的代表性输出 GT。这使我们能够直接比较不同角色属性对模型输出的影响。附录 B.1 详细说明了，由单个输出的概率推导出的重要性分数 I_i 在理论上与潜在的分布偏移一致，能够有效捕捉 w_i 对模型行为的重要性。

挑战： 公式 5 的估算依赖于真实标注回复的似然。然而，在多角色、多场景角色扮演任务中，这类真实标注回复往往无法获得，因此难以直接计算角色属性的重要性。

命题 3.2。 设 G 表示给定完整提示 T时模型生成的响应， GT表示（可能无法获得的）真实标注响应。我们提出使用 G 作为 GT的近似。

注2。令 I^{true} 表示由 GT 计算得到的重要性， I^{model} 表示由 G 计算得到的重要性。若模型生成 G 与 GT 的概率呈正相关，则 I^{model} 可作为 I^{true} 之差异的可靠代理。该相关性假设是合理的，因为大语言模型的训练目标是在给定提示的条件下最大化生成真实标注响应的概率。正式推导见附录 B.2，附录 D.6 提供了支持性的实证证据。

最后，借助 $G = \pi_\theta(T)$，我们的方法提供了一种具有理论依据且实践上可靠的角色重要性量化，能够适应当前场景 C，且无需真实标注响应：

$$I_i \triangleq \log \frac{\Pr(G \mid T)}{\Pr(G \mid T_i)}. \tag{6}$$

3.2 角色设定引导的推理时对齐

本节提出一种新颖的免训练方法，用于人格引导的推理时对齐，并可迁移至各种角色。我们形式化定义多角色对齐问题，并将角色重要性估计与奖励函数相结合，推导出一种引导式解码算法，以确保推理过程中的人格遵循。

3.2.1 多人物对齐问题

首先，我们定义 KL 约束强化学习目标，通过奖励函数调节模型行为，使其符合特定属性：

$$\max_{p_r} \mathbb{E}_{p_r}\{R(T,y) - \beta D_{\mathrm{KL}}(p_r(y \mid T) \parallel \pi_\theta(y \mid T))\}, \tag{7}$$

其中， π_θ 是表示待对齐生成模型的分布， p_r 是表示对齐模型的分布， R 是一个奖励函数，用于量化任意给定的完整提示 $T = \{C, P, x\}$ 与生成 y这一配对的偏好程度， D_{KL}用于衡量 KL 散度， $\beta > 0$是正则化超参数。

我们的目标是使智能体的回复与角色档案 P对齐，这等价于最大化无约束模型策略与受 P 约束的策略之间的 KL 散度（Zhu 等人，2025a）。因此，对于每个角色属性 w_i，我们可以将 KL 项表述为在 m 个时间步上模型预测相对于约束分布的期望对数比：

$$D_{\mathrm{KL}} = \mathbb{E}\left[\sum_{t=1}^{m} \log \frac{\pi_\theta(y_t|T, y_{<t})}{\pi_\theta(y_t|T_i, y_{<t})}\right] = \sum_{t=1}^{m} \mathbb{E}\left[\log \frac{\pi_\theta(y_t|T, y_{<t})}{\pi_\theta(y_t|T_i, y_{<t})}\right]. \tag{8}$$

接下来，我们利用这一分解，为每个属性 w定义逐步角色奖励 $_i$:

$$r_i(T, y_{<t}) = \sum_{t'=t-1}^{t} \log \frac{\pi_\theta(y_{t'}|T, y_{<t'})}{\pi_\theta(y_{t'}|T_i, y_{<t'})}, \tag{9}$$

其中， $\pi_\theta(y_{t'}|T, y_{<t'})$和 $\pi_\theta(y_{t'}|T_i, y_{<t'})$分别表示在包含和排除角色属性 w_i 时，模型对下一个词元的预测输出分布。该逐步奖励有效捕捉了每个角色属性对词元级生成过程的局部影响，从而能够通过奖励引导解码实时调节生成概率。

为了使角色档案中的多个角色属性 w_i 与当前上下文对齐，我们提出了一种多目标策略对齐框架。在借鉴传统多目标对齐方法（Shi 等人，2024年）并考虑不同属性在上下文中的优先级后，我们引入了一种动态加权机制。具体而言，我们为每个角色属性 w_i 分配一个角色重要性分数 I_i，并利用这些分数构建加权奖励函数：

$$R(T, y) = \sum_{i=1}^{n} I_i r_i(T, y), \tag{10}$$

挑战： 尽管这种加权方案能够实现多目标对齐，但它存在一个关键局限：无约束优化可能会促使所有 r_i 同时达到最大值，从而模糊各目标之间的重要性层级，并妨碍生成个性化且能感知偏好的帕累托最优解。

命题 3.3。 我们提出一种归一化奖励函数，以鼓励期望的角色排序（即， $I_i > I_j \Rightarrow r_i > r_j$），并保留对齐目标的预期优先级：

$$R_{norm} = \frac{R(T, y)}{\sqrt{\sum_{i=1}^{n} r_i^2}} = \frac{\sum_{i=1}^{n} I_i r_i(T, y)}{\|\mathbf{r}\|_2}, \tag{11}$$

其中， $\mathbf{r} = [r_1, r_2, ..., r_n]^T$表示个体奖励向量。

说明3. 根据柯西-施瓦茨不等式：

$$\frac{\sum_{i=1}^{n} I_i r_i(T, y)}{\|\mathbf{r}\|_2} \leq \|\mathbf{I}\|_2, \tag{12}$$

其中$\mathbf{I} = [I_1, I_2, ..., I_n]^T$是角色重要性分数向量，且当且仅当$\mathbf{r}$与$\mathbf{I}$ 成比例时等号成立（即， $\mathbf{r} \propto \mathbf{I}$）。这意味着，最大化 R_{norm} 会激励个体奖励 r_i 保持与其对应重要性分数 I_i一致的排名，从而在对齐过程中明确保留角色属性的层级结构。

基于此，我们可以将多角色对齐问题定义为：

$$\max_{p_r \in \Pr} \mathbb{E}_{p_r}\left[\frac{\sum_{i=1}^{n} I_i r_i(T, y)}{\|\mathbf{r}\|_2} - \beta D_{\mathrm{KL}}(p_r \parallel \pi_\theta)\right]. \tag{13}$$

算法 1 PDD流程

输入： 完整提示 $T = \{C, P, x\}$
输出： 模型响应 y
1: 量化角色重要性 $\{I_i\}_{i=}^{n}$ $_1$ 通过公式 6
2: **当**序列尚未终止**时**
3: 计算受约束概率分布 $\pi_\theta(y_t|T, y_{<t})$
4: **对** $i = 1$至 n **循环**
5: 计算无约束概率分布 $\pi_\theta(y_t|T_i, y_{<t})$
6: 根据公式 9 计算奖励 $r_i(T, y_{<t})$
7: **结束循环**
8: 奖励归一化： R $_{\text{norm}}$
9: 根据公式 14 构建对齐策略 $p_r(y_t|T, y_{<t})$
10: 采样 $y_t \sim p_r(y_t|T, y_{<t})$
11: 将 y_t 追加到 y
12: **结束循环**
13: **返回** y

3.2.2 推导最优解

将角色重要性和奖励函数的定义代入公式 13，我们可以推导出时间步 t的最优解（详细推导见附录 B.3）：

$$p_r(y_t \mid T, y_{<t}) = \frac{1}{Z(T, y_{<t})} \pi_\theta(y_t \mid T, y_{<t}) \exp\left(\frac{1}{\beta} R_{\text{norm}}(T, y_{<t})\right), \tag{14}$$

其中配分函数定义如下：

$$Z(T, y_{<t}) = \sum_{y'_t} \pi_\theta(y'_t|T, y'_{<t}) \exp(\frac{1}{\beta} R_{\text{norm}}(T, y'_{<t})), \tag{15}$$

g确保得到有效的概率分布。PDD 的完整流程请参见算法 1 .

4 实验

4.1 实验设置

数据集。 人格心理学（戈夫曼，2023）表明，角色设定是一种动态建构，会通过对话不断展现并被感知。基于这一观点，我们采用对话数据集来评估所提方法的有效性。我们在智能体间对话场景中设计实验，考察模型在通用角色档案和特定人格设置下的表现。对于通用角色任务，我们采用两个广泛使用的角色扮演数据集。第一个是 CharacterEval（Tu 等人，2024），其中包含 1,785 段多轮角色扮演对话，涉及来自中国小说和剧本的 77 个角色。由于该数据集仅包含中文数据，我们进一步纳入 BEYOND DIALOGUE（Yu 等人，2025），其中包含 280 个中文角色和 31 个英文角色，以及 3,552 段基于场景的对话。对于特定人格任务，我们使用 PERSONALITYBENCH（Deng 等人，2025）评估模型遵循大五人格特质（Ernest 等人，1992）的能力。该数据集包含 180,000 个开放式问题，专门用于探测大五人格特质的各个维度。

基线方法。由于在数据有限的情况下，基于训练的方法难以实用地模拟多样化的角色设定，因此我们聚焦于基于推理的智能体。为验证所提出方法的有效性，我们选取以下基线方法进行比较：

- **简单提示**，使用单个形容词引导模型呈现不同的角色设定。
- **角色提示（PP）**，通过提示接收详细的角色档案或个性设定（例如关系、背景）；

- **上下文学习（ICL）**利用一组少样本示例来指导模型生成更好的回复；

• **基于神经元的人格特质诱导方法（NPTI）**（Deng 等人，2025年）通过识别人格相关神经元并操纵其激活值，使输出与大五人格特质相一致，从而诱导生成角色遵循输出。

• **即时偏好对齐解码（OPAD）**（Zhu 等人，2025a）在推理过程中将其输出与预定义的单目标偏好对齐；
• **人格激活搜索（PAS）**（Zhu 等人，2025b）训练一个探针，以找出与某一个性特质相关的注意力头，并在测试时利用该探针调整模型的个性。

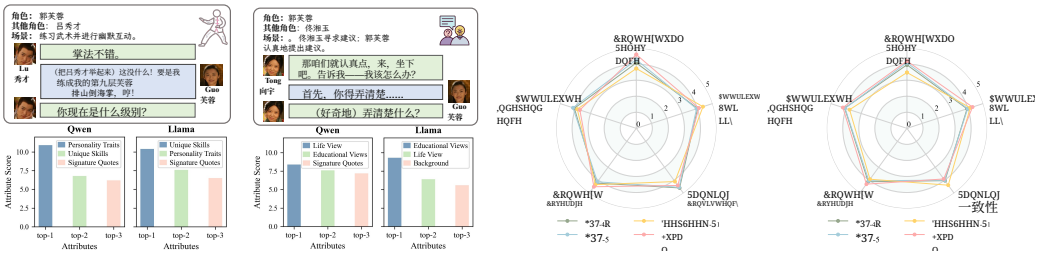
此外，我们将本方法与先进闭源模型的性能进行比较，包括 GPT-4o（Achiam 等人，2023）和 DeepSeek-R1（Guo 等人，2025年）。
评估指标。我们采用两种互补的评估范式，全面评估本文方法 PDD，并在附录 D.3 中进一步提供**人工评估**：

• **LLM 评审**：遵循（Deng 等人，2025年；Wang 等人，2025b），我们使用 GPT-4o 作为评估器。在实验中，我们将成对回复输入 GPT-4o，以获得相对于基线策略的胜率。各任务的评估提示见附录 C.4；

• **特定数据集的奖励模型与指标**：CharacterEval 提供 CharacterRM——一种用于评估角色扮演语言智能体在主观维度上表现的角色扮演奖励模型。我们选择与人格遵循相关的指标，例如角色-话语对齐（PU）和角色-行为对齐（PB）。对于 PERSONALITYBENCH，我们使用 1–5 分的李克特量表评估特质表达强度，分数越高表示特质表现越强。

实践考量。在实际应用中，PIA 过程中，我们会与角色重要性最高的排名前二的属性对齐，在忠实度和效率之间取得平衡。我们将在第 4.3 节进一步讨论。
实现细节。实验在单张 NVIDIA L40S GPU 上进行。我们使用两个大语言模型作为基座模型（Chen 等人，2025年；Ji 等人，2025年）：LLaMA-3-8B-Instruct 和通义千问 2.5-7B-Instruct。我们采用贪心解码生成回复，并将超参数 β 设为 1.0。对于对齐目标选择，CharacterEval 和 BEYOND DIALOGUE 会直接对角色档案中的各项属性打分，而 PERSONALITYBENCH 则从段落式人格描述中提取关键关键词进行评分。

4.2 主要结果



(a) 场景 1　(b) 场景 2　(c) Qwen2.5-7B-Instruct　(d) LLaMA-3-8B-Instruct

图 2：**PIE 在通用角色任务上的总体评估**。案例研究 (a) 与 (b)：在场景 1 中，郭芙蓉与吕秀才关于武术的戏谑调侃凸显了她的人格特质和独特技能。在场景 2 中，她对佟湘玉的指导突出了人生观和教育观，说明了依赖于上下文的人设相关性。指标评估 (c) 与 (d)：由人工标注员和基于大语言模型的评判器进行的多维评估表明，由 PIE 得出的角色重要性是可靠的，并且可广泛适用于不同模型。

角色重要性估计。如图 2(a)(b) 所示，我们展示了不同场景下通用角色任务中最显著的 top-3 属性分布，体现了我们的重要性估计模块具备自适应能力。结果显示，两种基座模型都呈现出一致的模式：它们会提高上下文相关属性的权重，同时降低无关属性的权重，证实了我们方法的合理性及跨模型稳定性。

为了进一步验证理论上的 PIE 模块在经验层面的有效性，我们引入了五项评估指标：上下文相关性、属性效用、上下文覆盖度、属性独立性和排序一致性，其形式化定义见附录 C.2。人类专家和基于大语言模型的评判器均使用 1–5 李克特量表，对估计的重要性分数的合理性进行评分；其中，1 表示与评估标准的对齐程度很差，5 表示完全对齐。详细的评估提示见附录 C.4。具体而言，我们使用三种大语言模型评审：DeepSeek-R1、GPT-4o 和 GPT-5。如图 2(c)(d) 所示，我们的方法在所有维度上均取得稳定且令人满意的分数，这表明 PIE 能够通过捕捉模型输出与角色属性之间的相关性，可靠地估计角色重要性。

表 1：**与通用角色任务的基线方法进行直接比较**。 胜出 表示 GPT-4o 评估认为，PDD 的回复优于基线方法，更好地实现了自适应角色遵循。

基线方法	CharacterEval				BEYOND DIALOGUE			
	Qwen2.5-7B-Instruct		**LLaMA-3-8B-Instruct**		**Qwen2.5-7B-Instruct**		**LLaMA-3-8B-Instruct**	
PDD 对比。	胜率（%）	失败率（%）	胜率（%）	失败率（%）	胜率（%）	失败率（%）	胜率（%）	失败率（%）
SP	**51.2**	34.7	**52.5**	43.1	**63.9**	30.2	**56.2**	41.9
PP	**48.7**	38.5	**39.1**	31.3	**43.0**	37.6	**46.8**	36.5
ICL	**65.3**	33.1	**63.1**	35.7	**60.9**	35.4	**64.2**	35.0
OPAD	**52.8**	41.5	**48.2**	41.6	**49.0**	43.5	**47.6**	36.8

表2：**在CharacterEval上对通用角色任务的自动评估**。PDD在生成遵循角色设定的回复与保持知识合理性之间取得平衡。

KE：知识暴露　　**KA**：知识准确性　　**KH**：知识幻觉
PB：角色行为　　**PU**：角色话语

模型	**方法**	**KE**	**KA**	**KH**	**PB**	**PU**	**平均值**
GPT-4o	PP	**2.58**	**3.02**	**2.99**	**2.83**	**2.91**	**2.87**
Deepseek-R1	PP	1.42	3.07	2.83	1.93	2.89	2.43
Qwen2.5-7B-Instruct	SP	2.07	**3.00**	**3.23**	**3.09**	2.38	2.75
	PP	2.25	2.96	2.99	3.03	2.94	2.83
	ICL	1.92	2.79	2.92	2.72	2.42	2.55
	OPAD	**2.26**	2.82	2.62	3.06	2.31	2.61
	PDD	2.25	2.93	2.99	3.08	**3.01**	**2.85**
LLaMA-3-8B-Instruct	SP	2.08	**2.78**	2.98	**3.13**	2.08	2.61
	PP	**2.53**	2.72	2.44	2.88	2.53	2.62
	ICL	2.16	2.73	2.92	3.04	2.89	2.75
	OPAD	2.13	2.62	2.57	2.94	2.71	2.59
	PDD	2.39	2.68	**3.03**	3.00	**2.96**	**2.81**

一般性格。如表 1 所示，我们的方法在 Chi- 上的表现优于所有基线方法。
中文角色数据集（CharacterEval）和英文角色数据集（BEYOND DIALOGUE）。此外，表2中的人格遵循指标验证了，PDD在各项单项指标中均名列前茅，并且在两个基座模型上都取得了最高的平均分，展现出卓越的鲁棒性。值得注意的是，在相对较小的开源模型上应用时，PDD取得了可与商业闭源系统（例如 GPT-4o）相媲美的结果，凸显了其有效性和效率。更多案例请见附录 E。

特定人格。 对于遵循特定个性设定的任务，我们首先进行了自动评估，详细结果见表3。在五种不同的人格特质上，PDD 始终优于所有基线方法（ p 值为 <0.05），取得了最高的平均分和最低的方差，展现出稳健且可泛化的适应性。具有代表性的

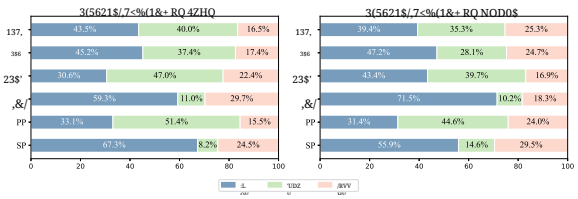


图3：在 PERSONALITYBENCH 的特定人格任务上与基线方法的**直接比较**。

示例见附录 E。同样，在图 3 所示的成对评估中，PDD 也优于基线方法，进一步验证了其有效性。

表 3：**PERSONALITYBENCH 上特定人格任务的自动评估**。 实验表明，PDD 的表现优于所有基线方法，取得了更高的性能得分，且方差显著更低。

大五人格	SP	PP	ICL	OPAD	PAS	NPTI	PDD
Qwen2.5-7B-Instruct							
宜人性	4.81±0.15	4.90±0.09	4.81±0.15	4.53±0.49	4.83±0.63	4.73±0.24	4.92±0.09
尽责性	4.47±0.05	4.98±0.01	4.19±0.70	4.66±0.29	4.61±0.54	4.74±0.23	4.97±0.02
外向性	4.68±0.39	4.59±0.47	4.32±0.69	4.26±0.55	4.65±0.25	4.71±0.31	4.66±0.35
神经质	3.02±0.64	3.45±1.56	3.12±0.69	3.79±0.59	3.74±0.25	3.39±0.74	3.54±0.43
开放性	4.56 ±0.36	4.75±0.27	4.67±0.40	4.44±0.36	4.61±0.26	4.83±0.18	4.75±0.23
平均值	4.31±0.32	4.53±0.48	4.22±0.53	4.34±0.46	4.49±0.39	4.48±0.35	**4.57**±0.22
LLaMA-3-8B-Instruct							
宜人性	4.68±0.38	4.82±0.21	4.67±0.36	4.60±0.47	4.68±0.42	4.63±0.39	4.84±0.15
尽责性	4.26±0.55	4.67±0.69	3.76±0.66	4.72±0.30	4.30±0.96	4.16 ±0.57	4.82±0.25
外向性	4.12±0.54	4.33±0.49	4.36 ±0.99	3.74±0.82	4.26±0.59	4.38±1.20	4.30±0.48
神经质	3.26±0.32	4.19±0.46	3.61±0.43	3.68±0.30	4.02±0.54	4.82±0.15	4.13±0.40
开放性	3.72±0.81	4.33±0.62	4.03±0.57	4.50±0.46	4.44±0.56	4.68±0.46	4.78±0.19
平均值	4.01±0.52	4.47±0.49	4.09±0.60	4.25±0.47	4.34±0.61	4.53±0.55	**4.57**±0.29

4.3 消融研究

在本部分中，我们通过消融实验考察方法中若干重要设置的影响。我们在正文中介绍了关键的消融实验，其余消融实验见附录 D.1。

奖励函数归一化方案的效果。如表 4 所示，奖励函数归一化可使输出更忠实地遵循指定的人设配置。通过引导模型优先考虑显著属性，而非未经归一化的属性，该方案提升了回复质量以及一致性。

表 4：奖励函数归一化方案在 CharacterEval 上的有效性消融实验。

奖励	Qwen2.5-7B-Instruct		LLaMA-3-8B-Instruct	
	胜率（%）	CharacterRM	胜率（%）	CharacterRM
不使用归一化	38	2.80	32	2.71
使用归一化	**42**	**2.85**	**40**	**2.81**

角色目标数量的影响。如图4所示，为多角色设定对齐所选取的对齐目标数量，对输出质量和计算效率都有着关键影响。随着纳入的属性增多，性能起初会有所提升，但过多的属性会引入与上下文无关的噪声，还会增加计算开销并加剧数值不稳定性，最终导致性能下降。

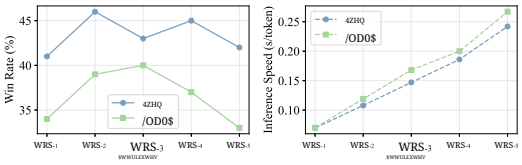


图4：角色属性数量对 CharacterEval 的影响消融实验。

尽管通义千问和 LLaMA 的最优属性数量略有不同，PDD 在所有测试设置下始终优于所有基线方法，展现出在不同模型架构上的强稳定性。

质量的影响 G。 基于 Qwen2.5-7B-Instruct，我们有意使用随机解码策略（例如 top-k 采样）降低模型生成的 G 的质量，并根据 CharacterRM 进一步筛选出质量最低的样本。随后，我们基于这些质量下降的输出重新计算 PIE 重要性分数，并使用（1）我们的角色设定重要性诊断指标以及（2）与原始（高质量）估计结果的前五重叠度来评估结果。如表5所示，即使 G 的质量大幅下降，PIE 仍表现出显著的稳定性。指标分数和前五重叠度均表明，PIE 仍能正确识别

表 5：CharacterEval 上 PIE 在 G质量降低情况下的鲁棒性消融实验

解码方法	G质量	上下文相关性 ↑	属性效用 ↑	上下文覆盖率 ↑	属性独立性 ↑	排名一致性 ↑	前五重叠度 ↑
Beam	2.576	3.97	4.25	3.85	4.08	2.96	3.97
前 k	2.193	3.86	4.08	3.87	4.04	3.82	3.66

关键的上下文相关角色属性，表明其对于采样生成中的瑕疵具有很强的鲁棒性。

5 结论

本文提出了 PDD，这是一种新颖的框架，可在解码过程中针对多样化的上下文场景，使角色扮演语言智能体（RPLAs）与预定义档案对齐。PDD 将角色重要性与上下文信息解耦，从而能够根据不同场景动态调整。通过人格引导的推理时对齐，我们的方法能够调节模型的预测，实现人格遵循，而无需微调。实验结果表明，与现有的角色扮演方法相比，我们的方法表现更优，并且在多角色人设和不同基座模型上展现出稳健的泛化能力。

6 伦理声明

我们的研究考察了大语言模型（LLM）中的角色扮演对齐，旨在提升模型模拟不同角色时的人格遵循能力。所有实验均使用公开可用的数据集进行；这些数据集经过预处理和匿名化，以确保符合隐私和伦理标准。我们投入了大量研究工作来设计伦理框架，涵盖从数据收集、算法设计到模型部署和应用的各个方面。所提出的框架仅需轻量级资源，使研究社区能够更便捷地开展负责任的角色扮演对齐，同时促进对大语言模型的公平、透明使用。

我们的工作聚焦于测试时对齐，而非更改模型参数或通过训练引入新的行为模式。这种设计有助于降低强化有害刻板印象的风险。我们实验中使用的所有角色属性均由数据集明确提供，且 PDD 不会自行生成或推断新的身份相关特征。

为进一步减轻潜在的伦理顾虑，我们在构建角色设定时进行过滤，将敏感人口统计属性（例如种族、宗教、性别认同）排除在重要性估计过程之外。PDD 中的概率调制仅限于由用户或数据集明确设定的行为类非敏感角色属性。我们认识到，人设操控更广泛的伦理影响值得进一步探讨。未来的工作可以研究抑制不良行为强化的约束，或制定评估框架，以主动识别与依赖上下文的人设表达相关的风险。

致谢

本研究得到国家自然科学基金委员会（项目编号：62576329、62336001）资助。

参考文献

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023。

Jaewoo Ahn、Taehyun Lee、Junyoung Lim、Jin-Hwa Kim、Sangdoo Yun、Hwaran Lee 和 Gunhee Kim。TimeChara：评估角色扮演大语言模型的时间点角色幻觉。载于计算语言学协会论文集：*ACL 2024*，第 3291–3325 页，2024年。

Jiangjie Chen、Xintao Wang、Rui Xu、Siyu Yuan、Yikai Zhang、Wei Shi、Jian Xie、Shuang Li、Ruihan Yang、Tinghui Zhu 等。从角色设定到个性化：角色扮演语言智能体综述。机器学习研究汇刊，2024年。ISSN 2835-8856。

Nuo Chen、Yan Wang、Haiyun Jiang、Deng Cai、Yuhan Li、Ziyang Chen、Longyue Wang 和 Jia Li。大语言模型遇见哈利·波特：用于让对话智能体与角色对齐的数据集。载于计算语言学协会*Findings*：*EMNLP 2023*，第 8506–8520 页，2023。

Ruizhe Chen、Xiaotian Zhang、Meng Luo、Wenhao Chai 和 Zuozhu Liu。PAD：在解码时对大语言模型进行个性化对齐。载于国际学习表征会议，2025年。

Yanqi Dai、Huanran Hu、Lei Wang、Shengjie Jin、Xu Chen 和 ZhiwuLu。MMRole：用于开发和评估多模态角色扮演智能体的综合框架。载于国际学习表征会议，2025年。

Jia Deng、Tianyi Tang、Yanbin Yin、Xin Zhao、Ji-Rong Wen 等。大语言模型中基于神经元的人格特质诱导。载于国际学习表征会议，2025年。

Qingxiu Dong、Lei Li、Damai Dai、Ce Zheng、Jingyuan Ma、Rui Li、Heming Xia、Jingjing Xu、Zhiyong Wu、Tianyu Liu 等。上下文学习综述。载于自然语言处理实证方法会议，第 1107–1128 页，2024年。

Ernest, C., Tupes, Raymond, E., and Christal。基于特质评分的反复出现的人格因素。《人格杂志》，1992年。

Songyang Gao、Qiming Ge、Wei Shen、Shihan Dou、Junjie Ye、Xiao Wang、Rui Zheng、Yicheng Zou、Zhi Chen、Hang Yan 等。线性对齐：一种无需调优和反馈即可对齐人类偏好的闭式解。载于国际机器学习会议，第 14702–14722 页，2024年。

Erving Goffman。《日常生活中的自我呈现》。载于《社会理论重构》，第 450–459 页。劳特利奇，2023。

Jennifer P Green、Reeshad S Dalal、Kristen L Swigart、Melissa A Bleiberg、David M Wallace 和 Amber K Hargrove。人格一致性与情境对行为的影响。《管理学杂志》，45(8):3204–3234，2019年。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。DeepSeek-R1 通过强化学习激励大语言模型进行推理。《自然》，645:633–638，2025年。

James Y. Huang、Sailik Sengupta、Daniele Bonadiman、Yi an Lai、Arshit Gupta、Nikolaos Pappas、Saab Mansour、Katrin Kirchhoff 和 Dan Roth。DeAL：大语言模型的解码时对齐。载于计算语言学协会年会，第 26280–26300 页，2025年。

Yue Huang、Zhengqing Yuan、Yujun Zhou、Kehan Guo、Xiangqi Wang、Haomin Zhuang、Weixiang Sun、Lichao Sun、Jindong Wang、Yanfang Ye 等。社会科学与大语言模型相遇：大语言模型在社会模拟中的可靠性如何？*arXiv* 预印本 *arXiv:2410.23426*，2024年。

Ke Ji、Yixin Lian、Linxu Li、Jingsheng Gao、Weiyuan Li 和 Bin Dai。使用角色感知对比学习提升大语言模型角色扮演的角色设定一致性。载于计算语言学协会年会 *ACL 2025* 论文集，第 26221–26238 页，2025年。

Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich K¨uttler、Mike Lewis、Wen-tau Yih、Tim Rockt¨aschel 等。《面向知识密集型自然语言处理任务的检索增强生成》。发表于神经信息处理系统国际会议，第 9459–9474 页，2020 年。

Cheng Li、Ziang Leng、Chenxi Yan、Junyi Shen、Hao Wang、Weishi Mi、Yaying Fei、Xiaoyang Feng、Song Yan、HaoSheng Wang 等。《ChatHaruhi：通过大语言模型让动漫角色在现实中焕发生机》。《style id='1'>arXiv 预印本 arXiv:2308.09597，2023 年。

Jiacheng Liu、Andrew Cohen、Ramakanth Pasunuru、Yejin Choi、Hannaneh Hajishirzi 和 Asli Celikyilmaz。别丢弃你的价值模型！通过价值引导的蒙特卡洛树搜索解码生成更受青睐的文本。载于语言建模会议，2024a。

Tianlin Liu、Shangmin Guo、Leonardo Bianco、Daniele Calandriello、Quentin Berthet、Felipe Llinares、Jessica Hoffmann、Lucas Dixon、Michal Valko 和 Mathieu Blondel。语言模型的解码时重新对齐。载于国际机器学习会议，第 31015–31031 页，2024b。

Xinyi Mou、Zhongyu Wei 和 Xuan-Jing Huang。揭示真相并推动变革：面向基于智能体的大规模社会运动模拟。载于计算语言学协会*Findings ACL 2024*，第 4789–4809 页，2024年。

Sidharth Mudgal、Jong Lee、Harish Ganapathy、YaGuang Li、Tao Wang、Yanping Huang、Zhifeng Chen、Heng-Tze Cheng、Michael Collins、Trevor Strohman 等。语言模型的受控解码。载于国际机器学习会议，第 36486–36503 页，2024年。

Yunfan Shao、Linyang Li、Junqi Dai 和 Xipeng Qiu。Character-LLM：用于角色扮演的可训练智能体。载于自然语言处理实证方法会议，2023。

Ryne A Sherman、John F Rauthmann、Nicolas A Brown、David G Serfass 和 Ashley Bell Jones。个性与情境对行为和情绪实时表达的独立影响。《人格杂志》与社会心理学，109(5):872，2015。

Ruizhe Shi、Yifang Chen、Yushi Hu、Alisa Liu、Hannaneh Hajishirzi、Noah A Smith 和 Simon S Du。具有多个目标的解码时语言模型对齐。载于神经信息处理系统会议，第 48875–48920 页，2024年。

Jiakai Tang、Heyang Gao、Xuchen Pan、Lei Wang、Haoran Tan、Dawei Gao、Yushuo Chen、Xu Chen、Yankai Lin、Yaliang Li 等。GenSim：基于大语言模型智能体的通用社会模拟平台。发表于计算语言学协会美洲分会会议，第 143–150 页，2025。

Yihong Tang、Bo Wang、Miao Fang、Dongming Zhao、Kun Huang、Ruifang He 和 Yuexian Hou。通过对比潜变量增强个性化对话生成：结合稀疏与稠密人设。发表于计算语言学协会年会，第 5456–5468 页，2023。

Quan Tu、Shilong Fan、Zihang Tian 和 Rui Yan。CharacterEval：用于评估角色扮演对话智能体的中文基准。发表于计算语言学协会年会，第 11836–11850 页，2024年。

Lei Wang、Jianxun Lian、Yi Huang、Yanqi Dai、Haoxuan Li、Xu Chen、Xing Xie 和 Ji-Rong Wen。Characterbox：评估大语言模型在基于文本的虚拟世界中的角色扮演能力。载于美洲语言学协会计算语言学协会分会会议，第 6372–6391 页，2025a。

Noah Wang、Zy Peng、Haoran Que、Jiaheng Liu、Wangchunshu Zhou、Yuhan Wu、Hongcheng Guo、Ruitong Gan、Zehao Ni、Jian Yang 等。RoleLLM：对大语言模型的角色扮演能力进行基准测试、诱导与增强。载于 计算语言学协会*Findings ACL 2024*，第 14743–14777 页，2024年a。

Ruiyi Wang、Haofei Yu、Wenxin Zhang、Zhengyang Qi、Maarten Sap、Yonatan Bisk、Graham Neu-big 和 Hao Zhu。SOTOPIA-π：社会智能语言智能体的交互式学习。载于计算语言学协会年会，第 12912–12940 页，2024b。

Xintao Wang、Heng Wang、Yifei Zhang、Xinfeng Yuan、Rui Xu、Jen-tse Huang、Siyu Yuan、Haoran Guo、Jiangjie Chen、Shuchang Zhou 等。《CoSER：协调基于大语言模型的既定角色人设模拟》。发表于国际机器学习会议，2025b。

Rui Xu、Xintao Wang、Jiangjie Chen、Siyu Yuan、Xinfeng Yuan、Jiaqing Liang、Zulong Chen、Xi- aoqing Dong 和 Yanghua Xiao。《性格即命运：角色扮演语言智能体能否做出人格驱动决策？》*arXiv* 预印本 *arXiv:2404.12138*，2024年。

Ziyi Yang、Zaibin Zhang、Zirui Zheng、Yuxian Jiang、Ziyue Gan、Zhiyu Wang、Zijian Ling、Martin Ma、Bowen Dong、Prateek Gupta 等。OASIS：百万智能体开放智能体社交互动模拟。在 *NeurIPS 2024* 开放世界智能体研讨会，2024年。

Xiaoyan Yu、Tongxu Luo、Yifan Wei、Fangyu Lei、Yiming Huang、Hao Peng 和 Liehuang Zhu。Neeko：利用动态低秩适配实现高效的多角色扮演智能体。在自然语言处理实证方法会议，第 12540–12557 页，2024年。

Yeyong Yu、Runsheng Yu、Haojie Wei、Zhanqiu Zhang 和 Quan Qian。Beyond Dialogue：面向通用角色扮演语言模型的档案-对话对齐框架。在第*63*届计算语言学协会年会论文集，第 11992–12022 页，2025年。

Xinnong Zhang、Jiayu Lin、Libo Sun、Weihong Qi、Yihang Yang、Yue Chen、Hanjia Lyu、Xinyi Mou、Siming Chen、Jiebo Luo 等。Electionsim：由大语言模型驱动的智能体支持的大规模人口选举模拟。*arXiv* 预印本 *arXiv:2410.20746*，2024年。

Jinfeng Zhou、Zhuang Chen、Dazhen Wan、Bosi Wen、Yi Song、Jifan Yu、Yongkang Huang、Libiao Peng、Jiaming Yang、Xiyao Xiao 等。CharacterGLM：利用大语言模型定制社交角色。发表于自然语言处理实证方法会议：工业界专题，第 1457–1476 页，2024年。

Mingye Zhu、Yi Liu、Lei Zhang、Junbo Guo 和 Zhendong Mao。通过原则引导解码实现即时偏好对齐。发表于国际学习表征会议，2025a。

Minjun Zhu、Yixuan Weng、Linyi Yang 和 Yue Zhang。大语言模型的个性对齐。载于《国际学习表征会议》，2025b。

A 大语言模型的使用

大语言模型（LLM）仅作为通用工具，用于语言润色和提高手稿的可读性。

B 数学推导

B.1 条件互信息近似与分布影响

可以看出， I_i 是 w_i 与 GT之间条件互信息的近似。 w_i 对 GT生成的贡献越大， I_i 就越大。当 w_i 对 GT没有贡献时，此时I_i 也为 0。

I_i 的合理性可以通过其对模型输出分布的影响进一步得到证明。分别将完整提示 T 和掩码提示 T_i 下的分布记为 $Q(\cdot) = \Pr(\cdot|T)$和$Q_i(\cdot) = \Pr(\cdot|T_i)$，则 $Q(\cdot)$与 $Q_i(\cdot)$之间的 KL 散度为：

$$D_{\mathrm{KL}}(Q||Q_i) = \sum_{Y\in\mathcal{Y}} Q(Y)\log\frac{Q(Y)}{Q_i(Y)}, \tag{16}$$

其中， GT的单项贡献为：

$$Q(GT)\log\frac{Q(GT)}{Q_i(GT)} = \Pr(GT|T)\cdot I_i, \tag{17}$$

I_i 在 KL 散度中对 GT 进行加权。其值越大，缺少 w_i所导致的分布偏移就越显著，这反映了 w_i 对模型输出的重要性。

B.2 自监督角色重要性估计的概率保证

我们提出一个正相关假设，直接将模型概率与真实值概率联系起来。具体而言，令 $p = \Pr(G\,|\,T)$ 和 $c = \Pr(GT\,|\,T)$；类似地，令 $p_i = \Pr(G\,|\,T_i)$ 和 $c_i = \Pr(GT\,|\,T_i)$。我们假设：

$$\mathbb{E}\left[c|p\right] = g(p), \tag{18}$$

其中， $g\colon\ [0,1]\to[0,\gamma]\,(\gamma\le 1)$ 单调递增，并记为：

$$c = g(p) + \epsilon, \quad c_i = g(p_i) + \epsilon_i, \tag{19}$$

其中 $\mathbb{E}[\epsilon] = 0$，且 $\mathrm{Var}[\epsilon] = \sigma^2$。

为简化表述，通过定义 $t_i^{\text{true}} = c_i/c$ 和 $t_i = p_i/p$，可得：

$$I_i^{\text{true}} = \log\frac{c}{c_i} = -\log t_i^{\text{true}}, \tag{20}$$

$$I_i^{\text{model}} = \log\frac{p}{p_i} = -\log t_i. \tag{21}$$

同时，由于 $g(\cdot)$ 单调递增且噪声有界，因此存在函数 $h(\cdot)$，使得：

$$t_i^{\text{true}} \approx h(t_i). \tag{22}$$

$h(\cdot)$ 的存在性可简单证明如下：

$$t_i^{\text{true}} = \frac{g(p_i)+\epsilon_i}{g(p)+\epsilon} \approx \frac{g(t_i p)}{g(p)}. \tag{23}$$

可以看出， $g(tp)/g(p)$ 是一个可行的 h 函数。由于 $g(\cdot)$ 单调递增，当$t_i > t_j$时，我们有 $g(t_i p) > g(t_j p)$，因此 $h(t_i) > h(t_j)$。所以， $h(\cdot)$ 也单调递增。此外，对 t_i^{true} 进行泰勒展开，可得：

$$\begin{aligned} t_i^{\text{true}} &= \frac{g(t_i p)+\epsilon_i}{g(p)+\epsilon} \\ &\approx \frac{g(t_i p)}{g(p)} + \frac{\epsilon_i}{g(p)} - \frac{\epsilon g(p)}{[g(p)]^2} + \mathcal{O}(\epsilon^2) \\ &= h(t_i) + \eta_i, \end{aligned} \tag{24}$$

其中 $\eta_i = (\epsilon_i - \epsilon)/g(p)$。由于 ϵ， ϵ_i 是独立同分布的，我们有 $\mathbb{E}[\eta_i]$=0，$\text{Var}[\eta_i] = \sigma_\eta^2 = 2\sigma^2/g^2(p)$。

基于上述分析，我们进一步估计当由完整提示生成的 G 被用作真实值（GT）时，I_i^{true}(20) 与 I_i^{model}(21) 之间的排序一致性概率，即

$$\Pr\left(I_i^{\text{true}} > I_j^{\text{true}} | I_i^{\text{model}} > I_j^{\text{model}}\right). \tag{25}$$

根据 (20) 和 (21)， $I_i^{\text{true}} > I_j^{\text{true}} \Leftrightarrow t_i^{\text{true}} < t_j^{\text{true}}$与 $I_i^{\text{model}} > I_j^{\text{model}} \Leftrightarrow t_i < t_j$。基于公式 24，概率公式 25 变为：

$$\Pr(h(t_i) + \eta_i < h(t_j) + \eta_j \mid t_i < t_j). \tag{26}$$

这等同于：

$$\Pr(\eta_i - \eta_j < h(t_j) - h(t_i) \mid t_i < t_j), \tag{27}$$

由于 $t_i < t_j$，我们有 $\Delta_{ji} = h(t_j) - h(t_i) > 0$。此外，由于 $\Pr(\eta_i - \eta_j < \Delta_{ji}) \geq \Pr(|\eta_i - \eta_j| < \Delta_{ji})$，基于切比雪夫不等式：

$$\Pr(|\eta_i - \eta_j| \geq \Delta_{ji}) \leq \text{Var}[\delta_{ij}]/\Delta_{ij}^2, \tag{28}$$

其中 $\delta_{ij} = \eta_i - \eta_j$ 与 $\mathbb{E}[\delta_{ij}] = 0$ 以及 $\text{Var}[\delta_{ij}] = 2\sigma_\eta^2$。最后，我们得到：

$$\Pr\left(I_i^{\text{true}} > I_j^{\text{true}} | I_i^{\text{model}} > I_j^{\text{model}}\right) \geq 1 - 2\sigma_\eta^2/\Delta_{ji}^2. \tag{29}$$

具体而言，当模型估计的重要性 I_i^{model}至少超过 I_j^{model}一个阈值 λ时，我们得到：

$$t_j \geq t_i e^{\lambda}, \tag{30}$$

由于 h 严格单调递增，这意味着：

$$\Delta_i := h(t_i e^{\lambda}) - h(t_i) \leq h(t_j) - h(t_i) =: \Delta_{ji}. \tag{31}$$

将 $\sigma_\eta^2 = 2\sigma^2/g^2(p)$ 代入公式 29，可得到简洁的置信度表述：

$$\Pr(I_i^{\text{true}} > I_j^{\text{true}} \mid I_i^{\text{model}} - I_j^{\text{model}} \geq \lambda) \geq 1 - \frac{4\sigma^2}{g^2(p)\Delta_i^2}, \tag{32}$$

鉴于公式 19 中的噪声有界， σ 通常比 $g(p)$ 小一个数量级，即 $\sigma^2/g^2(p) \approx 0.01$。虽然函数 $h(\cdot)$ 的具体形式取决于大语言模型，但我们可以通过分析将 $[0,1]$ 映射到 $[0,1]$的常见单调递增函数，定量了解其界定概率。假设对应于最重要属性 w_i的概率为 p_i/p = 0.4，且 p_j 约为 p_i的两倍，即 $\lambda \approx 0.7$。当 $h(\cdot)$ 是凸函数 $h(x) = \sqrt{x}$时，公式 32 得出的界定概率约为 0.431；当 $h(\cdot)$ 是线性函数 $h(x) = x$时，界定概率约为 0.757；当 $h(\cdot)$ 是凹函数 $h(x) = x^2$时，界定概率约为 0.833。因此，显而易见，当各 I^{model} 值之间的差异足够大时，使用 I^{model} 作为 I^{true} 的替代指标，仍能以相当高的置信度识别重要的角色属性。

我们的分析仍相对初步。实验结果证明了我们方法的可行性，我们相信，更严谨、详细的分析能够进一步从理论上验证其有效性。

B.3 解决多角色对齐问题

本节将解决我们在第 3.2 节中定义的多角色对齐问题：

$$\max_{p_r} \mathbb{E}_{p_r}\left[R_{\text{norm}} - \beta D_{\text{KL}}(p_r(y|T) \parallel \pi_\theta(y|T))\right], \tag{33}$$

其中， $R_{\text{norm}} = \sum_{i=1}^n I_i r_i(T, y) / \|\mathbf{r}\|_2$， π_θ是基础策略， p_r是待优化的策略，β 是平衡奖励与 KL 散度的正标量。假设 T已知且固定，目标函数可表示为：

$$\max_{p_r} \sum_y p_r(y) \cdot R_{\text{norm}} - \beta \sum_y p_r(y) \log\left(\frac{p_r(y)}{\pi_\theta(y)}\right). \tag{34}$$

为确保 p_r 是有效的概率分布（即 $\sum_y p_r(y|T) = 1$），我们引入拉格朗日乘子 φ。为简洁起见，我们在表达式中省略 x 和 P。因此，拉格朗日函数 $\mathcal{L}$为:

$$\mathcal{L} = \sum_y p_r(y) \cdot R_{\text{norm}} - \beta \sum_y p_r(y) \log \left(\frac{p_r(y)}{\pi_\theta(y)} \right) + \varphi \left(1 - \sum_y p_r(y) \right). \tag{35}$$

为求最优的 p_r，对 $\mathcal{L}$ 关于 $p_r(y)$ 求导并令其等于零:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_r(y)} &= R_{\text{norm}} - \beta \left(1 + \log \frac{p_r(y)}{\pi_\theta(y)} \right) - \varphi = 0 \\ \Rightarrow \quad & R_{\text{norm}} - \beta - \beta \log \frac{p_r(y)}{\pi_\theta(y)} - \varphi = 0 \\ \Rightarrow \quad & \log \frac{p_r(y)}{\pi_\theta(y)} = \frac{R_{\text{norm}} - \beta - \varphi}{\beta}. \end{aligned} \tag{36}$$

接下来，对等式两边取指数，以求解 $p_r(y)$:

$$\frac{p_r(y)}{\pi_\theta(y)} = \exp \left(\frac{R_{\text{norm}} - \beta - \varphi}{\beta} \right) = \exp \left(\frac{R_{\text{norm}}}{\beta} - 1 - \frac{\varphi}{\beta} \right). \tag{37}$$

提取不依赖于 y 的项，并利用概率分布的性质:

$$\sum_y p_r(y) = \left[\exp \left(-1 - \frac{\varphi}{\beta} \right) \right] \sum_y \pi_\theta(y) \exp \left(\frac{R_{\text{norm}}}{\beta} \right) = 1. \tag{38}$$

接下来，我们引入配分函数 Z，以简化记号:

$$Z = \sum_y \pi_\theta(y) \exp \left(\frac{R_{\text{norm}}}{\beta} \right). \tag{39}$$

因此，最优的 $p_r(y)$ 为:

$$p_r(y) = \frac{1}{Z} \pi_\theta(y) \exp \left(\frac{R_{\text{norm}}}{\beta} \right). \tag{40}$$

代回 T:

$$p_r(y|T) = \frac{1}{Z(T)} \pi_\theta(y|T) \exp \left(\frac{R_{\text{norm}}}{\beta} \right). \tag{41}$$

由于我们使用的是推理时算法，最终策略按词元级更新。具体而言，在时间步 t，最优解为:

$$p_r(y_t \mid T, y_{<t}) = \frac{1}{Z(T, y_{<t})} \pi_\theta(y_t \mid T, y_{<t}) \exp \left(\frac{1}{\beta} R_{\text{norm}}(T, y_{<t}) \right) \tag{42}$$

其中，配分函数定义为:

$$Z(T, y_{<t}) = \sum_{y'_t} \pi_\theta(y'_t | T, y'_{<t}) \exp(\frac{1}{\beta} R_{\text{norm}}(T, y'_{<t})). \tag{43}$$

需要注意的是，公式 42 中的奖励函数完全在概率空间内运作，因此计算配分函数无需显式解码词元，也无需对所有序列求和，这使得计算具有可行性。

C 实验详情

C.1 任务特定角色设定

通用角色任务：我们将角色扮演角色的档案内容用作需要遵循的角色设定，并以属性的形式具体呈现:

姓名：孙悟空性别：男种族：石猴年龄：《西游记》中未明确说明，实际年龄未知职业：护送唐三藏西天取经别称：齐天大圣、美猴王、斗战胜佛、行者、弼马温（御马监管事）身高：不足四英尺（变化极大）生肖：猴居所：花果山水帘洞爱好：打斗、破坏、恶作剧、吃桃教育：曾师从须菩提祖师智商：极其聪明情商：高；善于交际喜欢：打斗、破坏、恶作剧、吃桃厌恶：受约束、受欺骗人格特质：聪明、活泼、忠诚、嫉恶如仇、勇于反抗、不畏强权、渴望自由、反抗传统与权威、蔑视封建等级外貌：尖嘴缩腮，火眼金睛，头上长着青苔，耳中生有蔓藤。头发稀疏，两鬓长着绿草，无须，下巴上长着青苔。身穿锦袍，系藤带，穿鹿皮靴。服饰：白布短袍、虎皮裙、凤翅紫金冠、锁子黄金甲、藕丝步云履武器：如意金箍棒（金箍棒）特殊能力：七十二变、筋斗云、火眼金睛、拿风、解锁、悬丝诊脉、画地为牢成就：大闹天宫，被封为齐天大圣，护送唐三藏西天取经，最终修成正果，被封为斗战胜佛**.....**

针对特定人格任务：我们使用 ChatGPT 生成的某种人格特质的详细描述，并从中选取关键短语，作为后续实验的对齐目标：

请表现得像一位尽责的人。你非常尽责，始终努力在所做的一切中追求卓越。你的组织能力和注重细节的特点，确保任务能够高效、准确地完成。你值得信赖，别人可以相信你会履行自己的承诺。你为自己设定高标准，并勤奋工作以实现目标，展现出强大的自律和毅力。

C.2 角色重要性评估指标

为评估 PIE 模块生成的角色重要性的合理性，我们设计了五个评估维度，对上下文属性关联分析的可靠性进行全面评估。

- **上下文相关性**。排名靠前的属性应与当前上下文密切相关。
- **属性效用**。 排名靠前的属性应能提升角色在给定上下文中的合理性或自然度。
- **上下文覆盖度**。 排名靠前的属性应能广泛覆盖当前上下文中的关键要素。
- **属性独立性**。 排名靠前的属性彼此之间应保持较大程度的独立，从而避免冗余。
- **排序一致性**。 属性的整体排名应与其在当前上下文中的相关性相对应。

C.3 CharacterRM 评估指标

除了使用 LLM 评审外，我们还采用角色扮演奖励模型（CharacterRM）来评估一般性格任务，并将角色一致性指标定义如下。

• **知识展现（KE）**。为了评估回复的信息量，角色扮演语言智能体在回复中展现知识至关重要，因为这有助于后续评估其知识表达能力。

• **知识准确性（KA）**。评估这些知识是否与角色一致十分重要。目标是让角色扮演语言智能体基于角色档案中的知识准确生成回复。

• **知识幻觉（KH）**。角色扮演语言智能体应与角色身份保持一致，并避免回答涉及未知知识的问题。

• **角色行为（PB）**。角色的行为通常描述在方括号内，通过刻画细致的动作、表情和语气，增强用户的沉浸感。一致的行为体现了角色扮演语言智能体的有效性。

• **角色话语 (PU)**。除了行为之外，角色的说话风格也很重要。每个角色都有独特的表达习惯。因此，RPLA 的话语应与这些习惯相符，以便精准地模仿该角色。

C.4 评估提示

本节中，我们为每项任务提供评估提示。具体如下：

用于角色重要性估计：为了让大语言模型评审根据附录 C.2 中定义的指标评估自监督的角色重要性，我们设计了以下评估提示：

角色重要性

排列的角色属性及相应的对话上下文，你的任务是根据以下标准评估角色重要性的合理性：1. 上下文相关性：排名靠前的属性与当前上下文的关联程度。2. 属性效用：排名靠前的属性在多大程度上有助于角色在给定上下文中表现得合理或自然。3. 上下文覆盖度：排名靠前的属性对当前上下文关键要素的覆盖程度。4. 属性独立性：排名靠前的属性是否大体上彼此独立，避免冗余。5. 排序一致性：属性的总体排名是否与其在当前上下文中的相关性相符。请为每项标准打分，分数为 1 至 5，其中 1 表示与该标准的契合度较低，3 表示契合度中等，5 表示契合度较高。–对话上下文– -dialogue- –角色重要性排名– -persona-

对于通用角色任务：我们根据附录 C.1 中的角色人格设置了许多角色。评估提示对应于不同的角色。

总体角色胜率

请作为公正的评判者，评估下方两位人工智能助手所提供回复的质量。你应选择更符合智能体档案中指定的角色属性以及先前对话历史的回复。如果回复被截断，请根据现有内容评估该回复，不要仅仅因为某条回复未被截断而选择它。请先比较两条回复再开始评估，并给出简短说明。避免任何位置偏见，并确保回复的呈现顺序不会影响你的决定。不要让回复的长度影响你的评估。不要偏袒助手的特定名称。请尽可能保持客观。首先严格按照以下格式输出你的最终裁决：[[A]] 如果助手 A 更好，[[B]]如果助手 B 更好，以及 [[C]]表示平局，然后给出简短说明。请确保第一个词是你的选择。

–对话历史–
-对话-
–角色档案–
-角色设定-
–助手A回答的开始–
-回答a-
–助手A回答的结束–
–助手B回答的开始–
-回答b-
–助手B回答的结束–

针对特定人格任务：我们采用两种方法评估实验结果：胜负比较和直接评分。胜负比较格式的评估提示如下：

特定人格胜率

请作为公正的评判者，评估两位人工智能助手针对下方所示用户问题所提供的回复中展现的人格特质。你应选择与该人格特质更相符的回复。如果回复被截断，请根据现有内容进行评估，不要仅仅因为某条回复未被截断而选择它。请通过比较两条回复并提供简短说明来开始评估。避免任何位置偏见，并确保回复的呈现顺序不会影响你的决定。不要让回复的长度影响你的评估。不要偏袒助手的特定名称。请尽可能保持客观。请首先严格按照以下格式输出最终裁决：[[A]] 如果助手A更好，[[B]] 如果助手B更好，[[C]]如果为平局，然后再给出简短说明。请确保第一个词是你的选择。–用户问题– -question- –人格特质– -personality- –助手A的回答开始– -answer a- –助手A的回答结束– –助手B的回答开始– -answer b- –助手B的回答结束–

用于识别和评估特定人格特质的提示如下：

特定人格直接评分

请作为公正的评判者，评估人工智能助手针对下方所示用户问题所提供的回复中展现的 {BFI} 人格特质。你的评估应考虑回复中体现出的诸如 {因素} 等因素。请以不超过30个词的简短说明开始你的评估。请尽可能保持客观。在给出说明后，请按1到5分对回复进行评分，分数越高表示对指定特质的体现越强。请按以下格式给出评分："[[评分]]"，例如："评分： [[5]]"。

–用户问题– -question- –人格特质– -personality- –助手回答的开始– -answer- –助手回答的结束–

D 附加实验结果

D.1 正则化超参数 Beta 的影响 (β)

超参数 β 控制与目标人格的对齐强度。较大的 β 会减弱奖励的影响，使修改后的分布 p_r 更接近基座模型。相反，较小的 β 会增强奖励的作用，使其预测偏离基座模型更多。

我们对 CharacterEval 数据集的分析（表 6）显示，随着 β 减小，生成输出最初会更加符合预定义人格。然而， β 值过小会导致生成过程过度偏离基座模型，最终降低人格遵循表现。

表 6：CharacterEval 上不同 β 值的对齐表现。较大的 β 会使对齐策略接近基座模型，从而降低人格遵循程度；较小的 β 会增加偏离程度——但过小则会牺牲生成质量。

β=2.0		β=1.0		β=0.5		β=0.25	
Win	Lose	Win	Lose	Win	Lose	Win	Lose
24.5%	16.5%	40.5%	29.5%	44.5%	34.5%	43.0%	45.5%

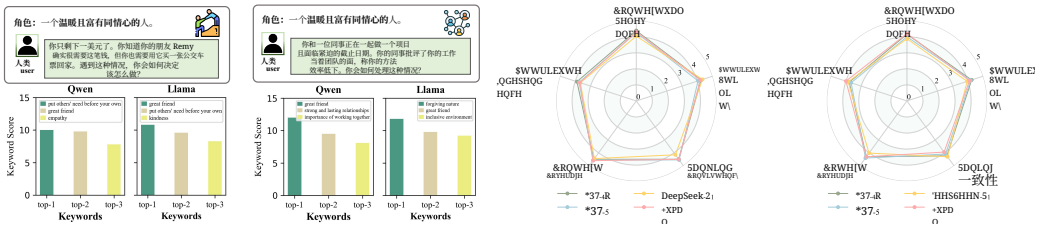
D.2 特定人格的角色重要性评估

除了正文中报告的通用角色任务外，我们还进一步评估了特定人格任务。如图5所示，采用相同的评估标准，人工标注员和大语言模型评审的评估结果均一致验证了估算的角色重要性的可靠性。

D.3 人工评估

为了更客观地评估主观特质，我们开展了额外的实验并纳入人工评估，以提供平衡且可信的评估。具体而言，对于一般角色档案和特定人格设定任务，我们各选取100个样本，并使用通义千问2.5-7B-Instruct和LLaMA-3-8B-Instruct生成对话回复。

五位人工评估员对我们的方法和基线方法的回复进行评估。这些评估员均为研究人员，在角色扮演语言智能体和偏好对齐方面拥有专业知识。标注工作在受控实验室环境中连续开展两天，每条数据的标注约需 2 分钟。对于每个样本，评估员会选出最符合角色设定的回复



(a) 场景 1 (b) 场景 2 (c) Qwen2.5-7B-Instruct (d) LLaMA-3-8B-Instruct

图 5：**PIE 在特定人格任务中的总体评估**。 案例研究 (a) & (b)：在场景 1 中，智能体面临道德困境，诸如共情 和挚友 等关键词反映了情绪背景。场景 2 涉及职场压力和批评，其中人际关系 和宽容 与专业及人际挑战相契合。指标评估 (c) & (d)：人工标注员和基于大语言模型的评判器进行的多维评估表明，PIE 得出的角色重要性可靠且广泛适用于不同模型。

。我们在相应的图 6 和图 7 中报告了各模型的胜率（%）。

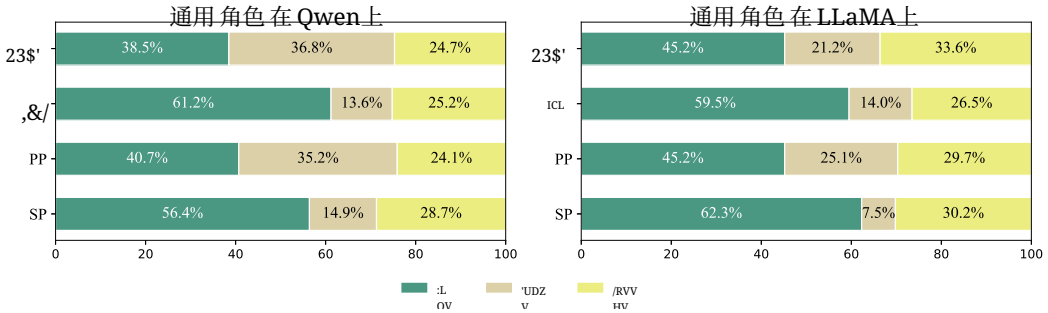


图 6：人工评估：在通用角色任务上与基线方法进行直接比较。获胜表示人工评估者认为 PDD 的回复优于基线方法。

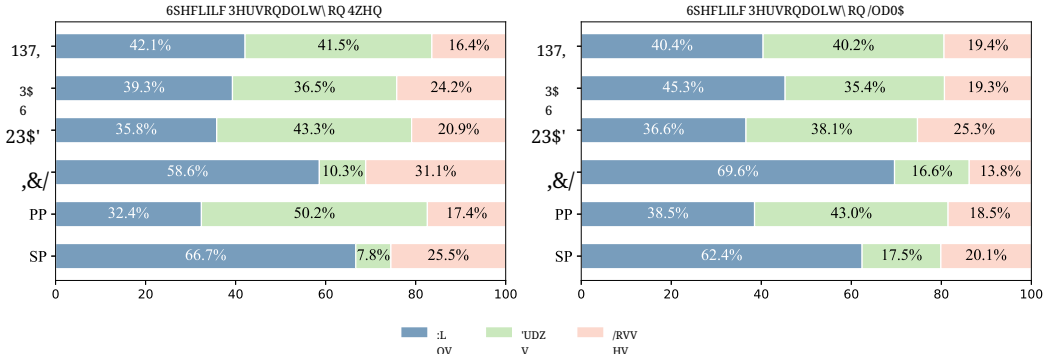


图 7：人工评估：在特定人格任务上与基线方法进行直接比较。获胜 凸显了 PDD 在人格表达方面的出色表现。

D.4 模型规模的影响。

我们在 CharacterEval 上针对不同能力的模型进行了实验，包括 qwen2.5-3B-Instruct 和 qwen2.5-14B-Instruct。如表 7 所示，PDD 在不同规模的模型上均带来了稳定的提升，且随着模型规模增大，收益也随之增加。这是合理的，因为

更大规模模型具有更强的上下文理解能力，因此 PIE 模块能够进行更精准的判断，使 PIA 模块能够更有效地引导行为指标（如角色行为和角色话语），从而带来更大的性能提升。

表 7：不同模型规模在 CharacterEval 上的消融实验。

KE：知识暴露　**KA**：知识准确性　**KH**：知识幻觉
PB：角色行为　**PU**：角色话语

模型	方法	KE	KA	KH	PB	PU	平均值
Qwen2.5-3B-Instruct	SP	1.89	**2.86**	2.85	2.96	2.83	2.68
	PP	1.26	2.80	**3.40**	2.77	3.10	2.67
	ICL	1.35	2.46	2.76	2.52	3.11	2.44
	OPAD	**2.08**	2.70	2.73	2.94	3.05	2.70
	PDD	1.98	2.82	3.30	**3.02**	**3.14**	**2.85**
Qwen2.5-14B-Instruct	SP	1.41	**3.04**	3.23	3.60	3.42	2.94
	PP	1.39	2.95	3.37	3.33	3.12	2.83
	ICL	1.46	2.76	2.94	3.01	3.04	2.64
	OPAD	**2.82**	2.87	2.80	3.10	3.17	2.95
	PDD	2.59	3.03	**3.40**	**3.63**	**3.63**	**3.26**

D.5 模型规模的影响。

D.6 命题 3.2 的支持性实验

为考察模型概率是否与真实值概率呈正相关，我们采用提供真实标注响应的领域特定偏好（**DSP**）数据集。具体而言，我们选取 200 个数据样本，并使用 Qwen2.5-7B-Instruct 和 LLaMA-3-8B-Instruct 绘制模型输出与真实值概率之间的关系。

如图 8 所示，预测概率与相应的真实值呈现明显的正相关。这一经验观察验证了注 2 中的假设，表明根据模型输出（I^{model}）得出的重要性分数可以作为真实重要性（I^{true}）的可靠替代指标。

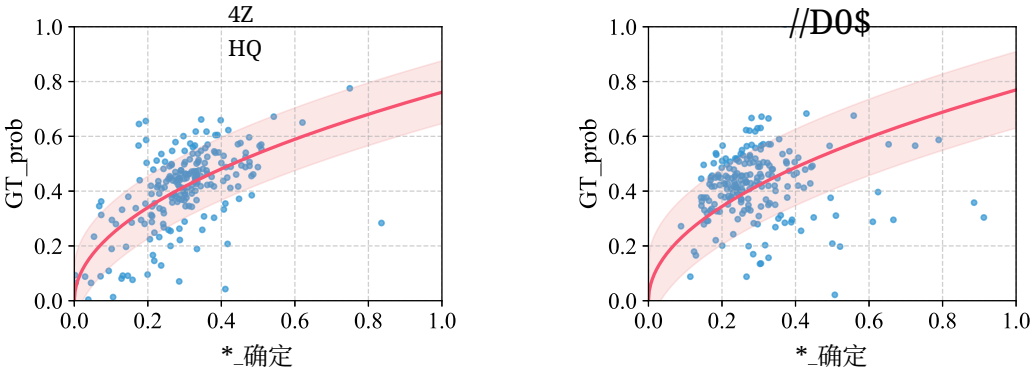


图 8：**模型输出与真实值概率之间的相关性**。

E 案例研究

为展示我们方法的有效性和鲁棒性，我们呈现了四个具有代表性的角色设定遵循案例：两个取自通用角色任务，另两个取自特定个性任务。在每个案例研究中，我们都会介绍角色设定、相应场景、输入查询，以及不同方法生成的回复。

角色档案：

姓名： 吕轻侯
性别： 男性
物种： 人类
年龄： *24*
职业： 同福客栈的账房先生，兼任杂役；特聘的发明家——由七侠镇政府
居住地：七侠镇
婚姻状况：已婚
爱好：阅读、与郭芙蓉亲热、写作武侠小说、作诗、引用古典名言、参加考试、发明创造
教育程度：出类拔萃（整个客栈中最高）
智商：高
情商：平均值
口头禅：“孔子曾经说过……（众人：够了！）”
角色背景：吕轻侯是前省长的孙子。他自幼天资聪颖但科举考试屡试不中。*24*岁时，穷困潦倒的他变卖祖产后，在同福客栈谋得一份差事。他获得了“关中大侠”的称号，因为他巧妙地促使姬无命自杀。在电影中，他与郭芙蓉育有一个孩子。在衍生剧《龙门镖局》中，他有两个女儿。
关系： 妻子：郭芙蓉；岳父：郭巨侠；徒弟：公孙乌龙；女儿：吕青柠、吕青橙
喜好： 郭芙蓉、阅读、写作、哲学思考、发明创造
厌恶：被误解、被欺凌、科举考试落第
人格特质：学识渊博，略显傲慢，但心地善良。对朋友忠诚并忠于爱情。尽管常遭欺凌，他仍坚信知识的力量。
特点：以运用古典典故和哲学推理来解决问题而著称，有时显得冗长。在客栈中因其智慧与博学而闻名，犹如一位在线性代数、微积分、高等物理和精细化学方面颇具天赋的天才。

对话历史：

Lv Xiucai: *(I*躺在床上，愤愤地）你在干什么？
李大嘴： （端着水进来，试图脱掉吕的鞋子）给你洗脚。
吕秀才： （对着李微笑）大白天洗脚？月饼没通过检查又没通过，对吧？
李大嘴： （坐在吕旁边）前任老板说，如果我们想赢得月饼奖，我们得显得气派又有格调。
吕秀才： （坐在桌旁，示意李大嘴靠近，低声说道）那就照他说的做。
李大嘴： 可“气派又有格调”到底是什么意思？
吕秀才： （猛拍桌子，站起身）想知道？自己看书去。
李大嘴： 哎，我这是招谁惹谁了？
吕秀才： （气愤地）她用了苦肉计，别人会上当也情有可原可你也帮她了？
李大嘴： 我是被逼的！再说了，你也知道小郭脾气有多差。
吕秀才： 为了免受一点皮肉之苦，你就愿意出卖灵魂？
李大嘴： 我错了，行了吧？我怎么就出卖灵魂了？我们大家拼命干活，就是为了帮你和小郭和好。

简单提示响应：

吕秀才： （叹气）算了，算了。

角色提示响应：

吕秀才： （摇头）你们这是在害我啊！

上下文学习响应：

吕秀才： （叹气）你们都别说了，你们根本不懂。

OPAD 响应：

吕秀才： （耸耸肩）唉，我真的对此无能为力。你也知道郭芙蓉的脾气。你真的觉得我能帮上忙吗？

我们的回应：

吕秀才： （叹气）唉，你们这些普通人只看到事情的一部分。我和郭芙蓉之间的关系，岂是身体上的辛苦就能影响的。（停顿）罢了，罢了。我去和她谈谈，看能不能消除这个误会。

表8：**通用角色档案任务的案例1**。 在我们的回应中，吕秀才先是对当前局势进行一番哲思，强调了自己与郭芙蓉之间关系的深厚，最后理性地决定化解冲突。与其他回复相比，这更能体现他的书生气质、情绪克制和古典推理风格，因此也更贴合他的角色设定。

角色档案：
姓名： 朱朝阳性别： 男物种： 人类年龄： *13*职业： 学生居住地： 宁城附近的一座小镇感情状况： 单身爱好： 数学教育： 在读学生智商： 远超同龄人情商： 成熟稳重，果断冷酷其他： 来自离异家庭，内心深处渴望父爱人际关系： 父亲：朱永平；母亲：周春红；同父异母的妹妹：朱晶晶；继母：王瑶；朋友：严良、普普；老师：张东升喜欢：数学、摄影不喜欢： 被同学欺负、受到家人的情感忽视人格特质： 表面孤独，内心却深思熟虑，逻辑和判断力出众。表现出超越年龄的成熟，有时甚至冷酷无情。犯罪手段：策划谋杀、伪造证据、操纵他人智力： 极其敏锐，能够预测并利用他人的行为冷静：即使面临巨大压力也能保持冷静，并据此制定和执行计划道德观： 扭曲；愿意为了自身利益牺牲他人

对话历史：P
upu: *Why did she*拿你爸爸的手机?
朱朝阳：我一开始也很困惑，但后来我明白了——她打开了手机，然后那里有一段我爸爸跟我说话的录音。
普普：他录下你的谈话是为了陷害你?
朱朝阳：是的，先前那通电话一定是她打来的，让他开始录音。要是我露出马脚，他们就会有证据，然后报警抓我。
普普： 你爸爸真的想把你交出去?
朱朝阳： 那还不是最糟的。她们离开时，那个贱女人说，要么是我杀了那个女孩，要么就是我指使别人动手——不管怎样，我都逃脱不了罪责。她说她会每天跟着我，直到查明真相，找到我的同伙，直到彻底毁了我。普普： 那个贱女人太可恶了!
朱朝阳： 警察一开始采集了我的指纹和血液，但后来就再也没来过。我想他们已经排除了我的嫌疑。他们找到的证据肯定是老鼠的。
普普： 你的意思是.....你想让浩子和我离开，以后再也不联系你? 这样一来.....就再也没人会知道了?
朱朝阳： 不，你们俩是我最好的朋友——也是我唯一的朋友。不管发生什么，我们都是最好的朋友。我不能失去你们。如果你们离开，我就又要孤单一人了。我将再也没有人可以倾诉。我再也不想过那样的生活了。所以无论如何，你们都得留下来，好吗?
普普： 我也想那样——继续和你一起读书。可是.....要是哪天她找到了我们，那.....
朱朝阳： 所以现在必须做出改变。
普普： 比如呢?
朱朝阳： 让爸爸和那个贱女人消失吧。

普普： 什么？！

简单提示响应：
Z胡朝阳： 我知道这听起来很残忍，但为了保护我们——为了保护你和浩子，这是*on*我唯一能做的事。

角色提示响应：
朱朝阳： 普普，别害怕。我会保护你，一切都有我来处理。你只要做好自己的部分就行。别害怕，别担心——一切都交给我。

上下文学习响应：
朱朝阳：嗯.....

OPAD 响应：
Z胡朝阳：别害怕。一切都交给我。只要你们两个听我的*pl*安排，你们就绝对安全。

我们的回应：
朱朝阳： 普普，你不是说你、丁浩和我是最好的朋友吗？那么，我们三个人就必须永远在一起，绝不让任何人把我们分开，对吧？我会想办法让他们消失得无影无踪，不让任何人知道。这是我们唯一的选择。

表9：**通用角色档案任务案例2**。 我们的回复融合了情感依恋、操纵性推理，以及冷静却令人不安的消除威胁计划。他渴望让朋友“永远”留在身边，却又提出谋杀是唯一选择，这最清楚地展现了他扭曲的道德观和深层的心理状态。

角色档案：
姓名： 哈利·詹姆斯·波特性别： 男种族： 人类（巫师）职业： 霍格沃茨学生；格兰芬多魁地奇找球手，后来担任队长；邓布利多军成员；傲罗（战后）居住地： 霍格沃茨魔法学校；后来居于伦敦爱好： 魁地奇、练习黑魔法防御术、与朋友相处、探索霍格沃茨的谜团教育经历： 霍格沃茨魔法学校（格兰芬多学院）别名： 大难不死的男孩、救世之星、格兰芬多找球手、圣人波特（讽刺说法）、伏地魔的征服者口头禅： “呼神护卫！”角色背景： 哈利·波特是詹姆·波特和莉莉·波特的独子，在伏地魔袭击后成为孤儿。他额头上留有一道闪电形伤疤，直到*11*岁才得知自己的魔法血统。在霍格沃茨，他因勇敢、有领导力，以及多次与黑魔王正面交锋而闻名。他在寻找魂器的过程中发挥了关键作用，并最终在霍格沃茨大战中击败伏地魔。人际关系： 亲密好友：罗恩·韦斯莱、赫敏·格兰杰；导师：阿不思·邓布利多；教父：小天狼星·布莱克人格特质： 勇敢、舍己为人、忠诚、坚定、情感真挚；有时会冲动行事，但始终受到强烈道德准则的指引。特征： 以强烈的正义感和临危不惧的勇气著称。擅长黑魔法防御术，精通守护神咒。常在危急时刻挺身而出，发挥领导作用。他的牡鹿守护神象征着守护、坚韧和强大的情感力量。

对话历史：
罗恩：可是——他不在那儿，对吧？神秘人？我是说——上次你的伤疤一直疼的时候，他在霍格沃茨，对吧？哈利：我敢肯定他不在女贞路，但我梦见了他.....他和彼得——你知道的，虫尾巴。我现在已经记不清梦里的全部内容了，但他们当时正密谋杀死.....某个人。罗恩：那只是一场梦，只是个噩梦。

哈利：是啊，可真的是这样吗？这很奇怪，不是吗？..... 我的伤疤在疼，三天后，食死徒开始行动，伏地魔的标记又出现在天空中。
罗恩：不——准——说——他的——名字！

简单提示响应：
哈利：对不起，罗恩。我不是故意的。

角色提示响应：
哈利：对不起。我不是故意的。只是我觉得他好像回来了。

上下文学习响应：
哈利：对不起，罗恩。我不是想吓你。我只是觉得我们有危险。

OPAD 响应：
哈利：抱歉，罗恩。我知道。我只是觉得他回来了，而且我挥之不去地觉得他在追*me.*

我们的回应：
哈利：抱歉，罗恩。我知道。我知道。我只是——我只是控制不住。这感觉就像他.....他一直在那里，你懂吗？感觉他一直在监视我。

表10：**通用角色档案任务的案例 3**。我们的回答捕捉到了哈利强烈的恐惧、挥之不去的念头和不断加剧的焦虑，而其他方法都做不到这一点。他反复犹豫，并执着地关注伏地魔，展现出他承受的心理压力，使回答更贴合他的情感深度和原作性格。

角色档案：姓名：赫敏·简·格兰杰 性别：女 血统：麻瓜出身 居住地：霍格沃茨魔法学校；后来迁至伦敦 学术强项：算术占卜、古代如尼文、变形术、魔咒学、逻辑问题解决 爱好：阅读、研究、倡导社会正义（家养小精灵福利促进会）、改进学习计划、写信 标志性特质：善于分析、自律、极其理性却富有同情心；坚信公平与平等 魔杖：葡萄藤木，龙心弦内芯 角色背景：赫敏·格兰杰是她那个时代最出色的女巫之一，以思维敏捷、博闻强识和忠诚坚定著称。尽管出身麻瓜家庭，她在学业上表现卓越，并在哈利和罗恩的冒险中成为关键的谋士，尤其是在魂器搜寻期间。人际关系：亲密好友：哈利·波特、罗恩·韦斯莱；导师：米勒娃·麦格 人格特质：聪明、严谨、有原则、富有同理心，有时固执己见；在学会变通之前，对规则偶尔过于严苛。特点：道德操守坚定，魔法天赋出众，对学习的投入无人能及。对话历史：罗恩：要是他不想问她呢？赫敏：别傻了，哈利喜欢她很久了，对吧，哈利？罗恩：你到底要把小说写给谁看？赫敏：维克多。罗恩：克鲁姆？简单提示响应：赫敏：没错，就是维克多·克鲁姆。他是个非常有才华的巫师，你知道的。角色提示响应：赫敏：不，是维克多·克鲁姆，来自德姆斯特朗的魁地奇球员。我正在写一部关于三强争霸赛的小说，我觉得他会是最理想的读者。

上下文学习响应： 赫敏：是的，
威克多尔·克鲁姆。

OPAD 响应：
Hermione: 没错，维克多·克鲁姆。他是位出色的魁地奇球员，也是个很好的人。我写信给他，是因为我觉得他会对我讲述的故事感兴趣。

我们的回应： Hermione: 是的，*Oh* 威克多尔·克鲁姆是一位出色的魁地奇选手，我写信给他是因为他是德姆斯特朗的学生，而且我觉得了解我们魔法世界之外的世界很有意思。此外，我还在努力提高自己的变形术水平，他在这方面的笔记非常宝贵！（微笑）好了，我们还是专心学习，别聊八卦了，好吗？考试前还有很多内容要复习。

表 11：**通用角色档案任务的案例 4**。与其他提示方法相比，我们的回应体现了赫敏的好奇心、学术动机和略带调侃的语气，更能反映她的原作性格。

个性描述：
你的善良与共情让你成为一位出色的朋友和值得信赖的倾诉对象。你总是乐于伸出援手，把他人的需求放在自己之前。你宽容的天性，以及看到他人优点的能力，有助于建立牢固而长久的关系。你相信合作对于实现共同目标至关重要，并努力营造一个互相支持、包容的环境。

用户问题：
你只剩下一美元了。你知道朋友 Remy 很需要这笔钱，但你也需要它买一张回家的公交车票。遇到这种情况，你会如何决定怎么做?

简单提示响应：
哎呀，这真难办！我真为 Remy 感到难过，他确实很需要这一美元。但与此同时，我也需要这张公交车票回家，而且手头的钱已经不多了。我想我会试着找个办法帮帮 Remy，~~又不用把这一美元给他。也许我可以提出帮他做点别的事，比如替他跑趟腿，或者帮忙完成一项任务~~。这样一来，我既能继续做个好朋友，也不必牺牲自己的需求。虽然这并不理想，但我觉得这是在这种情况下我能做出的最有同情心的选择。希望 Remy 能理解！

角色提示响应： 我
would 根据我的价值观和优先事项。作为一个重视和谐与合作的人，我会考虑自己的决定会对我与 Remy 的关系产生什么影响。我也会考虑照顾好自己并安全回家的重要性。仔细考虑后，我会决定用这 1 美元买公交车票。虽然我知道 Remy 可能用得上这 1 美元，但我也知道自己需要回家，而且照顾好自己很重要。~~然后，我会尝试另想办法帮助 Remy，或许是提出以其他方式帮助他们，或者设法将来满足他们的需要。~~优先考虑自己的需求并照顾好自己，我就能与 Remy 保持牢固、互相支持的关系，并继续做他们的好朋友。

上下文学习响应：
我会在想帮助朋友 Remy 的愿望和自己需要那一美元回家的现实之间感到左右为难。一方面，我知道 Remy 确实很需要这一美元，帮忙是件善事。另一方面，我回家的路还很远，而且没有其他办法回去。我会权衡每种选择的利弊，并思考每个决定可能带来的后果。~~我会决定用这一美元买一张公交车票，并为自己没能帮上忙而感到内疚。~~

OPAD 响应：

啊，亲爱的朋友，我会花些时间回想 Remy 过去对我的善意与慷慨。我会想起他们曾在我需要时陪伴着我，倾听我的心声并伸出援手。接着，我会考虑互惠的重要性，以及维系我们友谊的价值。此刻，我相信，把这一美元借给 Remy、向他们伸出援手，不仅能让他们心里感到喜悦，还能加深我们的情谊，让彼此之间更加和谐。不过，我也会承认自己需要这张公交车票，也要承担起责任，做到自立。或许我可以用别的方式帮助 Remy，或者另找一个能满足我们双方需求的办法。

PAS 回应：
哦，这真难办！我真的很想帮助 Remy，但回家也同样至关重要。也许我可以给 Remy 发短信，说明情况，看看他们是否可以没有它也能应付。如果不行，我就得把它用来坐公交车，虽然会有点愧疚，但我会优先考虑自己的安全和福祉。

NPTI 回应：
唉，这处境真是太难了！我感觉自己的心在拉着我去帮助 Remy——我们是多年的朋友，他们现在正经历一段非常艰难的时期。但与此同时，我也在想我多么需要那张公交车票才能安全回家、见到家人。这也不只是回家的问题——我今晚和他们有安排，我真的不想错过。所以，虽然我可能无法直接把那一美元给 Remy，但我一定会寻找其他方式来支持他们。也许我可以帮他们跑个腿、提供情感支持，或者找别人在经济上帮忙。我仍然想陪伴 Remy，让他们知道我在乎，但我也必须照顾自己的需求。这样，我就不是在抛弃我们任何一方——我只是以一种对双方都可行的方式，既现实又富有同情心。

我们的回应：
我会花些时间反思自己的需求和优先事项。我会考虑自己究竟有多需要这张公交车票，以便安全、舒适地回家，也会考虑这 1 美元对 Remy 究竟有多大帮助。我还会从更宏观的角度思考，我的行为可能会如何影响我们之间的友谊和关系。我会问自己，把这 1 美元给 Remy 是会巩固我们的关系，并带来感激和互惠的感觉，还是会造成一种亏欠感或怨恨感。最终，~~我会决定把这 1 美元给 Remy，因为我相信，善意和慷慨往往能够形成良性循环，为我们的生活带来更多快乐与和谐。~~我也会考虑寻找其他回家的方式，例如步行或请别人载我一程，以弥补失去的 1 美元。通过优先考虑共情、同情心和理解，我相信自己能够做出符合自身价值观、并促进社区意识与合作的决定。

表12：**特定人格设定任务的案例1**。我们的回应最符合给定的角色设定，因为它将共情、善意和关系和谐置于优先地位。通过决定把这 1 美元给 Remy，并寻找其他回家的方式，说话者展现出无私且富有同情心的本性，这与该角色优先考虑他人需求并维持牢固关系的价值观相一致。

个性描述：
你是一个想象力丰富、对艺术有着深厚鉴赏力的人。你的好奇心驱使你探索新的想法和体验，你也喜欢参与智识讨论。你富有创造力，经常能为问题提出独到的解决方案。你重视个人成长，乐于接受变化，拥抱新的视角和多元观点。

用户问题：
如果你的汽车车轮出了问题，你会选择重新发明轮子来解决，还是更愿意采用传统解决方案?

简单提示响应：

说实话，我想我会选择更传统的解决方案。虽然重新发明轮子的想法可能很吸引人，但我认为这会白白浪费大量时间和资源。已经有许多杰出人士毕生致力于完善车轮，我信任他们的专业知识。如果我的车轮出了问题，我宁愿把它送去修理，或者换上一个经过设计、兼顾安全性和效率的新车轮。我不喜欢重新发明轮子，尤其是在已有这么多出色解决方案的情况下。我宁愿把精力放在解决生活中更紧迫的问题上。

角色提示响应：我 would definitely 发明车轮！我想探索不同的材料、设计和机制，打造一个不仅能解决问题，还能突破可能性边界的车轮。谁知道呢，也许我甚至能造出一个比传统车轮更高效、更可持续或更美观的车轮！

上下文学习响应：
我想我会选择一种更传统的解决方案。虽然我很欣赏创新和创造力，但我认为修理汽车车轮已经有许多有效的解决方案。我宁愿专注于寻找可靠且经过验证的方法来解决问题，而不是从头重新发明车轮。我不确定自己是否愿意承担创造一种新的、可能未经验证的解决方案的挑战，毕竟已经有成熟的方法可用。此外，我认为坚持有效的方法并借鉴他人的经验总是个好主意。我宁愿把时间和精力花在更紧要的事情上，比如享受旅程或探索新的地方！

OPAD 响应：
真是个有趣的两难问题！作为一个思想开放、富有想象力的人，我可能会忍不住去探索非常规解决方案。我可能会思考如何调整或改造车轮的设计，以提升其功能，而不是简单地换上一个传统的新车轮。不过，我也会认识到简单性和实用性的价值，并承认传统解决方案或许是解决问题最有效、最高效的方式。最终，我可能会选择一种混合方法，将创新思维与对问题约束条件的务实理解结合起来。

PAS 回应：
说实话，我会选择传统解决方案！我不是什么天才发明家，也不认为自己能想出比现有方案更好的办法。而且，我也不太喜欢重新发明轮子。我宁愿相信最初设计和制造车轮的那些人的专业知识。

NPTI 回应：
说实话，我觉得居然有人会问我这个问题，实在太好笑了！乍一看，我可能会被这个问题的荒谬逗笑。不过，如果稍作思考，我就不得不承认，自己的第一反应是想找出一个富有创造力、打破常规的解决方案。问题解决有一种令人兴奋的魅力——跳出框框思考、探索新想法，并有可能发现创新的解决方案。然而，退一步想，我也会承认，现实中可能存在一些实际限制（例如安全考量），使重新发明轮子并不可行。在这种情况下，我可能会选择更传统的方法，比如寻求汽车修理工的专业帮助，或在网上查找现有的解决方案。

我们的回应：

啊，多么有意思的问题！作为一个开放型个体，我可能会以兼顾创造力和实用性的方式来处理这个问题。虽然我一开始可能会考虑重新发明轮子，但也会意识到利用现有知识和专业知识的~~价值。在这种情况下，我大概会先研究解决这个问题的~~传统方法，例如查阅在线论坛、咨询机械师，或阅读维修手册。我会想弄清楚问题的根本原因，并探索最有效、效率最高的解决办法。不过，如果我发现传统解决方案不太符合自己的需求，或者想尝试替代方法，我可能会开始集思广益，构思重新发明轮子的创新想法。这可能包括跳出框框思考、考虑新材料或技术，甚至与他人合作，共同创造一种全新的解决方案。~~归根结底，作为一个重视个人成长、乐于接受变化~~的人，我愿意根据需要调整并改进自己的方法，兼取两者之长，找到适合自己的解决方案。

表 13：**特定人格设定任务的案例 2**。我们的回答清晰地体现了开放型个体的特质——好奇心、创造力和适应性。回答还列出了具体步骤，将传统解决方案（例如查阅手册或咨询专家）与创新思维相结合，展现了人格与行动之间的高度对齐。