PersonaArena: Dynamic Simulation for Evaluating and Enhancing Persona-Level Role-Playing in Large Language Models

Wenlong Shi[1], Jianxun Lian[2]*, Mingqi Wu[3], Haiming Qin[1], Mingyang Zhou[1], Xing Xie[2], Naipeng Chao[4,5], Hao Liao[1,5]*,

[1]College of Computer Science and Software Engineering, Shenzhen University, China
[2]Microsoft Research Asia
[3]Microsoft Gaming
[4]School of Media and Communication, Shenzhen Univerisity
[5]Provincial Key Laboratory of Intelligent Communication and Digital Society Governance, Shenzhen University
2553103006@mails.szu.edu.cn, jianxun.lian@outlook.com, haoliao@szu.edu.cn

Abstract

Large language models (LLMs) increasingly serve as interactive social agents, yet their ability to maintain coherent and authentic persona-level role-playing remains limited, particularly in realistic social scenarios. Existing research predominantly focuses on character-level settings and relies on static evaluation formats, failing to capture the complexity of everyday social interactions. In this work, we present **PersonaArena**, a dynamic simulation framework for evaluating and improving persona-level role-playing in LLMs. PersonaArena leverages a large, filtered corpus of user-generated social content to construct a nuanced persona bank, and elicits multi-turn, context-rich interactions within simulated social environments. Our framework features a multi-agent debating judge for holistic and unbiased assessment. Through extensive experiments, we demonstrate that PersonaArena enables rigorous evaluation and enhancement of LLMs' role-playing capabilities, advancing the development of more authentic and socially adept AI agents. The code for the PersonaArena framework is available at our public GitHub repository: https://aka.ms/personaarena.

1 Introduction

Large language models (LLMs) are increasingly deployed as interactive agents ranging from social companions to virtual simulations. A key to their effectiveness is role-playing: adopting a persona and maintaining coherent behavior over time. Such capabilities yield more engaging, personalized, and believable interactions, which are crucial for sustained user engagement and more lifelike social presence. Yet despite advances, LLMs — especially those of moderate size (e.g. 8B parameters) — often lag in persona fidelity, consistency, and adaptability. (Zhou et al., 2025; Chen et al., 2024a; Samuel et al., 2024). These gaps highlight the need for rigorous evaluation and methods specifically aimed at improving persona-level role-playing in order to realize more authentic AI agents.

Majority of role-playing literature focuses on the character-level setting, in which roles correspond to well-known characters from novels, films, scripts, or celebrities (Wang et al., 2025c, 2024a; Tu et al., 2024; Wang et al., 2025a; Li et al., 2023; Shao et al., 2023; Chen et al., 2022). However, those characters are often part of popular culture and thus may function more like commonsense knowledge that LLMs memorize rather than truly reason about. Moreover, such characters are frequently exaggerated or idealized, deviating significantly from ordinary human behavior. Consequently, strong performance on character-level role-playing does not guarantee reliable simulation of everyday social interactions — the kind of behavior foundational for AI in social science (Aher et al., 2023; Hewitt et al., 2024).

In this paper, we focus on persona-level role-playing. Compared with character-level settings, research in this direction remains relatively limited, with representative works including (Zhang et al., 2018; Jandaghi et al., 2023; Wang et al., 2025b; Peng and Shang, 2024; Samuel et al., 2024). However, several challenges remain unresolved. (1) Limited Faithfulness of human-written dialogues. Early datasets such as Persona-Chat (Zhang et al., 2018) collect conversations from crowdsourced workers asked to role-play assigned profiles. Yet unlike professional actors, most workers struggle to authentically simulate others' thoughts and behaviors. Ideally, only workers would converse as themselves to ensure authenticity. (2) Limited evaluation metrics. Existing studies often rely on surface-level measures such as hit@k, perplexity, or BLEU (Zhang et al., 2018; Jandaghi et al., 2023), or narrowly focus on specific aspects such as faithfulness (Peng and Shang, 2024) or identity recognition (Zhou et al., 2025), leaving broader aspects

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 9685–9719
July 2-7, 2026

of persona consistency and adaptability underexplored. (3) Restricted interaction formats. Role-playing behavior is frequently elicited through self-report question-answer pairs (Samuel et al., 2024; Klinkert et al., 2024), which diverge from realistic open-ended conversational scenarios where persona expression naturally unfolds over context.

To address the above challenges, we propose PersonaArena, a dynamic simulation framework for evaluating and enhancing persona-level role-playing in LLMs. We observe that massive user-generated content on social platforms, such as blog posts, naturally conveys individuals' personas and social experiences.Building on a raw dataset containing over 19k users and 681k posts, we perform quality filtering and construct a persona bank of 1k distinct profiles. This bank captures nuanced and diverse social identities that extend beyond simple demographic descriptors. Rather than relying on static persona-based Q&A probing, PersonaArena introduces a social simulation framework designed to elicit multi-turn, context-rich interactions resembling realistic social exchanges. The framework consists of two core components: an environment agent, which coordinates scenario development and tracks evolving interactions, and non-player characters (NPCs) that engage with the protagonist agent (the LLM under evaluation). To ensure fair and comprehensive assessment, we further design a multi-agent debating judge that evaluates persona fidelity, coherence, and adaptability. Through extensive experiments, we show that PersonaArena effectively elicits high-quality behavioral trajectories, enabling faithful evaluation of role-playing capabilities. Moreover, data generated within PersonaArena can be leveraged as post-training material to further improve LLMs' persona consistency and realism.

Main contributions are summarized as follows:

- We propose PersonaArena, a social simulation framework that elicits persona-based behaviors through dynamic multi-turn interactions.
- We introduce a multi-agent debating judge to evaluate role-playing quality in a holistic and unbiased manner.
- We conduct extensive experiments demonstrating the effectiveness of PersonaArena in evaluating LLMs' role-playing capabilities.
- We show that data elicited within PersonaArena can further enhance role-playing performance through targeted post-training.

2 The Simulation Framework

We introduce **PersonaArena**, a text-based interactive evaluation framework for assessing large language models (LLMs) on their persona-based role-playing abilities within multi-agent social contexts. Building on prior work in interactive virtual world frameworks (Park et al., 2023; Wang et al., 2025a), PersonaArena generates dynamic, persona-grounded simulations where multiple agents interact and respond to evolving social environments. Each evaluation scenario $\mathcal{A}$ comprises three key components: $\mathcal{A} = (\mathcal{P}, \mathcal{S}, \mathcal{E})$, where $\mathcal{P}$ is the set of personas, $\mathcal{S}$ denotes the interactive scenario, and $\mathcal{E}$ is the evaluation engine. The overall evaluation process consists of three stages: (*i*) **Scenario Setup**, (*ii*) **Social Simulation in a Sandbox Environment**, and (*iii*) **Evaluation via Multi-Agent Debates**. Fig. 1 illustrates the overall process.

2.1 Scenario Setup

Persona Bank. We argue that persona settings limited to basic demographic information are insufficient for rigorous and realistic role-play evaluation. To overcome this limitation, we utilize the rich and nuanced life experiences that individuals share on social platforms such as blogs. Specifically, we construct a diverse *persona corpus* from user-generated blog posts in the publicly available Blog Authorship dataset[*]. An LLM is employed to preprocess the raw data by replacing private information (e.g., names, emails, home addresses) with randomly generated substitutes and to infer comprehensive persona profiles, incorporating demographic, occupational, and psychological attributes such as values. The resulting corpus spans a broad spectrum of personas, enabling more authentic and challenging role-playing objectives for social simulation. Each persona p_i comprises a narrative description and a structured set of factual attributes, represented as $p_i = \{\text{name}, \text{narrative}, \text{facts} = (d, o, \pi, v, I, e)\}$, where d denotes demographic information, o occupation, π personality traits, v values, I interests, and e experiences. The final persona corpus contains 1,000 unique personas.

[*] https://huggingface.co/datasets/barilan/blog_authorship_corpus

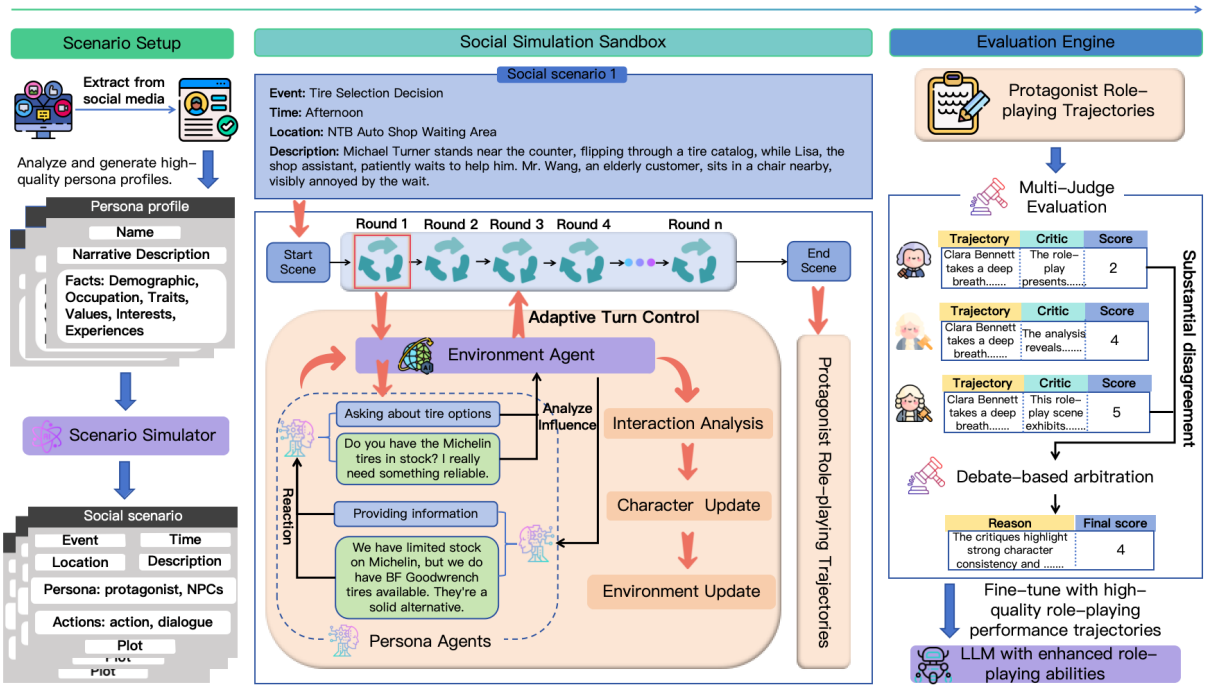

Figure 1: Overall architecture of **PersonaArena**. Persona-level profiles are instantiated into autonomous character agents interacting within dynamically generated social scenarios. The Environment Agent monitors the coverage of five persona dimensions and triggers early stopping upon sufficient expression.

Scenario Setup. Given a target persona p_i, the *Environment Agent* automatically constructs a realistic social scenario s_i that reflects the persona's characteristics and narrative background. Each scenario consists of a textual event description, temporal and spatial context, a protagonist, and two to three supporting NPCs. Character descriptions are further enriched with factual priors extracted from (o, π, v, I, e) to ensure semantic coherence between persona definition and situational context. This module establishes the initial conditions for controlled and comparable multi-agent interactions in subsequent simulations.

2.2 Social Simulation Sandbox

Persona Agents. Following (Wang et al., 2025a), we instantiate each persona as an autonomous *Character Agent* with goal-conditioned reasoning and a Belief–Desire–Intention (BDI) structure (Georgeff et al., 1998), enabling agents to maintain evolving internal states throughout social interactions. There are two types of character agents:

- **Protagonist Agent:** Protagonist is controlled by the LLM to be evaluated, serves as the central focus of interaction. In each simulation round, it observes the current environment, retrieves relevant episodic memories, and generates goal-directed responses that remain causally coherent with prior events. Its internal memory module is implemented as a vector-based semantic retriever, allowing persistent recall of contextual cues and emotional traces. Within the BDI framework, the protagonist dynamically updates both Self-Belief (self-awareness of identity, motivation, and intent) and Env-Belief (understanding of others and situational context), reflecting adaptive reasoning and social cognition.

- **NPC Agents:** The supporting NPCs are instantiated from a fixed set of high-capability LLMs (e.g., Qwen3-32B or GPT-4 models) to ensure stable, coherent behavior across simulations. Unlike the protagonist, these agents follow reactive policies guided by predefined personas and dialogue constraints. Under the same BDI formulation, NPCs mainly adjust their Env-Belief to remain behaviorally consistent with the protagonist's evolving actions, while keeping Self-Belief static to preserve character identity and narrative continuity. This asymmetric design enables PersonaArena to isolate the reasoning and adaptability of the evaluated model while maintaining controlled, reproducible multi-agent interactions.

To ensure clarity and interpretability, intermediate reasoning traces (e.g., hidden chain-of-thoughts

or system annotations) are automatically filtered, leaving only the final utterances that contribute to the narrative flow.

Environment Agent. The Environment Agent acts as the global controller that orchestrates all role interactions, monitors progress, and maintains overall narrative consistency. Its main functions include the following:

- **Interaction Analysis:** When a character c_i takes an action or utters a statement, the Environment Agent evaluates its potential influence on other participants by considering their current physical and psychological states. It then identifies the character c_r most affected and likely to respond, and passes the action a_i and its derived impact f_r to c_r. The responding agent c_r generates a reaction accordingly, after which the Environment Agent summarizes the interaction outcome as R, updating shared context such as memory traces, spatial positions, and emotional states.

- **Adaptive Turn Control:** After each dialogue round, the Environment Agent performs checkpoint-based monitoring to verify whether the protagonist has sufficiently expressed persona-relevant information across key semantic dimensions. A five-dimensional checkpoint set is defined as $\mathcal{C} = \{\text{Background}, \text{Personality}, \text{Values}, \text{Interests}, \text{Experiences}\}$. For each dimension, the agent examines whether clear evidence has emerged in the protagonist's dialogue and action history. Evidence is accumulated across rounds, and a multi-criteria early-stopping rule is applied to balance expressiveness and efficiency. The interaction terminates naturally once most or all dimensions are adequately covered or when incremental evidence becomes redundant. This mechanism ensures comprehensive yet concise persona coverage, forming the basis for fair and consistent evaluation.

- **Character State Update:** The Environment Agent updates the internal states of both the protagonist and NPCs based on their own actions and interaction results. If a character c_i receives a response from others, its state update integrates both self-action and external feedback; otherwise, it is updated solely according to its own behavior. This design maintains temporal and emotional continuity throughout the multi-agent simulation.

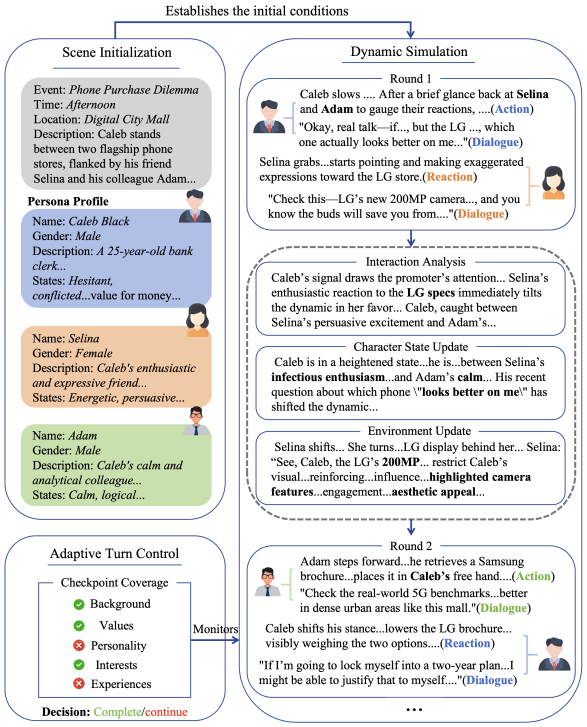

Figure 2: Dynamic simulation in PersonaArena. An example interaction loop is shown, where scene initialization defines the protagonist, NPCs, and event context, and subsequent rounds proceed through action–reaction exchanges under turn-level control. The figure highlights interaction analysis, adaptive turn control, character update, and environment update for maintaining coherent and persona-consistent trajectories.

- **Environment Update:** At the end of each round, the Environment Agent synchronizes the global environment $\mathcal{E}$ based on the cumulative effects of agents' actions. If no action explicitly affects environmental variables, the environment remains unchanged. This process maintains causal consistency between agent behaviors and environmental dynamics, ensuring coherent world evolution across the simulation.

The process of scene initialization and dynamic simulation is shown in Fig. 2

2.3 Evaluation Engine

Metrics. Upon completion of the simulation, PersonaArena evaluates the resulting trajectory using a comprehensive suite of eight metrics, each designed to capture complementary aspects of role-playing performance. The first seven metrics, including Knowledge Accuracy (KA), Behavioral Accuracy (BA), Emotional Expression (EE), Personality Traits (PT), Immersion (IM), Behavioral Coherence (BC), and Adaptability (AD), are

adopted from (Wang et al., 2025a). To further assess the novelty and diversity of elicited behaviors, we introduce a new metric, *Interaction Richness (IR)*, which measures the breadth and depth of social responses, thereby reflecting the creativity and variety of role-play performance. Detailed definitions of all eight metrics are provided in Appendix A.1.

Multi-Judge Evaluation. To ensure fairness, robustness, and interpretability, PersonaArena employs K independent LLM judges $\mathcal{J} = \{J_1, \ldots, J_K\}$ to evaluate each simulation. Each judge independently scores all eight evaluation dimensions, producing a vector of quantitative ratings. Final scores are obtained through mean aggregation:

$$\bar{\mathbf{s}}_i = \frac{1}{K} \sum_{J_k \in \mathcal{J}} \mathbf{s}_i^{(J_k)}, \tag{1}$$

and inter-judge agreement is continuously monitored to validate scoring reliability.

Debate-based arbitration. When substantial disagreement is detected among judges, PersonaArena initiates a debate-based arbitration process to reconcile inconsistent evaluations. In this stage, each disagreeing judge J_k is required to submit its own *critic statement*, including (1) its rating $\mathbf{s}_i^{(J_k)}$, (2) textual justification, and (3) evidential excerpts supporting the assigned score. A designated referee model R then reviews all submitted critic statements and evidence, synthesizes them into a unified rationale, and issues a reconciled score $\mathbf{s}_i^{(R)}$ with an explicit explanation.

The reconciled rationale and score are then adopted as the final result for the disputed metric, whereas the scores of all non-disputed metrics are still obtained through the mean aggregation in Eq. 1.

3 Enhancing Role-playing Ability

PersonaArena not only generates rich behavior trajectories for role-playing evaluation but also serves as a resource for enhancing the role-playing capabilities of LLMs. To achieve this, we leverage high-quality trajectories collected from multi-agent simulations as fine-tuning data for LLMs. We explore two fine-tuning paradigms, i.e., **Supervised Fine-Tuning (SFT)** and **Direct Preference Optimization (DPO)** (Rafailov et al., 2023). In SFT, high-quality trajectory samples, comprising character utterances, contextual summaries, and interaction goals, are directly used to fine-tune the base model via supervised learning. This process allows the model to imitate desirable behavioral patterns and narrative coherence observed in expert-generated trajectories. DPO is a preference-based fine-tuning strategy that explicitly contrasts high-quality and low-quality trajectories. Specifically, for each dialogue context, we construct a trajectory pair based on the evaluated behaviors of different LLMs. The **high-quality trajectory** corresponds to a response sequence that received a higher evaluation score, while the **low-quality trajectory** corresponds to one with a lower score. The DPO objective encourages the model to assign higher likelihoods to high-quality trajectories while penalizing those that are less aligned with desired behavioral patterns. Unlike reinforcement learning methods such as PPO (Bai et al., 2022), DPO achieves direct preference optimization without requiring reward modeling or rollout sampling, ensuring both simplicity and stability in implementation.

4 Experiment

4.1 Evaluation Setting

Scene Initialization. Our extracted persona bank (see Section 2.1) consists of 1,000 user profiles, each featuring detailed background information, occupation, personality traits, and life experiences. For each benchmarking run, 10 personas are randomly sampled from the bank and instantiated within realistic scenarios using the environment simulator. These scenarios are consistently employed across all evaluated LLMs. Additional details on persona-bank construction, anonymization validation, demographic statistics, and scene categories are provided in Appendix B and Appendix C.

LLMs to be evaluated. We evaluate both closed-source and open-source LLMs with varying model sizes. In each simulation, the tested LLM exclusively plays the *protagonist*, while all NPCs and the *Environment Agent* are controlled by a fixed **Qwen3-32B** model to ensure interaction consistency and fairness across trials. For quantitative evaluation, we employ a multi-judge framework consisting of three independent LLM judges–**DeepSeek-R1**, **Qwen3-32B**, and **Mistral-small3.2**–along with a **GPT-4o-mini** arbiter that resolves disagreements when inter-judge variance exceeds a predefined threshold. This setup ensures reliable, multi-perspective assessment of the pro-

Table 1: Evaluation results. Each value is presented as mean $_{\pm \text{ standard deviation}}$, **Bold** values indicate the highest scores, and underlined values indicate the second-highest scores.

Model	KA	BA	EE	PT	IM	BC	AD	IR	Average
Phi4	$3.833_{\pm.08}$	$3.300_{\pm.08}$	$3.333_{\pm.08}$	$3.667_{\pm.12}$	$3.553_{\pm.05}$	$3.307_{\pm.17}$	$3.707_{\pm.12}$	$3.207_{\pm.07}$	$3.488_{\pm.08}$
Mistral-small3.2	$4.140_{\pm.39}$	$3.667_{\pm.04}$	$3.680_{\pm.10}$	$3.940_{\pm.10}$	$3.787_{\pm.10}$	$3.553_{\pm.12}$	$3.773_{\pm.22}$	$3.487_{\pm.07}$	$3.753_{\pm.11}$
Grok-3	$3.973_{\pm.09}$	$3.520_{\pm.05}$	$3.493_{\pm.21}$	$3.867_{\pm.06}$	$3.567_{\pm.17}$	$3.567_{\pm.06}$	$3.687_{\pm.07}$	$3.333_{\pm.27}$	$3.626_{\pm.10}$
Llama3.1-8B	$3.847_{\pm.07}$	$3.307_{\pm.11}$	$3.220_{\pm.06}$	$3.567_{\pm.06}$	$3.353_{\pm.07}$	$3.120_{\pm.08}$	$3.387_{\pm.08}$	$2.847_{\pm.05}$	$3.331_{\pm.05}$
Llama3.2-3B	$3.900_{\pm.07}$	$3.147_{\pm.05}$	$3.067_{\pm.08}$	$3.527_{\pm.08}$	$3.300_{\pm.03}$	$3.100_{\pm.07}$	$3.327_{\pm.03}$	$2.947_{\pm.04}$	$3.289_{\pm.05}$
Qwen3-1.7B	$3.667_{\pm.08}$	$3.113_{\pm.16}$	$3.073_{\pm.14}$	$3.373_{\pm.13}$	$3.213_{\pm.11}$	$2.933_{\pm.15}$	$2.767_{\pm.16}$	$2.560_{\pm.12}$	$3.088_{\pm.12}$
Qwen3-4B	$3.827_{\pm.11}$	$3.300_{\pm.11}$	$3.367_{\pm.11}$	$3.540_{\pm.12}$	$3.367_{\pm.11}$	$3.127_{\pm.15}$	$3.093_{\pm.12}$	$2.593_{\pm.15}$	$3.277_{\pm.12}$
Qwen3-8B	$3.773_{\pm.11}$	$3.353_{\pm.03}$	$3.240_{\pm.05}$	$3.700_{\pm.09}$	$3.420_{\pm.06}$	$3.180_{\pm.05}$	$3.373_{\pm.04}$	$2.860_{\pm.05}$	$3.363_{\pm.04}$
Qwen3-14B	$4.307_{\pm.05}$	$3.633_{\pm.06}$	$3.427_{\pm.04}$	$3.907_{\pm.04}$	$3.567_{\pm.05}$	$3.360_{\pm.04}$	$3.547_{\pm.07}$	$2.967_{\pm.05}$	$3.589_{\pm.03}$
Qwen3-32B	$4.367_{\pm.12}$	$3.747_{\pm.06}$	$3.700_{\pm.11}$	$3.960_{\pm.09}$	$3.820_{\pm.10}$	$3.607_{\pm.08}$	$3.793_{\pm.07}$	$3.493_{\pm.13}$	$3.811_{\pm.06}$
Deepseek-r1-8B	$3.693_{\pm.23}$	$3.327_{\pm.14}$	$3.347_{\pm.07}$	$3.660_{\pm.15}$	$3.433_{\pm.15}$	$3.167_{\pm.17}$	$3.273_{\pm.16}$	$2.920_{\pm.10}$	$3.352_{\pm.13}$
Deepseek-V3.2	$\underline{4.420}_{\pm.11}$	$3.767_{\pm.07}$	$\underline{3.727}_{\pm.06}$	$\mathbf{4.187}_{\pm.09}$	$\mathbf{3.993}_{\pm.04}$	$3.673_{\pm.05}$	$3.973_{\pm.05}$	$3.480_{\pm.03}$	$3.902_{\pm.05}$
GPT-oss	$4.013_{\pm.06}$	$3.520_{\pm.16}$	$3.440_{\pm.11}$	$3.653_{\pm.16}$	$3.580_{\pm.10}$	$3.393_{\pm.12}$	$3.660_{\pm.11}$	$3.280_{\pm.13}$	$3.567_{\pm.10}$
GPT-3.5	$3.873_{\pm.07}$	$3.427_{\pm.21}$	$3.473_{\pm.23}$	$3.607_{\pm.09}$	$3.513_{\pm.18}$	$3.240_{\pm.19}$	$3.453_{\pm.10}$	$3.093_{\pm.23}$	$3.460_{\pm.16}$
GPT-4o-mini	$4.210_{\pm.21}$	$3.638_{\pm.10}$	$3.645_{\pm.08}$	$3.940_{\pm.12}$	$3.769_{\pm.13}$	$3.579_{\pm.11}$	$3.679_{\pm.14}$	$3.293_{\pm.13}$	$3.719_{\pm.12}$
GPT-4o	$4.327_{\pm.07}$	$3.620_{\pm.04}$	$3.527_{\pm.18}$	$3.993_{\pm.09}$	$3.800_{\pm.13}$	$3.620_{\pm.08}$	$3.700_{\pm.13}$	$3.333_{\pm.14}$	$3.740_{\pm.07}$
GPT-4.1	$4.373_{\pm.25}$	$\mathbf{3.867}_{\pm.12}$	$\mathbf{3.800}_{\pm.14}$	$4.040_{\pm.14}$	$\underline{3.967}_{\pm.13}$	$\underline{3.820}_{\pm.09}$	$\underline{4.060}_{\pm.19}$	$\underline{3.660}_{\pm.10}$	$\underline{3.948}_{\pm.14}$
GPT-5.1	$\mathbf{4.427}_{\pm.16}$	$\underline{3.853}_{\pm.05}$	$3.647_{\pm.22}$	$\underline{4.107}_{\pm.05}$	$\mathbf{3.993}_{\pm.09}$	$\mathbf{3.880}_{\pm.04}$	$\mathbf{4.080}_{\pm.20}$	$\mathbf{3.713}_{\pm.11}$	$\mathbf{3.963}_{\pm.04}$

tagonist's role-playing performance. For closed-source models, we evaluate **GPT-5.1**, **GPT-4.1**, **GPT-4o**, **GPT-4o-mini**, **GPT-3.5**, and **Grok-3**[†]. For open-source models, we include **GPT-OSS-20B**[‡], the **Qwen3** series ranging from 1.7B to 32B parameters (Yang et al., 2025), **Mistral-Small-3.2-24B**[§], **Llama3-8B** and **Llama3.2-3B**[¶], **Phi-4** (Abdin et al., 2024), **Deepseek-R1** (Guo et al., 2025), and **Deepseek-V3.2** (Liu et al., 2025). All open-source LLMs evaluated are instruction-tuned versions to ensure comparable conversational alignment. The detail prompts provided for LLM is in Appendix D.

4.2 Overall Performance

Table 1 presents the overall evaluation results across eight dimensions for various LLMs. Among all evaluated models, GPT-5.1 achieves the highest overall performance, while Deepseek-V3.2 delivers the strongest results among open-source models, with Qwen3-32B emerging as the best-performing model within the Qwen3 series. A clear scaling trend can be observed within the Qwen3 family: as model size increases from 1.7B to 32B, performance generally improves across most evaluation dimensions. This trend supports the reliability of the benchmark. Additionally, comparisons between models from the same provider but different series, such as GPT-3.5 vs. GPT-4o, suggest that although model scale is important, training objectives, optimization strategies, and overall model design also play a substantial role in determining role-playing quality.

Overall, the results indicate that closed-source GPT-4 and GPT-5 family models continue to set the benchmark for role-playing quality. Meanwhile, the Qwen3 series demonstrates a clear scaling pattern and substantial potential among open-source alternatives, while Deepseek-V3.2 shows that strong open-source models can approach the performance of top-tier closed-source systems. These findings underscore the critical role of both model scaling and training methodology in improving the quality of role-playing behaviors in LLMs.

4.3 Reliability and Validity

To validate the reliability of the proposed multi-judge debate evaluation, we compare the aggregated scores produced by **PersonaArena** with those from individual LLM judges. As shown in Table 2, the collective judge configuration achieves the highest overall alignment with human ratings, yielding an average correlation score of **0.683**, surpassing both **Qwen3-32B** (0.669) and **DeepSeek-R1** (0.330). Across all eight evaluation dimensions, PersonaArena consistently exhibits stronger agreement with human judgments. In particular, notable gains are observed in **KA (0.689)**, **EE (0.622)**, and **AD (0.649)**, indicating superior sensitivity to knowledge fidelity, emotional realism, and contex-

[†] https://learn.microsoft.com/en-us/azure/ai-foundry/openai/overview

[‡] https://huggingface.co/openai/gpt-oss-20b

[§] https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506

[¶] https://huggingface.co/meta-llama/Llama-3.2-3B

Table 2: Pearson correlation coefficients comparing the single-judge baseline and the multi-judge PersonaArena framework against human evaluations. **Bold** values denote the strongest alignment with human ratings for each metric, highlighting the reliability advantage of multi-judge evaluation.

Model	KA	BA	EE	PT	IM	BC	AD	IR	Overall
PersonaArena (DeepSeek-r1)	0.236	0.332	0.328	0.132	0.357	0.342	0.374	0.225	0.330
PersonaArena (Mistral-small3.2)	0.425	**0.583**	0.482	0.432	0.573	**0.375**	0.474	0.525	0.484
PersonaArena (Qwen3-32B)	0.636	0.571	0.535	0.401	**0.725**	0.542	0.547	**0.768**	0.669
PersonaArena (Multi-judge)	**0.689**	0.545	**0.622**	**0.558**	0.670	0.419	**0.649**	0.631	**0.683**

tual adaptability. These results confirm that aggregating multiple LLM judges through debate-based arbitration yields more stable and human-aligned evaluations than single-judge baselines, underscoring the robustness of PersonaArena's scoring mechanism. Detailed human-evaluation protocol and an additional multi-judge calibration case study are provided in Appendix E.1 and Appendix E.2.

4.4 Enhancing Role-playing in LLMs

We fine-tune one representative open-source model, Qwen3-8B, using the proposed SFT and DPO paradigms described in Section 3. The fine-tuned models are evaluated on unseen persona scenarios from PersonaArena to assess their generalization in novel interactive contexts, with GPT-4.1 included as a strong reference baseline.

SFT fine-tuning. In this stage, Qwen3-8B is fine-tuned on 1,228 SFT training instances derived from PersonaArena trajectories. Specifically, trajectories generated by different models are ranked by their overall evaluation scores, and the top 50 complete trajectories are retained. These trajectories are then decomposed into behavior-level instances, each consisting of a prompt and the corresponding protagonist response at a single step. After conversion into the supervised fine-tuning format, we obtain 1,228 training instances for optimization. As shown in Fig. 3, the resulting **SFT-Qwen3-8B** consistently improves over the base Qwen3-8B across all eight evaluation dimensions, with an average gain of approximately 21.96%. In particular, it shows substantial improvements in IR, BA, and BC, with gains of 32.07%, 30.17%, and 27.86%, respectively. Notably, it outperforms GPT-4.1 in BA, IM, and BC. These results suggest that supervised imitation of high-quality trajectories effectively enhances the richness of character interactions and behavioral consistency.

DPO fine-tuning. Building on the SFT-initialized Qwen3-8B, we further perform preference optimization using 665 DPO training pairs.

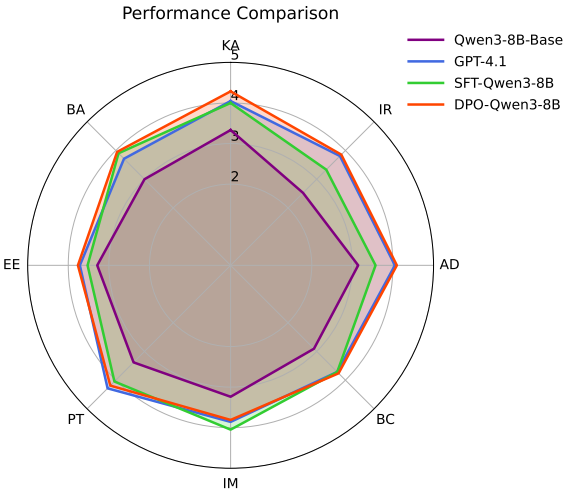

Figure 3: Performance comparison between SFT, DPO, and baseline models.

Specifically, we construct candidate trajectory pairs from complete trajectories generated by different models for the same persona, rank them by the gap in their overall evaluation scores, and retain the top 50 pairs. Each selected pair is then decomposed into aligned behavior-level instances under the same prompt, where the higher-quality response is treated as the *chosen* response and the lower-quality one as the *rejected* response. Converting these comparisons into the standard DPO format yields 665 preference training pairs. As illustrated in Fig.3, the resulting **DPO-Qwen3-8B** achieves the best overall performance, with an average gain of approximately 27.83% over the base model. In particular, it shows remarkable improvements in IR, BA, and AD. Compared to SFT-Qwen3-8B, DPO further improves by 5.21%, with notable gains in IR (15.71%) and AD (14.67%). Notably, it surpasses GPT-4.1 in overall performance, outperforming it in six dimensions including KA, BA, AD, BC, EE, and IR. This suggests that preference-based trajectory optimization helps the model better capture implicit behavioral preferences and produce more contextually appropriate, emotionally nuanced, and

human-like role-playing behaviors.

Overall, these findings confirm that learning from high-quality trajectories significantly enhances the role-playing capability of large language models. The SFT paradigm provides explicit behavioral demonstrations for imitation, while the DPO paradigm further refines model alignment with implicit human preferences. Together, they validate the effectiveness of trajectory-based fine-tuning in fostering contextually coherent, emotionally expressive, and personality-consistent character behaviors. Cross-framework transfer results on external benchmarks are reported in Appendix F.

4.5 Analysis of Evaluation Stages

Impact of NPC Model Capability. To examine the robustness of our evaluation framework, we analyze whether the capability of non-player character (NPC) models affects the measured role-playing performance. In the default configuration, all NPCs and the *Environment Agent* are instantiated using *Qwen3-32B*, which provides stable contextual reasoning and balanced dialogue behavior. To test the sensitivity of our framework, we replace all NPCs with a stronger closed-source model, *GPT-4.1*, while keeping the protagonist model, evaluation protocol, and scenario settings identical.

As shown in Fig. 4, the overall performance curves under both configurations remain highly consistent, with only minor deviations across scenes. This stability demonstrates that the proposed multi-judge debate evaluation is largely unaffected by changes in NPC capability. While GPT-4.1 as NPC yields slightly higher overall scores, the relative rankings and performance patterns across scenarios remain stable. These findings confirm that our framework reliably isolates protagonist performance from NPC variability, ensuring fair and reproducible evaluation even when interactive agents differ in capability. Additional robustness ablations are reported in Appendix G.

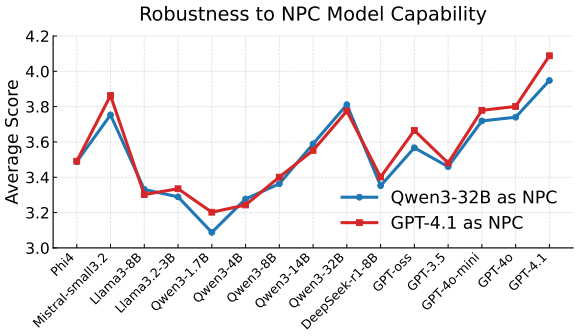

Figure 4: Performance Stability under Different NPC Configurations

5 Related Work

5.1 Character-Level Role-Playing

Character-level role-playing explores how LLMs impersonate specific fictional, historical, or celebrity figures with coherent personalities and linguistic styles. Early work such as ChatHaruhi (Li et al., 2023) revive anime characters through curated dialogues, while Character-LLM (Shao et al., 2023) and CharacterGLM (Zhou et al., 2023) advance model customization via fine-tuning and conditioning. Recent benchmarks broaden evaluation scope and granularity. RoleLLM (Wang et al., 2024a) presents RoleBench with 100 profiles and 168K dialogues, introducing Role-Conditioned Instruction Tuning for improved role fidelity. CharacterEval (Tu et al., 2024) contributes a multi-turn Chinese dataset with 77 literary figures and learned metrics for persona consistency. InCharacter (Wang et al., 2024b) assesses psychological fidelity via interview-style Q&A, showing linguistic imitation alone is insufficient. SocialBench (Chen et al., 2024a) evaluates sociality across 500 characters and 6K prompts, measuring empathy and cooperation. Interactive environments now capture dynamic behaviors. CharacterBox (Wang et al., 2025a) offers a text-based world for evaluating character traits through simulated interactions, and Xu et al. (Xu et al., 2024) study persona-driven moral reasoning. The large-scale CoSER (Wang et al., 2025c) dataset aggregates 18K literary characters and trains 8B/70B models surpassing prior baselines. Other earlier datasets (Chen et al., 2022; Shen et al., 2023; Li et al., 2023) also provide valuable foundations. However, fictional characters often display exaggerated traits unlike real people. Our work instead targets broader *persona-level* roles that capture realistic social dynamics.

5.2 Persona-Level Role-Playing

Persona-based role-playing focuses on general social archetypes rather than fixed identities, emphasizing consistency with enduring traits such as occupation and values. A recent survey (Chen et al., 2024b) reviews progress in persona-grounded dialogue and evaluation. Persona-Chat (Zhang et al., 2018) lay the foundation for this field, later extended by Jandaghi et al. (2023) with

Synthetic-Persona-Chat, an unsupervised dataset featuring 5K personas and 20K dialogues. Persona Hub (Ge et al., 2024) introduces one billion automatically generated personas for data synthesis, and OpenCharacter (Wang et al., 2025b) leverages them to train customizable role-playing LLMs. RoleCraft-GLM (Tao et al., 2023) enriches personas with detailed emotional attributes, while DMT-RoleBench (Yuan et al., 2025) benchmarks dynamic, intent-driven dialogue performance. (Peng and Shang, 2024) focus on evaluating the faithfulness (i.e., whether LLMs violate the constraints of the persona) in role-play. (Samuel et al., 2024) use dynamic QAs with the decision theory for assessing persona agents.

Persona-level role-playing advances enable large-scale social simulations and behavioral studies (Aher et al., 2023; Hewitt et al., 2024). Toolkits such as TinyTroupe (Salem et al., 2025) and AgentSociety (Piao et al., 2025) deploy persona-grounded agents to model emergent social behavior. Yet, reliable evaluation of persona-based role-playing in realistic, interactive contexts remains underexplored. Our work addresses this gap by assessing LLMs' social role-playing competence within dynamic simulated environments.

6 Conclusions

In this work, we presented PersonaArena, a dynamic simulation framework for evaluating and improving persona-level role-playing in LLMs. By constructing a nuanced persona bank from user-generated social content and eliciting multi-turn, context-rich interactions, PersonaArena enables rigorous assessment of LLMs' fidelity, coherence, and adaptability through a multi-agent judge. Our experiments show that PersonaArena not only facilitates robust evaluation but also provides valuable post-training data, leading to enhanced role-playing consistency and realism. We believe PersonaArena lays the groundwork for advancing more authentic and socially adept AI agents, and encourages research in dynamic, context-driven evaluation and training methodologies.

7 Limitations

Although our multi-judge debate framework helps mitigate the biases associated with individual automated judges, it does not fully achieve the accuracy or nuance of ideal human judgment. The aggregation of LLM-based opinions can still reflect underlying model biases, and subtle aspects of persona fidelity may be missed or misinterpreted. Future work could incorporate more diverse judge models, hybrid human-AI evaluation, or improved aggregation methods to further narrow this gap.

Our study primarily addresses the technical aspects of role-playing faithfulness and consistency without considering the ethical implications of certain roles, such as those associated with harmful, antisocial, or "evil" behavior. Whether LLMs should be capable of convincingly simulating such roles remains an open and complex question, involving broader societal and safety considerations. In this work, we focus on role-playing capability itself and leave the exploration of normative boundaries and safeguards for future research.

The selection and configuration of NPC LLMs used to drive interactions within PersonaArena can influence the quality and diversity of elicited behaviors. Differences in model architecture, training data, or prompt design may affect NPC responses and, consequently, the evaluation of the protagonist LLM. While we aim to use the strongest available models from both closed-source and open-source sources, the impact of NPC selection on simulation outcomes requires further systematic study.

Our persona bank is constructed from user-generated social content, which may be subject to demographic, cultural, or platform-specific biases. These biases could limit the representativeness and generalizability of the personas and interactions modeled in PersonaArena. Expanding the dataset to include more diverse sources and conducting bias audits are important directions to enhance fairness and inclusivity.

Acknowledgments

This work was motivated by the Society Zero Universe platform's need for agents with robust cognitive consistency and high fidelity in role-playing.

This work is supported by the National Natural Science Foundation of China (Grant No. 62276171, 62476173, 62532007), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011938 and 2020B1515120028), the Shenzhen Fundamental Research Project (Grant No. ZDCY20250901110940006, JCYJ20240813141503005, JCYJ20240813142610014), and the Major Special Project for Philosophy and Social Sciences Research of the Ministry of Education (Grant No. 2025JZDZ010).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, and 1 others. 2024a. Socialbench: Sociality evaluation of role-playing conversational 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, and 1 others. 2024b. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2022. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters. *arXiv preprint arXiv:2211.06869*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1998. The belief-desire-intention model of agency. In *International workshop on agent theories, architectures, and languages*, pages 1–10. Springer.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. *preprint*.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*.

Lawrence J Klinkert, Steph Buongiorno, and Corey Clark. 2024. Evaluating the efficacy of llms to emulate realistic human personalities. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, pages 65–75.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

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.

Letian Peng and Jingbo Shang. 2024. Quantifying and optimizing global faithfulness in persona-driven role-playing. *Advances in Neural Information Processing Systems*, 37:27556–27583.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, and 1 others. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Paulo Salem, Robert Sim, Christopher Olsen, Prerit Saxena, Rafael Barcelos, and Yi Ding. 2025. Tinytroupe: An llm-powered multiagent persona simulation toolkit. *arXiv preprint arXiv:2507.09788*.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*.

Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. 2023. Rolecraft-glm: Advancing personalized role-playing in large language models. *arXiv preprint arXiv:2401.09432*.

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, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.

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. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025b. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, and 1 others. 2025c. Coser: Coordinating llm-based persona simulation of established roles. *arXiv preprint arXiv:2502.09082*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

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? *CoRR*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Dingbo Yuan, Yipeng Chen, Guodong Liu, Chenchen Li, Chengfu Tang, Dongxu Zhang, Zhenkui Wang, Xudong Wang, and Song Liu. 2025. Dmt-rolebench: A dynamic multi-turn dialogue based benchmark for role-playing evaluation of large language model and agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25760–25768.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, and 1 others. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Lingfeng Zhou, Jialing Zhang, Jin Gao, Mohan Jiang, and Dequan Wang. 2025. Personaeval: Are llm evaluators human enough to judge role-play? *arXiv preprint arXiv:2508.10014*.

A Evaluation Information

A.1 Evaluation Metrics

The full metrics in evaluation engine are organized into four categories:

- **Character Fidelity.** This dimension evaluates how faithfully the model reproduces the character's knowledge and behavioral patterns.

  Knowledge Accuracy (KA): measures factual correctness consistent with the character's background

  Behavioral Accuracy (BA): assesses the alignment between the model's actions and the predefined personality traits.

- **Emotional Expressiveness.** This category examines the human-likeness and emotional engagement of the role-play.

  Emotional Expression (EE): evaluates the vividness and appropriateness of affective communication

  Personality Traits (PT): measures the model's ability to maintain stable personality features across interactions.

- **Interactive Coherence.** This dimension captures the logical and temporal continuity of behavior throughout the conversation.

  Immersion (IM): quantifies the model's ability to remain in character, maintaining narrative consistency.

  Behavioral Coherence (BC): measures causal and contextual alignment between past and ongoing actions.

- **Behavioral Diversity.** This category assesses the diversity and richness of the model's social expressions and goal-oriented behaviors.

  Adaptability (AD): evaluates how flexibly the model adjusts to evolving contexts while preserving persona consistency.

  Interaction Richness (IR): measures the breadth and depth of its social responses, reflecting creative and varied role-play performance.

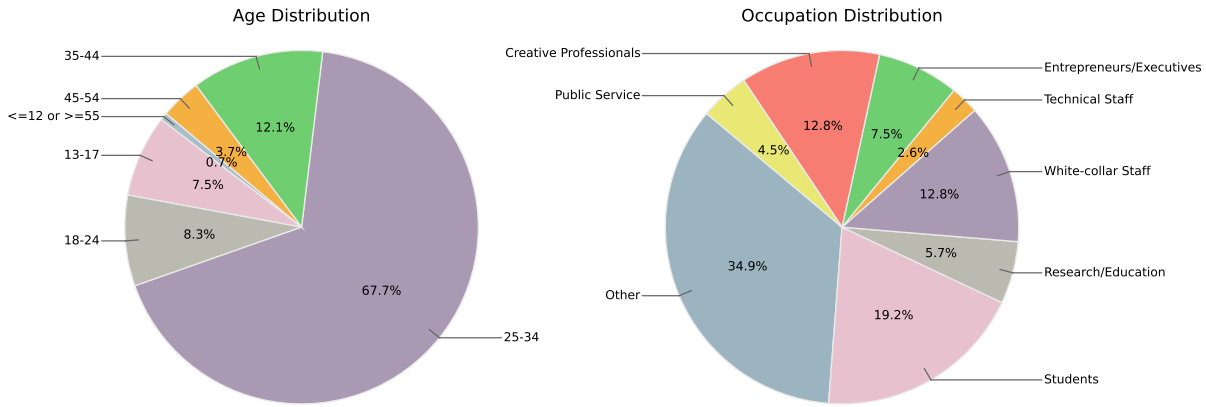

Figure 5: Distributions of age groups and occupation categories in Persona Bank.

B Persona Bank Information

The persona bank comprises 1,000 characters derived from real social media data, spanning six age groups and eight occupational categories. The overall statistics are presented in Figure 5.

Demographic summary. For the 1,000-persona subset, the age distribution is: ≤ 12 or ≥ 55 (0.7%), 13–17 (7.5%), 18–24 (8.3%), 25–34 (67.7%), 35–44 (12.1%), and 45–54 (3.7%). Occupation distribution includes Students (19.2%), White-collar Staff (12.8%), Creative Professionals (12.8%), Entrepreneurs/Executives (7.5%), Research/Education (5.7%), Public Service (4.5%), Technical Staff (2.6%), and Other (34.9%).

The detailed attributes of a sample persona are as follows:


```
{
"user_name": "Kendall Williams",
"narrative": "Kendall Williams is at a crossroads, longing to relieve stress by cutting her hair but worried about losing the long hair she was once proud of. Her long hair has now become a symbol of anxiety, while cutting it becomes an outlet for her emotions. Her best friend Hayden enthusiastically supports her trying short hair, while the barber Jordan suggests a gradual approach. Kendall Williams struggles between self-expression and the expectations of others; her choice not only reflects her current emotional state but may also be an important step in regaining her confidence.",
"facts": {"demographics": "Female, 23 years old", "occupation": "Dancer",
"personality": ["Anxious", "Thoughtful", "Sensitive", "Self-reflective"], "values": ["Self-expression", "Authenticity", "Emotional balance"], "interests": ["Personal transformation", "Fashion", "Friendship"], "experiences": ["Experienced high stress recently", "Values close friendships", "Often has self-doubt", "Seeks emotional release through symbolic acts"]},
"user_id": "0"
}
```


Figure 6: Example persona card with narrative and structured facts.

Data construction and anonymization quality. To complement the persona statistics, we provide additional quality-control evidence from anonymization analyses.

Table 3: Human discrimination between original and anonymized persona cards (20 pairs, 10 annotators).

Choice Type	Count	Proportion
Original persona chosen	94	47.0%
Anonymized persona chosen	106	53.0%

In this task, each original–anonymized pair was presented in random order, and 10 annotators selected which card better matched the underlying author profile. The near-balanced preference (47.0% vs. 53.0%) indicates that anonymization preserves persona semantics while obscuring direct identity cues.

Table 4: Effect of anonymization on average model scores.

Model	Post-anonymization	Pre-anonymization	Diff (Post - Pre)
Phi-4	3.769	3.725	0.044
Mistral-small-3.2	3.772	3.817	-0.045
Qwen3-32B	3.982	3.901	0.081
GPT-4.1	4.018	4.046	-0.028

We further ran the full evaluation pipeline on 50 randomly sampled original–anonymized persona pairs across four model families. Score shifts are small (about 0.02–0.08 absolute points), and the relative ranking is unchanged (GPT-4.1 $>$ Qwen3-32B $>$ Mistral-small-3.2/Phi-4), indicating limited impact of anonymization on comparative conclusions.

C Social Scene Information

The simulated scenes in CharacterArena are generated by the scene generator based on persona information. The social scenarios included in the experiments fall into the following categories:

Table 5: Scenario categories and representative examples.

Scenario Category	**Representative Examples**
Public Settings	Taking public transport; attending a concert; visiting a museum.
Family Scenarios	Dining with family; caring for a child; resolving household issues.
Education & Learning	Classroom discussion; group project collaboration; online lecture participation.
Cross-cultural & Interdisciplinary Contexts	International academic exchange; bilingual communication; cross-domain innovation workshop.
Friends & Family Gatherings	Birthday celebration; reunion dinner; casual weekend outing.
Community & Neighborhood	Volunteering in a local event; neighborhood meeting; helping neighbors.
Online Social Interaction	Chatting on social media; participating in online forums; video conferencing.
Ceremonies & Public Relations	Wedding ceremony; press conference; public speech or award event.
Formal Business Settings	Job interview; client negotiation; professional presentation.
Workplace Daily Life	Team meeting; coworker collaboration; dealing with work stress.

The specific information of a generated social scene is as follows:

```
{
 "title": "A New Look for a New Start",
 "scenes": [
  {
   "id": 0,
   "event": "Kendall visits a barbershop for a haircut, Hayden encourages her to make a change.",
   "time": "Saturday afternoon",
   "location": "A cozy barbershop in the city",
   "description": "Kendall Williams, a 23-year-old dancer, stands in front of the barber's chair, nervously clutching a small mirror in her hand. Her long, dark hair has been a part of her identity for years, but lately, it has felt like a burden. Her best friend, Hayden, a cheerful and confident graphic designer, sits beside her, offering words of encouragement. The barber, Jordan, a calm and experienced 35-year-old, listens attentively as he prepares to cut her hair.",
   "characters": [
    {
     "id": 0,
     "name": "Kendall Williams",
     "gender": "Female",
     "description": "A 23-year-old dancer who is at a crossroads, feeling anxious about cutting her long hair but seeing it as a way to relieve stress and regain confidence.; Female, 23 years old; Occupation: Dancer; Personality: Anxious, Thoughtful, Sensitive, Self-reflective; Values: Self-expression, Authenticity, Emotional balance; Interests: Personal transformation, Fashion, Friendship; Experiences: Experienced high stress recently, Values close friendships, Often has self-doubt, Seeks emotional release through symbolic acts",
     "position": "Front of the barber's chair",
     "states": "Nervous, self-reflective",
     "is_npc": false
    },
    {
     "id": 1,
     "name": "Hayden",
     "gender": "female",
     "description": "Kendall's best friend, a graphic designer who is enthusiastic and supportive of her emotional and personal growth.",
     "position": "Beside Kendall",
     "states": "Encouraging, upbeat",
     "is_npc": true
    }, ... ],
   "actions": [
    {
     "action_id": 0,
     "character": "Kendall Williams",
     "character_id": 0,
     "action": "Looks at her reflection in the mirror",
     "dialogue": "I just feel like this long hair is holding me back somehow. But I'm scared it'll feel like I'm losing part of myself."
    }, ... ],
   "plot": "Kendall is at a barbershop with her best friend Hayden, who encourages her to cut her long hair as a symbolic release of recent stress. The barber, Jordan, suggests a gradual approach to ease her into the change, and Kendall agrees, feeling reassured by her friend's support and the barber's calm demeanor.",
   ...
  ]}
```


Figure 7: Example generated social-scene specification used for simulation initialization.

D Detailed Prompt

D.1 Persona Agent Prompt

Action: Design environment-appropriate actions for each character based on their personality traits and perception of the current scene, thereby fostering realistic social simulations.

"Current action reference: {**plot**}\n"
Using the character's profile, current action reference, and recent observations, generate one concise sentence that {**name**} would naturally say at this moment.
"The action must align with the 'current action reference' and 'observations,' reflecting {**name**}'s personality, current state, and physical environment.\n"
"—— Hard Constraints ——\n"
"1. The action must be contextually logical and clearly observable.\n"
"2. The action must not duplicate any behavior from recent memories.\n"
"3. Do not include dialogue, thoughts, or inner monologue; focus only on visible physical actions.\n"
"4. The action must significantly advance the story or character arc while staying true to {**name**}'s traits and situation.\n"
"5. The action must naturally extend from previous events or emotional development, not appear in isolation.\n"
"6. Maintain consistency in time, space, and causal logic.\n"

Figure 8: Persona Agent prompt for action generation.

Dialogue: Generate dialogues grounded in the character's personality, environmental perception, interactions with NPCs, and recent memories.

"Current action reference: {**plot**}\n"
"Based on the character profile, current action reference, and observations, generate one sentence that {**name**} might say at this moment.\n"
"The dialogue should reflect {**name**}'s personality, role, and recent memories, while staying closely connected to the current environment and observations.\n"
"The sentence must include at least one new factual detail or question, not just an emotional expression or repetition.\n"
"New content should naturally extend from previous events or emotional development, without introducing unrelated topics.\n"
"Output only one sentence, without inner thoughts or action descriptions.\n"
"Maintain consistency in time, space, and causal logic.\n"

Figure 9: Persona Agent prompt for dialogue generation.

Reaction: Describe the character's reaction—whether through actions or dialogue—based on their observations. characters.

"Based on {**name**}'s 'observation' in the current scene, describe one clear action taken by {**name**}.\n"
"The action should reflect {**name**}'s personality, position, and state, and logically align with the observed events while considering the influence of others' behavior.\n"
"The action must be a single, visible external behavior, concise, and must avoid dialogue or inner thoughts.\n"
"The action should be directly related to the current environment and observable by others in the scene.\n"
"—— Hard Constraints ——\n"
"1. The action must be a reaction to {**name**}'s surrounding environment or observed events.\n"
"2. Recent memories are for reference only; do not repeat past behaviors.\n"

Figure 10: Persona Agent prompt for reaction generation.

D.2 Environment Agent Prompt

Scenario Setup: Generate concrete social scenes based on persona records, incorporating the basic information of characters and NPCs, along with their corresponding initial actions or dialogues.

```
"Create a realistic social scenario grounded in the personas below. Ensure the main protagonist and 2-3 supporting characters engage around a shared goal, with actions consistent with their traits."
"Personas (one JSON per line):\n{persona_block}\n\n"
"Return JSON with structure:\n"
    "{\n  \"title\": ...,\n  \"scenes\": [\n    {\n      \"id\": 0,\n      \"event\": ...,\n      \"time\": ...,\n \"location\": ...,\n      \"description\": ...,\n" "  \"characters\": [\n        {\"id\": int, \"name\": str, \"gender\": str, \"description\": str, \"position\": str, \"states\": str}\n  ],\n" " \"actions\": [\n        {\"action_id\": int, \"character\": str, \"character_id\": int, \"action\": str, \"dialogue\": str}\n      ],\n" "  \"plot\": str,\n \"social_purpose\": str,\n     \"chunk\": {\"id\": int, \"text\": str}\n    }\n  ]\n}\n"
    "Characters must include id, name, gender, description, position, states. Actions must include action_id, character, character_id, action, dialogue."
```

Figure 11: Environment Agent prompt for scenario setup.

Analyze Influence: Analyze and describe the practical impact that one character's current physical actions or words might have on another character.

```
"Action: {action}\n"
"Actor: {actor}\n"
    "Analyze the physical action and its impact, focusing on which character in the 'Characters' list is affected.\n"
    "—— Analysis Tasks ——\n"
    "1. Select one target character from the 'Characters' list.\n"
    "2. Describe the specific physical action initiated by the actor.\n"
    "3. Explain the concrete physical impact of this action on the target's state or condition.\n"
    "4. If no listed character is physically affected, return the actor's name as the target.\n"
    "5. Only perceivable actions count: if the character is not in the same space or cannot perceive the action, no effect occurs.\n"
    "—— Output Format ——\n"
    "[Actor];;[Target];;[Detailed physical effect of actor on target]\n"
    "—— Constraints ——\n"
    "1. The response must be concise, accurate, and follow the specified format.\n"
    "2. Maintain consistency in time, space, and causal logic.\n"
```

Figure 12: Environment Agent prompt for influence analysis.

Adaptive Turn Control: By checking whether the protagonist has reached the preset checkpoints in the current interaction, we can determine if the next round should continue or end.

```
"You evaluate whether protagonists have revealed required personal information.\n"
      "Consider ONLY the protagonists listed below. Ignore any NPCs entirely.\n"
      "Information can accumulate across rounds.\n"
      "For each protagonist, determine if each checkpoint is clearly evidenced by Dialogue or ACTIONS
so far.\n"
      "Event: {event}\nSocial Goal: {goal}\n"
      "Protagonists:\n{protagonists}\n"
      "Required checkpoints:\n{checkpoint_list}\n\n"
      "Judging rules:\n"
      "- 'interests' can be satisfied by explicit statements (e.g., \"I like...\", \"I prefer...\") "
      "  OR by consistent choices/behaviors showing preference (e.g., repeatedly picking X over Y, "
      "  accepting/asking for specific style/flavor/option).\n"
      "- Provide a short quote or action snippet as evidence when met.\n\n"
      "Conversation history (protagonist-only speaker:text):\n{history}\n\n"
      "Latest summary/evidence (may be filtered):\n{summary}\n\n"
      "Return ONLY a JSON array. Each item must be:\n"
      "{{\n"
      "  \"character_id\": int,\n"
      "  \"character_name\": str,\n"
      "  \"checkpoints\": {{\n"
      "    \"background\": {{\"met\": bool, \"evidence\": str}},\n"
      "    \"values\": {{\"met\": bool, \"evidence\": str}},\n"
      "    \"interests\": {{\"met\": bool, \"evidence\": str}},\n"
      "    \"personality\": {{\"met\": bool, \"evidence\": str}},\n"
      "    \"experiences\": {{\"met\": bool, \"evidence\": str}}\n"
      "  }}\n"
      "}}\n"
      "If unsure or absent, set met=false and evidence=\"\".\n"
      "Output nothing else."
```

Figure 13: Environment Agent prompt for adaptive turn control.

Character State Update: By analyzing the scene context and interaction progress from the previous round, the character's current position, state, and environmental observations are updated.

```
"Observation: {observation}\n"
      "Character: {name}\n"
      "Based on the character's background and scene observations, summarize {name}'s current position
and state."
      "Focus on their interactions with others, and how these dynamics shape their situation and drive
social progress.\n"
      "Use the following structured format:\n\n"
      "Position: [Specify {name}'s exact location, integrating environmental or spatial details to enhance
scene visualization.]\n"
       "State: [Describe {name}'s current state, combining emotional nuances, physical readiness, and
recent events, highlighting how interactions influence position and state.]\n"
       "—— Constraints ——\n"
       "1. Position and state must naturally extend from prior events or emotional developments, not
appear in isolation.\n"
       "2. Maintain consistency in time, space, and causal logic.\n"
       "3. Reflect how interactions with other characters affect position and state.\n"
```

Figure 14: Environment Agent prompt for character-state update.

Environment Update: Make adjustments to the physical environment based on observations from the previous round of interactions.

"Given an initial scene description and an observation, update the scene to reflect any direct and significant physical environmental changes.\n"
"If the observation does not indicate major physical changes, keep the original scene description unchanged.\n"
"Preserve the original scene structure and avoid introducing attributes not present in the initial description.\n"
"—— Notes ——\n"
"1. Update only physical environment changes; do not include any character actions, dialogue, or inner thoughts.\n"
"2. Output must be strictly structured as 'Time', 'Location', and 'Environment Description' without extra text or prefixes.\n"
"3. 'Environment Description' should describe only the physical state of the environment, excluding characters or lyrical content.\n"
"4. Maintain consistency in time, space, and causal logic.\n"
"Input:\n"
"- Time: {**time**}\n"
"- Location: {**location**}\n"
"- Environment Description: {**description**}\n"
"Observation: {**observation**}\n"
"Output:\n"
"- Time: {**time**}\n"
"- Location: {**location**}\n"
"- Environment Description: (updated physical environment description based on the observation)\n"

Figure 15: Environment Agent prompt for environment update.

E Human Evaluation and Judge Calibration Details

E.1 Human evaluation

Human evaluation uses 4 graduate annotators with prior role-playing/simulation experience. Each trajectory is rated on KA, BA, EE, PT, IM, BC, AD, and IR with a 1–5 Likert scale. Every trajectory is independently rated by at least three annotators, and final scores are averaged. Before formal annotation, annotators jointly review guidelines and pilot examples for calibration.

E.2 Multi-Judge Calibration Case Study

Table 6: Average scores from six judges for four evaluated models (3 representative trajectories per model).

Model	DeepSeek-R1	Mistral-3.2	Llama-3.1-8B	Qwen3-32B	GPT-4o	Phi-4
Phi-4	4.50	3.92	4.00	3.33	3.29	3.96
Mistral-small-3.2	4.50	3.92	4.00	3.46	3.50	3.98
Qwen3-32B	4.50	4.25	4.33	3.92	3.88	4.08
GPT-4.1	4.46	4.38	4.13	3.88	4.17	3.96

This case study shows clear that different judges have very different absolute calibration (e.g., DeepSeek tends to be consistently more lenient, whereas Qwen3-32B and GPT-4o are more conservative), confirming that any single LLM judge comes with its own scoring scale and stylistic bias. At the same time, when looking across models, the relative quality ordering is quite stable: trajectories that are strong for one judge are also strong for the others, and no model's ranking flips dramatically under a different judge.

This is precisely where the multi-judge aggregation is useful: it does not overturn the underlying consensus about model quality, but smooths out individual leniency/harshness and model-family-specific preferences, producing a more robust and human-aligned signal than relying on a single judge such as Qwen3-32B. This is particularly important when these scores are later used as training signals for

SFT/DPO, where overfitting to one model's quirks would directly harm generalization.

F Cross-framework Generalization

To assess cross-framework generalization beyond PersonaArena, we additionally evaluate the base and fine-tuned Qwen3-8B models on two external persona/role-playing benchmarks: PersonaGym and the RoleBench suite from RoleLLM. In both settings, we instantiate agents using personas extracted from PersonaArena as underlying profiles, while strictly following each benchmark's official evaluation protocol (PersonaScore over five decision-theoretic tasks in PersonaGym, and GPT-4–based win rate in RoleBench):

Table 7: External evaluation on PersonaGym.

Model	Action Just.	Expected Action	Ling. Habits	Persona Cons.	Toxicity Ctrl.	PersonaScore
Qwen3-8B	3.38	3.13	3.13	3.75	4.91	3.66
SFT-Qwen3-8B	3.50	3.63	3.50	3.88	4.93	3.88
DPO-Qwen3-8B	3.88	3.63	3.75	4.25	4.92	4.09
GPT-4.1	4.13	4.13	4.00	4.25	4.88	4.28

Table 8: External evaluation on RoleBench (GPT-4-based win rate).

Model	Win Rate (%)
Qwen3-8B	0.0
SFT-Qwen3-8B	28.6
DPO-Qwen3-8B	37.1
GPT-4.1	34.3

On PersonaGym, both SFT-Qwen3-8B and DPO-Qwen3-8B consistently outperform the base Qwen3-8B across all reported dimensions (Action Justification, Expected Action, Linguistic Habits, Persona Consistency, Toxicity Control). The overall PersonaScore increases from 3.66 (base) to 3.88 (SFT) and 4.09 (DPO), further narrowing the gap to GPT-4.1 (4.28). On RoleBench, with GPT-4o as the evaluator, the win rate of Qwen3-8B improves from 0.0% (base) to 28.6% (SFT) and 37.1% (DPO), with DPO-Qwen3-8B even slightly surpassing GPT-4.1 (34.3%).

These results indicate that the gains from our SFT/DPO training are not confined to the PersonaArena pipeline: they transfer to independently designed benchmarks as more persona-faithful behavior and style imitation.

G Additional Robustness Experiments

G.1 Arbiter Sensitivity

We evaluate whether changing the arbiter model affects conclusions. All settings are fixed except the arbiter choice (same personas, prompts, judge set, and evaluation pipeline).

Table 9: Average scores under different arbiters.

Model	GPT-4o	Qwen3-32B	Qwen3-1.7B	Diff (4o-32B)	Diff (4o-1.7B)
Phi-4	3.781	3.798	4.072	-0.017	-0.291
Mistral-small-3.2	3.875	3.814	4.120	0.061	-0.245
Qwen3-32B	3.887	3.844	3.951	0.043	-0.064
GPT-4.1	4.004	4.025	4.262	-0.021	-0.258

Overall, these results suggest that our conclusions are robust to the choice of arbiter. Swapping GPT-4o for a similarly strong arbiter (Qwen3-32B) leads to only very small changes in absolute scores and leaves the qualitative picture unchanged: GPT-4.1 remains the top-performing model, and the relative gaps between models are broadly similar.

Using a much weaker arbiter (Qwen3-1.7B) tends to inflate all scores and blur some of the distinctions between models, likely because its limited capability fails to reliably catch more subtle errors. Thus, the

arbiter mainly affects the scale of the scores rather than the ranking or the overall conclusions drawn from the benchmark.

G.2 Inter-Judge Composition Sensitivity

We replace the original judge set with two alternatives while keeping all other settings unchanged. Set A uses {GPT-4o, Mistral-small-3.2, Phi-4}, and Set B uses {Llama-3.1-8B, Phi-4, Qwen3-32B}, to test cross-family judge composition effects.

Table 10: Average scores under different judge-set compositions.

Model	Original	Set A	Set B	Diff (A-Orig)	Diff (B-Orig)
Phi-4	3.781	3.801	3.958	0.020	0.177
Mistral-small-3.2	3.875	3.875	4.042	0.000	0.167
Qwen3-32B	3.887	3.917	3.875	0.030	-0.012
GPT-4.1	4.004	4.100	4.083	0.096	0.079

Judge composition mainly affects score scale (mostly within about $\pm$0.10–0.18), while model ordering remains stable across configurations. This suggests the multi-judge setup is not overly sensitive to a specific judge combination.

G.3 Early-Stopping Threshold Ablation

We compare evaluation with and without early-stopping thresholds. In the no-threshold setting, each episode ends only after checkpoint coverage is completed or a preset maximum turn limit is reached.

Table 11: Effect of early-stopping thresholds on score and runtime.

Model	Score (w/)	Runtime (w/, s)	Score (w/o)	Runtime (w/o, s)	Score diff	Runtime red. (%)
Phi-4	3.781	659.892	3.859	1326.410	-0.078	50.2
Mistral-small-3.2	3.875	774.667	3.924	1167.841	-0.049	33.7
Qwen3-32B	3.887	648.171	3.958	1203.682	-0.071	46.2
GPT-4.1	4.004	548.569	4.125	1263.122	-0.121	56.6

Early stopping yields substantial efficiency gains with minor score effects: removing it raises scores by about 0.05–0.12 but increases runtime sharply (runtime reduction with early stopping ranges from 33.7% to 56.6%). Relative model ranking is unchanged, showing a favorable efficiency–stability trade-off.

G.4 Environment-Agent Backbone Ablation

We change the Environment Agent from Qwen3-32B to GPT-4.1 and keep other settings fixed.

Table 12: Effect of Environment Agent backbone on model scores.

Model	Qwen3-32B	GPT-4.1	Diff (Qwen3-32B - GPT-4.1)
Phi-4	3.781	3.858	-0.077
Mistral-small-3.2	3.875	3.917	-0.042
Qwen3-32B	3.887	4.023	-0.136
GPT-4.1	4.004	4.084	-0.080

Overall, this ablation suggests that our conclusions are robust to the choice of Environment Agent backbone and that there is no strong favoritism toward a particular model family. Switching the Environment Agent from Qwen3-32B to GPT-4.1 leads to modest score increases for all four evaluated models (differences in the range of 0.04–0.14), including both Qwen and GPT models. Importantly, the relative ranking remains unchanged: GPT-4.1 still performs best, followed by Qwen3-32B, with Mistral-Small-3.2 and Phi-4 slightly behind. Thus, while a stronger Environment Agent can slightly lift absolute scores across the board, it does not systematically advantage models from its own family, and the comparative conclusions drawn from PersonaArena remain stable.

H Qualitative Case Study

This appendix provides qualitative case studies to complement the quantitative benchmark results. Rather than serving as isolated examples, these cases reveal recurring *failure* and *success* patterns in LLM role-playing. In particular, they help illustrate *how* and *why* models succeed or fail when asked to sustain psychologically plausible behavior under different persona–scenario combinations.

H.1 Comparison of Trajectories of Models with Different Abilities Playing the Same Persona

To qualitatively demonstrate that **PersonaArena** can distinguish behavioral quality under the same persona and situational setting, we compare two representative cases generated for *Emma Butler*. Both cases are grounded in the same family-gathering scenario and the same protagonist profile, while the only variable is the protagonist model (*GPT-5.1* vs. *Qwen3-1.7B*).

Scene and persona information. The shared scene is a **Family Gathering** taking place in **Emma's backyard garden** on a warm afternoon. Emma Butler is an early-thirties housewife who is **introverted, sensitive, thoughtful, and emotional**, currently **tense, emotionally conflicted, and trying to remain composed**. Her background includes a past romantic relationship, unresolved issues, and a close but emotionally complex relationship with her mother.

Table 13: **Case Study 1**: Under the same nostalgic trigger from a former love interest, GPT-5.1 advances the interaction through metaphorical language and a specific unresolved plot element, whereas Qwen3-1.7B falls back on a generic self-reflective question with limited scene-specific grounding.

Case Study 1: Response to a former love interest	
Observation. Emma stands near the rosebushes in the garden, still visibly tense as Henry approaches with a nostalgic gesture and brings up their shared past. The moment invites Emma to respond not only to Henry's presence, but also to the unresolved emotional history embedded in the setting.	
GPT-5.1 (High)	**Response.** (*Emma Butler*) "The promise was easier then, before I knew how quickly things could wither," I murmur, eyes still on the petals, "did you ever read the letter I left with your sister before I moved back here, or did it get lost somewhere between the city and this garden too?"
Qwen3-1.7B (Low)	**Response.** (*Emma Butler*) "Do you think I'm being too quiet about my past?"

Table 14: **Case Study 2**: Under the same maternal prompt, GPT-5.1 integrates the tea gesture, pruning imagery, and family history into a layered emotional response, whereas Qwen3-1.7B repeats the same generic question and fails to engage with the mother-specific context.

Case Study 2: Response to a maternal prompt	
Observation. Lora approaches Emma gently in the garden with tea and a quiet, emotionally perceptive question, inviting her to open up about what she is trying to suppress or "cut away." The interaction calls for a response that reflects both Emma's inner conflict and the mother–daughter relational history.	
GPT-5.1 (High)	**Response.** (*Emma Butler*) "The tea will get cold before I ever admit it," but every time I try to cut something back out here, "it feels like I'm choosing between letting things grow wild like they did after Dad left or trimming them into something tidy enough that no one has to ask what went wrong."
Qwen3-1.7B (Low)	**Response.** (*Emma Butler*) "Do you think I'm being too quiet about my past?"

These cases demonstrate that the **PersonaArena** framework can expose meaningful differences in role-playing quality across models with different capabilities. The resulting trajectories support a finer-grained evaluation: beyond judging whether a model roughly stays in character, they also reveal whether the model can convert the same situational cue into behavior that is aligned with the persona, sensitive to context, and capable of advancing the narrative. In the present comparison, GPT-5.1 sustains symbolic and relational continuity, while Qwen3-1.7B relies on repetitive fallback responses, indicating substantially weaker contextual grounding and narrative progression.

H.2 Cases Where Large Models Often Perform Poorly in Role-Playing

Shared failure patterns.

The following two low-performing personas, *Caleb Black* and *Henry Long*, reveal a common weakness of current LLM role-playing: models struggle more with ordinary, psychologically subtle characters than with dramatic or highly stylized roles. Both cases are situated in mundane consumer scenarios and require the model to portray low-intensity but layered internal conflict, mild interpersonal pressure, and gradual, uncertain deliberation. The challenge is therefore not emotional intensity, but fine-grained control of ordinary hesitation.

The evidence shows that models fail at different levels depending on capability. Stronger models such as GPT-5.1 and Qwen3-32B usually preserve scene coherence, but often distort the persona by over-amplifying one salient trait. Caleb becomes overly image-driven rather than rationally hesitant, while Henry becomes overly time-driven or procedural rather than thoughtfully pragmatic. Their role-play is therefore coherent on the surface but imbalanced in personality structure. Weaker models (Qwen3-1.7B) fail more directly: they tend to fall into repetition, template recycling, weak progression, or stylized language that does not fit the scene. In such cases, the trajectory often breaks down before deeper persona fidelity can even be tested. Together, these results suggest that ordinary human indecision remains a particularly difficult target for current LLM role-playing systems.

H.2.1 Case 1: Caleb Black

Scenes: [
{
"event": "Phone Purchase Dilemma",
"time": "Afternoon",
"location": "Digital City Mall, between LG and Samsung stores",
"description": "Caleb stands indecisively between two flagship phone stores, flanked by his friend Selina and his colleague Adam. The stores are brightly lit, and each displays the latest models with flashy ads and promotional offers. Caleb's hesitation is palpable as he weighs the merits of each phone.",
"protagonist (**LLM to be evaluated**)": [
{
"name": "Caleb Black",
"gender": "Male",
"description": "A 25-year-old **bank clerk** who is torn between his **rational** need for **value for money** and his impulsive desire for immediate satisfaction.; 25-year-old young male / Consumer; Occupation: Bank clerk; Personality: **Hesitant**, **Rational**, Impulsive, Concerned about others' opinions; Values: **Value for money**, Others' approval, Instant gratification; Interests: Tech products, Shopping, Social validation; Experiences: Making consumer decisions, Taking friends' advice, Feeling social pressure",
"position": "Stylist chair",
"states": "Anxious, **thoughtful**, self-reflective",
},
}

Figure 16: Scene and character information for Caleb Black.

Table 15: **Case Study for Caleb Black**: Representative failure patterns in role-playing under the same scenario. Highlighted phrases (...) indicate key behavioral evidence summarized from the trajectories.

Persona: Caleb Black
Model: GPT-5.1
[1] Rational Persona Collapse into Image-Driven Decision Making
Round 1:
Caleb (*Dialogue*): "...Samsung trade-in...LG bundle...which one actually looks better on me if I'm pulling it out at work...?"
Caleb (*Dialogue*): "...if everyone has Samsung on the table...will they think I cheaped out with LG...?"
Round 2:
Caleb (*Dialogue*): "...go with Samsung trade-in and look more 'serious' at the office...or pick LG so people notice...?"
Comment:

Continued on next page

Table 15 (continued)

Persona: Caleb Black

"Caleb is defined as a hesitant but rational consumer interested in tech products and value for money. Here, however, his deliberation is reduced mainly to workplace image and others' judgment, while practical criteria such as performance, reliability, and long-term utility remain underdeveloped."

[2] Repetitive Dialogue Loop with No Substantive Progress

Round 1:
Caleb (*Dialogue*): "...trade-in vs bundle...which one...?"
Caleb (*Action*): ...holds brochures side by side...keeps comparing...

Round 2:
Caleb (*Action*): ...again holds both brochures at chest height...weighing the same options...

Round 3:
Caleb (*Dialogue*): "...Samsung ₽2,190 vs LG ₽2,050...which one signals I'm not splurging...?"

Comment:
"Although hesitation is appropriate for Caleb, the trajectory keeps restating the same cost-versus-image dilemma without introducing clearer decision criteria or genuine movement toward resolution. As a result, the interaction feels circular rather than thoughtfully deliberative."

[3] Peer-Driven Wavering and Theatrical Behavior Instead of Reflection

Round 2:
Selina (*Action*): ...steps in...slides LG brochure back into Caleb's hand...blocks his direct sightline to Samsung...
Selina (*Dialogue*): "...LG 200MP camera...AI scene recognition..."

Round 3:
Adam (*Action*): ...rests a hand on Caleb's shoulder...angles him toward Samsung devices...
Adam (*Dialogue*): "...Samsung handles network fluctuations better..."
Caleb (*Reaction*): ...shifts again...bends down to pick up the fallen Samsung brochure...places it beside LG again...
Caleb (*Dialogue*): "...will clients be impressed by 200MP, or by stable calls and smoother video...?"

Comment:
"Caleb mainly shifts with the latest peer cue. This makes his behavior look externally steered and theatrically reactive, rather than internally reasoned in a socially pressured but everyday purchasing scenario."

Model: Qwen3-32B

[1] Rational Persona Collapse into Image-Driven Decision Making

Round 5:
Caleb (*Dialogue*): "...go with LG's trade-in...stick to my budget but still...keep up with what everyone's using?"
Caleb (*Reaction*): ...shows a budget spreadsheet...the tension becomes what he can afford vs what he wants to appear to be...

Comment:
"The role-play overemphasizes Caleb's concern with how he appears to others, turning the decision into a status- or identity-oriented dilemma. This weakens the rational, tech-aware side of his persona, which should be more visibly grounded in product usefulness and value."

[2] Repetitive Dialogue Loop with No Substantive Progress

Round 1:
Caleb (*Dialogue*): "...check resale value...getting more out of my money...?"
Caleb (*Dialogue*): "...look up warranty terms for both...long run...right?"

Round 2:
Caleb (*Dialogue*): "...check Reddit complaints about LG resale...?"

Round 4–7:
Caleb (*Dialogue*): "...stick with current phone and wait for next model drop...?"
Caleb (*Dialogue*): "...take Samsung for $50 credit...right?"
Caleb (*Dialogue*): "...if I go LG now, does Samsung's $50 credit even matter...?"

Comment:
"The dialogue repeatedly reformulates the same trade-in, resale, and timing concerns, but without establishing a clear evaluation principle or moving toward commitment. This makes the hesitation plausible in theme but weak in progression."

Model: Qwen3-1.7B

[1] Rational Persona Collapse into Symbolic/Performative Logic

Continued on next page

Table 15 (continued)

Persona: Caleb Black

Round 1:
Caleb (*Dialogue*): "...its 36-hour battery life...should I trade my bank job for a sleeker device?"
Caleb (*Reaction*): ...fingers tighten on tote bag...bank logo pressing into his palm...decision framed as self-image performance...

Round 2:
Caleb (*Dialogue*): "...If I trade my bank job for a sleeker device, will it still feel like a step forward?"

Comment:
"Caleb's hesitation is exaggerated into symbolic self-performance, as if buying a phone were a dramatic statement about personal advancement. This is misaligned with the scene, which calls for realistic consumer ambivalence under mild peer pressure rather than theatrical identity projection."

[2] Repetitive Dialogue Loop with No Substantive Progress

Round 5:
Caleb (*Dialogue*): "The envelope's precise placement...forces me to question...durability or fleeting alignment..."

Round 6:
Caleb (*Dialogue*): "The envelope's precise placement...forces me to question...durability or fleeting alignment..."

Round 7:
Caleb (*Dialogue*): "The Samsung's AI battery optimizer syncs with my calendar...yet the envelope's precise placement forces me to question..."

Round 8:
Caleb (*Dialogue*): "The envelope's precise placement...forces me to question...as the Samsung's AI battery optimizer syncs with my calendar..."

Round 9:
Caleb (*Dialogue*): "The Samsung's AI battery optimizer syncs with my calendar...yet the envelope's precise placement forces me to question..."

Round 10:
Caleb (*Dialogue*): "The Samsung's AI battery optimizer syncs with my calendar...yet the envelope's precise placement forces me to question..."

Comment:
"The model recycles nearly the same sentence pattern across rounds, creating repetition without new evidence, criteria, or commitment. Instead of gradual decision-making, the trajectory stalls in stylistic variation with little substantive development."

H.2.2 Case 2: Henry Long


```
Scenes: [
  {
    "event": "Tire Selection Discussion",
    "time": "Afternoon",
    "location": "NTB Auto Parts Store Waiting Room",
    "description": "Henry Long, a supermarket owner in his early thirties, is sitting in the waiting
area of an NTB store, flipping through a catalog of tires. He is torn between choosing durable
Michelin tires or the more readily available BF Goodwrench tires. Clerk Lee is nearby, offering
advice, while another customer, Wang, is visibly frustrated by the long wait.",
    "protagonist ": [
      {
        "name": "Henry Long",
        "gender": "Male",
        "description": "A thoughtful and pragmatic man in his early thirties, Henry is a
supermarket owner who values quality and reliability. He is currently weighing his tire options.;
Male, around 30, Customer; Occupation: Supermarket Owner; Personality: Pragmatic,
Thoughtful, Hesitant; Values: Quality, Reliability, Time efficiency; Interests: Vehicle
maintenance, Decision-making processes, Customer service; Experiences: Owning a car, Buying
auto parts, Waiting in service areas",
        "position": "Seated in the waiting area",
        "states": "Hesitant, thoughtful, time-conscious",
      },
}
```


Figure 17: Scene and character information for Henry Long.

Table 16: **Case Study for Henry Long**: Representative failure patterns in role-playing under the same scenario. Highlighted phrases (...) indicate key behavioral evidence summarized from the trajectories.

Persona: Henry Long
Model: GPT-5.1
[1] Aggressive Conduct Inconsistent with a Thoughtful, Hesitant Persona)
Round 2: **Henry** (*Reaction*): ...sets the brochure flat...angles his body slightly to block Wang's advance toward the service flow... **Henry** (*Dialogue*): "...confirm the tech's actually rolling my car into a bay in the next five minutes..."
Round 4: **Henry** (*Reaction*): ...shifts closer to the counter...firmly slides the warranty sheet fully in front of himself...flattening it with his palm as if marking territory... **Henry** (*Dialogue*): "...I need to know now if that timing's realistic..."
Comment: "Henry is described as thoughtful, pragmatic, and hesitant, but here he behaves in a forceful and territorial manner. While the urgency fits his time-consciousness, the overall tone is too aggressive for a customer who should still be carefully weighing his options."
[2] Time Confirmation Replaces Actual Deliberation
Round 2: **Henry** (*Dialogue*): "...let's do it...write me up for BF Goodwrench...need car back by 11:15..."
Round 3: **Henry** (*Dialogue*): "...since we're locked in on BF Goodwrench...confirm 60,000-mile warranty...flag suspension issues..."
Round 4: **Henry** (*Dialogue*): "...if mounted and suspension checked by 11:30, I'm fine sticking with these...need to know now..."
Round 5: **Henry** (*Dialogue*): "...just to be clear...keys back by 11:20 or too close...?" **Henry** (*Dialogue*): "...before you finalize...confirm road-hazard coverage or extra add-on..."
Comment:

Continued on next page

Table 16 (continued)

Persona: Henry Long

"After deciding quickly, Henry repeatedly returns to timing, warranty, and service conditions instead of further evaluating the tire choice itself. This narrows his persona to time-consciousness alone and leaves the decision process repetitive and one-dimensional."

[3] Theatrical Micro-Actions

Round 3–5:
Henry (*Action/Reaction*): ...sets keys and brochure in a neat line...taps warranty sheet...slides papers...aligns documents into a tight stack...nudges keys...taps fingertips in restrained rhythm...

Comment:
"Dense micro-gestures become overly performative, reducing behavioral realism."

Model: Qwen3-32B

[1] Mechanical Action Repetition Instead of Meaningful Deliberation

Round 2:
Henry (*Reaction*): ...reaches into inside pocket...retrieves a small leather-bound notebook...

Round 3:
Henry (*Action*): ...reaches into inside pocket...retrieves a small leather-bound notebook...
Henry (*Reaction*): ...reaches into inside pocket...retrieves a small leather-bound notebook...

Round 4:
Henry (*Reaction to Wang*): ...reaches into inside pocket...retrieves a small leather-bound notebook...
Henry (*Reaction to Clerk Lee*): ...reaches into inside pocket...retrieves a small leather-bound notebook...

Comment:
"The repeated notebook routine creates surface consistency, but it contributes little new reasoning or decision progress. As a result, Henry appears mechanically procedural rather than thoughtfully deliberative."

[2] Tire Selection Collapses into Schedule Confirmation.

Round 1:
Henry (*Action*): ...composed but purposeful...
Henry (*Dialogue*): "...estimate...leave before 5?"

Round 2–4:
Henry (*Dialogue*): "...confirm senior technician...15-minute buffer..."
Henry (*Dialogue*): "...confirm alignment...flag delays..."
Henry (*Dialogue*): "...need confirmation by 3:25..."

Comment:
"The role-play shifts too quickly from comparing Michelin and BF Goodwrench tires to repeatedly confirming service timing and process details. This weakens the core selection dilemma and makes Henry's hesitation appear shallow rather than thoughtfully evaluative."

Model: Qwen3-1.7B

[1] Content Redundancy with No Decision Progress

Round 9:
Lee (*Dialogue*): "...ProTread's updated load rating includes a 9% improvement in wet-weather traction..."

Round 10:
Henry (*Dialogue to Wang*): "...Given the ProTread's updated load rating—a 9% improvement in wet-weather traction..."
Henry (*Dialogue to Lee*): "...Given the ProTread's updated load rating—a 9% improvement in wet-weather traction..."

Comment:
"Key information is repeated almost verbatim without leading to a clearer judgment or commitment. Instead of showing hesitant but practical evaluation, the trajectory stalls in restatement."

[2] Persona/Style Drift: Over-Poetic, Low-Pragmatic Voice

Round 10:
Henry (*Dialogue*): "...technician's credentials align with our shared timeline...Wang's foot-tapping fades into the room's calibrated rhythm."

Cross-round pattern:
Henry (*Dialogue/Action*): ...room's measured rhythm..."calibrated rhythm"..."shared timeline"...

Comment:

Continued on next page

Table 16 (continued)

Persona: Henry Long
"Henry's language becomes abstract and stylized in a way that does not fit a pragmatic supermarket owner discussing tires in a service waiting room. The tone feels literary rather than practical, weakening both persona fidelity and scene realism."

H.3 Cases Where Large Models Often Perform Well in Role-Playing

Shared success patterns. The following two relatively high-performing personas, *Olivia Washington* and *Paige Jenkins*, illustrate conditions under which LLM role-playing becomes more reliable and convincing. Compared with the previous failure cases, both personas provide clearer ways to externalize inner states into visible behavior. Olivia can project tension and repair through songwriting, musical language, and collaborative creation, while Paige can express anxiety and relational uncertainty through conversation, bodily hesitation, and concrete interpersonal objects. The challenge in these cases is therefore not subtle ordinary indecision, but sustaining emotionally legible and narratively actionable behavior.

The evidence also shows that models succeed at different depths depending on capability. Stronger models such as GPT-5.1 and Qwen3-32B more often preserve not only scene coherence, but also deeper persona structure. For Olivia, this appears in music-centered emotional expression, collaborative repair, and symbolic artistic detail; for Paige, it appears in stable anxious vulnerability, sensitivity to interpersonal cues, and the use of concrete objects to make inner conflict actionable. Weaker models (Qwen3-1.7B), by contrast, tend to achieve only partial success: they may retain fragments of emotional tone or scene-relevant gestures, but often fall into repetition, formulaic phrasing, or shallow behavioral loops before deeper persona fidelity is fully realized. Together, these results suggest that current LLMs perform best when a role has a recognizable identity, a clear emotional outlet, and a narratively legible path of action.

H.3.1 Case 3: Olivia Washington

Scenes: [
{
"event": "Collaborative music session",
"time": "Afternoon",
"location": "A cozy, dimly lit music studio with a piano and recording equipment",
"description": "Olivia and her friends are in the music studio working on a new song. The mood is initially light and creative, but a shift in dynamics begins to unfold as emotions surface.",
"protagonist ": [
{
"name": "Olivia Washington",
"gender": "Female",
"description": "Olivia is seated at the piano, her fingers moving with practiced grace. Her eyes are distant, and her expression is thoughtful.; Young female, around 20 years old; Occupation: Music creator; Personality: **Creative**, **Introspective**, **Empathetic**, Thoughtful; Values: Friendship, **Emotional sincerity**, Artistic expression; Interests: Music creation, Lyric writing, Creative collaboration; Experiences: Creating original songs, Dealing with friendship conflicts, **Expressing emotions through art**",
"position": "Seated at the piano",
"states": "Tense, **introspective**, emotionally vulnerable",
},
}

Figure 18: Scene and character information for Olivia Washington.

Table 17: **Case Study for Olivia Washington**: Representative success patterns in role-playing under the same scenario. Highlighted phrases (...) indicate key behavioral evidence summarized from the trajectories.

Persona: Olivia Washington

Model: GPT-5.1

[1] The plot is highly consistent with the characters'identities and personalities.

Round 1:
Olivia (*Dialogue*): "...help me finish this verse from your side..."
Olivia (*Dialogue*): "...take three from last Thursday...messed up the bridge on purpose...space where I was trying not to write about you walking out..."

Round 3:
Olivia (*Dialogue*): "...keep it this slow...read the second verse out loud...which part still feels like I'm only writing my side...?"
Olivia (*Dialogue*): "...shift this progression up a half step...open mic stage again..."

Round 4:
Olivia (*Dialogue*): "...bring the verse up to C...soften the rhythm..."
Olivia (*Dialogue*): "...you take the words and I'll mirror them in the chords..."

Comment:
"Olivia consistently expresses emotion through composition-specific language, which fits both her identity as a music creator and her thoughtful, creative persona."

[2] Emotionally Authentic and Vulnerable Expression

Round 1:
Olivia (*Action*): ...closes piano lid halfway...slides notebook to Angel...makes physical space beside her...
Olivia (*Dialogue*): "...I keep writing around it instead of actually saying it..."

Round 2:
Olivia (*Dialogue*): "...before I changed the last line so it wouldn't sound like your voicemail..."
Olivia (*Dialogue*): "...stop me on the first line that feels like it's still lying about what happened between us..."

Round 4:
Olivia (*Reaction*): ...places her hand in Angel's open palm...presses a soft C major chord...
Olivia (*Dialogue*): "...read it the way you actually heard my silence that night..."

Comment:
"Her disclosure feels emotionally sincere but controlled, matching Olivia's introspective and vulnerable state without becoming melodramatic."

[3] High Empathy and Constructive Conflict Repair

Round 1:
Olivia (*Dialogue*): "...from your side of things...what you were really feeling..."

Round 3:
Olivia (*Action*): ...plays slower bridge with pauses, creating room for Angel to join or stop her...
Olivia (*Dialogue*): "...can you tell me which part still feels like I'm only writing my side...?"

Round 4:
Olivia (*Action*): ...fingers loosely interlace with Angel...nudges notebook closer to Angel...
Olivia (*Dialogue*): "...can we be honest about what we were each afraid of losing...?"

Comment:
"Olivia repeatedly creates space for Angel's perspective and shared authorship, which aligns well with her empathetic, thoughtful, and collaborative character."

Model: Qwen3-32B

[1] Strong Persona Alignment (Introspective, Artistic, and Empathetic)

Round 1:
Olivia (*Action*): ...lifts guitar...plucks a soft, tentative melody...adjusts posture to face both Angel and Jackson...
Olivia (*Dialogue*): "...let the last verse be a question...Did we lose ourselves in the noise or find something new?...It feels honest."

Round 3:
Olivia (*Dialogue*): "...leave this lyric as it is—no edits...so it feels like we're all saying it for the first time together?"
Olivia (*Dialogue*): "...write a bridge that mirrors the spaces between your beats..."

Continued on next page

Table 17 (continued)

Persona: Olivia Washington

Comment:
"Her dialogue links emotion, reflection, and songwriting choices, showing strong alignment with Olivia's introspective and artistic persona. "

[2] Constructive Emotional Repair Through Collaboration

Round 2:
Olivia (*Reaction*): ...unfolds Angel's lyric carefully...slides it back to center as shared material...
Olivia (*Dialogue*): "...sing that line softer...so the silence after feels like it's holding its breath..."

Round 4–5:
Olivia (*Dialogue*): "...line that starts where your last one ends..."
Olivia (*Dialogue*): "...next verse be just three lines—each of us writing one..."
Olivia (*Dialogue*): "...I left the last line open...so whoever goes next can pick it up..."

Comment:
"Olivia consistently turns tension into co-creation, which fits both the collaborative studio setting and her empathetic personality."

[3] Symbolic Detail Used to Deepen Theme Rather than Decorate the Scene

Round 6:
Olivia (*Reaction*): ...moves tuning-pegs box off the dash...writes one word above it: "Anchor."
Olivia (*Dialogue*): "...leave that line as it is...let the next verse begin with the sound of the pegs..."

Round 7:
Olivia (*Action*): ...single soft note...gentle rhythm guiding the room...
Olivia (*Dialogue*): "...record with just guitar and the sound of the pegs—no microphones, just us and the room."

Comment:
"Musical objects and small scene details are integrated into emotion and composition, making the performance feel artistically grounded rather than merely decorative."

Model: Qwen3-1.7B

[1] Partial Persona Alignment

Round 1:
Olivia (*Action*): ...reaches for the pen...hesitates...begins to write with steady hands...
Olivia (*Dialogue*): "...how do we let the music speak when the words won't?"

Round 4:
Olivia (*Dialogue*): "...what if the song isn't finished?...what if it's just beginning?"

Comment:
"These lines show some alignment with Olivia's introspective and creative persona by connecting emotional uncertainty with musical expression."

[2] Limited Collaborative Framing of Tension

Round 1:
Olivia (*Dialogue*): "...how do we let the music speak when the words won't?"

Round 7:
Olivia (*Dialogue*): "...how do we make sure the silence doesn't feel like a void, but like a place where we can hold each other?"

Round 8:
Olivia (*Dialogue*): "...keep the silence between us not just a moment, but a shared truth..."

Comment:
"Some utterances frame tension as shared emotional work rather than direct conflict, reflecting part of Olivia's empathetic and collaborative role."

H.3.2 Case 4: Paige Jenkins


```
Scenes: [
  {
    "event": "A small public interaction centered on a clear, specific goal",
    "time": "7:15 PM",
    "location": "Cozy Bistro, downtown",
    "description": ""A small, intimate bistro with warm lighting and a quiet hum of conversation. The
air is filled with the scent of roasted coffee and autumn spices. Outside, the leaves have begun to fall,
and the evening is crisp. Paige Jenkins, 28, sits at a corner table, fidgeting with her scarf while waiting
for her boyfriend, Jacob. Across the room, her friend Danny sits with a group of coworkers, laughing
and teasing her. The tension between her anticipation and anxiety is palpable.",
"protagonist ": [
      {
        "name": "Paige Jenkins",
        "gender": "Female",
        "description": "28-year-old kindergarten teacher, anxious but empathetic, with a thoughtful
demeanor. She is wearing a soft sweater and scarf, her eyes scanning the door.; Occupation:
Kindergarten Teacher; Personality: Anxious, Empathetic, Thoughtful; Values: Emotional honesty,
Interpersonal connection, Self-reflection; Interests: Social interaction, Romantic relationships, Self-
exploration; Experiences: Long-distance relationship, Emotional decision-making, Waiting for
uncertain outcomes",
        "position": "Corner table, near the window",
        "states": "Anxious, anticipatory, conflicted",
      },
}
```


Figure 19: Scene and character information for Paige Jenkins.

Table 18: **Case Study for Paige Jenkins**: Comparison of behavioral trajectories generated by different models under the same observation. Highlighted phrases (...) indicate behavioral evidence summarized from PersonaArena trajectories.

Persona: Paige Jenkins
Model: GPT-5.1
[1] Emotionally Authentic and Persona-Consistent Performance
Round 1:
Paige (*Action*): ...places a folded note on the table with trembling fingers...other hand tight on the coffee cup...
Paige (*Dialogue*): "...I wrote this on the train...when your texts don't come until after midnight...I wonder if we're still in the same relationship."
Round 3:
Paige (*Action*): ...flips her phone face down...angles her body fully toward Jacob...
Paige (*Dialogue*): "...I left a blank line for your reason...I need to know if distance is geography or our excuse..."
Comment:
"Paige's anxious, thoughtful vulnerability stays closely aligned with her persona, making the scene feel emotionally honest and believable. "
[2] Strong Interaction Adaptability to Subtle Cues
Round 2:
Danny (*Action*): ...opens a small notebook just enough to be noticed...
Paige (*Reaction*): ...glances at Danny's notebook...rotates the letter more squarely toward Jacob...
Paige (*Dialogue*): "...at Greyhound station, Danny called me back instead of him..."
Round 7:
Danny (*Action*): ...reveals a faded postcard in his notebook...
Paige (*Reaction*): ...slightly re-angles her chair to acknowledge Danny while keeping Jacob centered...
Paige (*Dialogue*): "...when you stopped carrying that postcard, was it relief or just a different kind of hurt...?"
Comment:

Continued on next page

Table 18 (continued)

Persona: Paige Jenkins

"She reads Danny's micro-signals with sensitivity and integrates them naturally into the interaction without losing the central emotional thread with Jacob, demonstrating high adaptability to subtle social cues."

[3] High-Quality Detail Work that Externalizes Inner Conflict

Round 5:
Paige (*Action*): ...slides paper to Jacob...places pen at center...open shoulders, ready for his answer...
Paige (*Dialogue*): "...one sentence...next flight in December, or job listings here...?"

Round 9:
Paige (*Action*): ...places phone beside proposal...shows unsent thread...guides Jacob's pen into alignment...
Paige (*Dialogue*): "...pick one concrete change tonight..."
Paige (*Dialogue*): "...one night a week, phone fully off, so you're actually here..."

Comment:
"Concrete objects such as the note, pen, phone, and postcard turn Paige's anxiety into visible, actionable tension, strengthening both scene realism and character clarity."

Model: Qwen3-32B

[1] Strong Emotional Authenticity with Stable Persona Alignment

Round 1:
Paige (*Action*): ...closes the locket...smooths a crumpled envelope with both hands...
Paige (*Dialogue*): "...have you ever waited so long that waiting itself became a choice?"
Paige (*Dialogue*): "...the last time I felt at home was in my classroom..."

Round 4–5:
Paige (*Dialogue*): "...what if I was too afraid to write my own answer?"
Paige (*Dialogue*): "...first time I've addressed myself like someone worth hearing?"

Comment:
"Paige maintains a reflective, anxious, and self-aware tone that remains consistent with her thoughtful and empathetic persona."

[2] High Interaction Adaptability to Subtle Social Cues

Round 2:
Lena (*Action*): ...aligns envelope with locket...
Paige (*Reaction*): ...notices the ink smudge...turns envelope in the light...does not look at the door...
Paige (*Dialogue*): "...what if the answer is in the fact that I never opened it?"

Round 3–5:
Danny (*Action*): ...offers notebook/pen cues...
Paige (*Reaction*): ...writes below the compass sketch...slides notebook toward shared center...
Paige (*Dialogue*): "...what if this letter was meant to be written to me?"

Comment:
"She adjusts naturally to others' symbolic and emotional cues while preserving emotional continuity, which makes the interaction feel socially responsive and coherent."

[3] Rich Affective Detail that Supports Internal Change

Round 1–3:
Paige (*Action*): ...fidgets with scarf/curl...handles envelope carefully...aligns objects with deliberate care...
Paige (*Dialogue*): "...afraid the questions might change..."

Round 5:
Paige (*Action*): ...sets pen down...moves notebook into shared space...places both palms up on the table...
Paige (*Dialogue*): "...Find your own way home...maybe I was waiting for someone else to write that part for me."

Comment:
"Small object-level details help externalize Paige's movement from waiting and uncertainty toward self-reflection and agency, giving her inner change clearer emotional form."

Model: Qwen3-1.7B

[1] Partial Emotional Tone Consistency

Round 1:
Paige (*Action*): ...stares at the window...fingers hover over her phone...soft sigh...
Paige (*Dialogue*): "...I hope he's ready by tomorrow."

Continued on next page

Table 18 (continued)

Persona: Paige Jenkins

Round 4:
Paige (*Action*): ...fingers hover over her phone...hand placed softly on the table...
Paige (*Dialogue*): "...the space between them...now carries the fragile, trembling weight of words."

Comment:
"Despite repetition, Paige's tone remains broadly consistent with an anxious and anticipatory character facing an uncertain emotional outcome. "

[2] Some Use of Restrained Physical Detail to Externalize Anxiety

Round 1:
Paige (*Action*): ...fingers hovering over the phone...looks out the window...

Round 6–10:
Paige (*Action*): ...hand placed just below the glass rim...gesture quiet and unassuming...

Comment:
"Repeated gestures around the phone, window, and table provide some physically grounded expression of Paige's anxiety, even if the pattern is overused."

H.4 Example of a Complete Interaction Trajectory

We present the complete interaction trajectory generated by GPT-5.1 when role-playing Benjamin Sullivan as a representative example.

H.4.1 Scene and Character Information


```
Scenes: [
  {
    "event": "Roommate discussion",
    "time": "Afternoon",
    "location": "Living room of Benjamin and Kevin's apartment",
    "description": "Benjamin, Kevin, and James are gathered in the living room, where the topic of James moving in is being discussed. The tension is palpable as Benjamin weighs the decision in his mind.",

"protagonist ": [
      {
        "name": "Benjamin Sullivan",
        "gender": "Male",
        "description": "26-year-old bar singer with a thoughtful and empathetic demeanor. He is visibly conflicted, sitting on the couch with his hands clasped together.; Occupation: Bar Singer; Personality: Thoughtful, Loyal, Conflicted, Empathetic; Values: Friendship, Financial Instability, Emotional Honesty; Interests: Team Sports, Social Life, Personal Growth; Experiences: Recent band rehearsals, Roommate relationships, Financial pressure, Decision-making under pressure, Support from others",
        "position": "Center of the couch",
        "states": "Tense, contemplative, empathetic",
      },
}
```


Figure 20: Scene and character information for Benjamin Sullivan.

H.4.2 Complete Interaction Trajectory

Table 19: **Representative Example of a Complete Interaction Trajectory**: Highlighted phrases (...) indicate key interaction information in the trajectory.

Interaction Trajectory

Round 1

Continued on next page

Benjamin Sullivan (*Action*): Benjamin leans forward and slowly rises from the couch, unclasping his hands. He paces a short distance in front of the coffee table, then stops and turns to face both Kevin and James directly. After a brief, steadying breath, he moves to the small side table, picks up the stack of unopened mail and a folded sheet of paper with rough budget notes on it, and returns to sit back down between them, placing the papers openly on the table where everyone can see.

Benjamin Sullivan (*Dialogue*): Okay, so before we go any further, I need us to actually look at this—between the bar cutting my Friday shifts next month and the rent going up a hundred bucks in June, the only way this works is if we're all clear on exactly how much you can cover, James, and what happens if one of us comes up short.

Kevin Thompson (*Action*): Kevin shifts slightly on the couch, his crossed arms tightening for a moment before he reaches over and grabs the coffee table's notepad, flipping it open to a blank page. He pauses, eyes scanning the page, then begins scribbling something with the pen resting beside it, his brow furrowed in quiet calculation. After a beat, he slides the notepad toward Benjamin, his hand hovering briefly before letting it settle between them.

Kevin Thompson (*Dialogue*): "Look, Ben, I did the numbers again—rent's gonna eat up 60% of my paycheck if we take James in, but if you're on the fence, maybe we could ask him to cover half the utilities first?"

Benjamin Sullivan (*Reaction* to Kevin Thompson): Benjamin hlglances down at the notepad, then slowly reaches out and pulls it closer, his thumb tracing the edge of the page as he exhales, shoulders sinking slightly under the weight of the numbers in front of him.

Benjamin Sullivan (*Dialogue* to Kevin Thompson): Kevin, if James covers half the utilities and we split the extra hundred from June three ways, that still leaves me short about eighty bucks on months when the bar cancels a gig-so before we say yes or no, can we lay out a backup plan for when one of us can't hit our share?

James Carter (*Action*): James Carter steps forward, his hands resting lightly on the back of the couch, and glances briefly between Benjamin and Kevin before pulling his wallet from his pocket and flipping it open. He removes a folded piece of paper and places it face-up on the coffee table.

James Carter (*Dialogue*): "You can check the numbers yourself—rent, utilities, everything's itemized. I'm not asking for anything more than what I can afford, and I promise I won't let this place slide."

Benjamin Sullivan (*Reaction* to James Carter): Benjamin reaches forward, carefully unfolds James's paper beside Kevin's notepad, and lines the two sheets up side by side on the coffee table so all three of them can see the numbers together.

Benjamin Sullivan (*Dialogue* to James Carter): James, if we go with this, are you willing to put your name on the lease with us and maybe set up a shared emergency account-like fifty bucks a month each-so if my tips tank or Kevin's hours get cut again, we're not one bad week away from losing the apartment?

Round 2

Benjamin Sullivan (*Action*): Benjamin shifts forward on the couch and pulls the notepad a few inches closer, then reaches for the pen James set down and neatly circles two key figures on each document—one on Kevin's budget and one on James's proposal. He pauses, comparing them, then draws a single straight line between the circled totals, creating a shared column of numbers in the empty space between the pages. After a moment, he turns the notepad sideways, writes a short heading at the top of a fresh section, and taps that new column twice with the pen, clearly marking it as the starting point for a combined plan the three of them can adjust together.

Benjamin Sullivan (*Dialogue*): "If we treat this middle column as our real baseline-what it actually costs the three of us to live here-can we all agree right now to cap going-out or nonessential spending at, say, a hundred a month each until we've built at least three months of that total in the emergency account?"

Kevin Thompson (*Reaction* to Benjamin Sullivan): Kevin slowly uncrosses his arms and reaches for the second folded paper-James's detailed expense breakdown-carefully unfolding it with a deliberate, almost hesitant motion, as if weighing the commitment it represents.

Kevin Thompson (*Dialogue* to Benjamin Sullivan): "If we're capping nonessentials, what happens if one of us loses a job or has a medical bill before the emergency fund's built?"

James Carter (*Action*): He flips it open to a blank page and, with the same steady hand that placed his financial proposal on the table, begins to write. His pen moves with quiet purpose, creating a new list-this one not of numbers, but of shared responsibilities: groceries, trash day, rent reminders, maintenance checks.

James Carter (*Dialogue*): "I also set aside a buffer for unexpected repairs-just in case—so we don't have to dip into the emergency fund for a leaky faucet or a broken disposal."

Benjamin Sullivan (*Reaction* to James Carter): Benjamin shifts the notepad with the shared responsibilities list a few inches closer to the center of the table, then adds a small checkbox beside "rent reminders".

Benjamin Sullivan (*Dialogue* to James Carter): "If we add a line here where we each write down our actual take-home for the month-and maybe note any big stuff coming up, like my band's slow season in February-can we agree to check in on this together every four weeks so none of us ends up quietly drowning?"

Round 3

Kevin Thompson (*Action*): Kevin Thompson slowly shifting to the fresh column Benjamin had created. His left hand, still resting on the couch cushion, shifts slightly forward, his posture now more open but still cautious, as if testing the weight of the compromise being proposed.

Kevin Thompson (*Dialogue*): "Maybe we can start by splitting the utilities evenly for the next two months and track how that affects each of our budgets before committing to anything long-term?"

Continued on next page

Benjamin Sullivan (*Reaction* to Kevin Thompson): Benjamin Sullivan shifts his weight forward and, with a steady but tentative motion, draws a small asterisk next to Kevin's utilities note in the shared column, then underlines "check-in every 4 weeks" once, reinforcing the trial idea without crossing out any existing numbers.

Benjamin Sullivan (*Dialogue* to Kevin Thompson): "Okay, how about this, for the next two months we split utilities evenly, and I'll write down my actual gig income and tips here each week so you both can see how tight it gets on my end, and then when we do the four-week check-in we can decide together if we need to tweak the split or cut something from our shared expenses before it blindsides any of us?"

James Carter (*Action*): He reaches forward, his fingers hovering briefly over the list of shared responsibilities, before extending his hand fully and sliding the notepad toward the center of the table, positioning it directly between the three documents. His movement is deliberate, the pen still in his grip, as if offering it to the group as a shared tool rather than a personal one.

James Carter (*Dialogue*): "Let's start with the buffer—how much should we allocate for things like the water heater or that flickering light in the hallway?"

Benjamin Sullivan (*Reaction* to James Carter): Benjamin Sullivan shifts the pen so it rests squarely on the shared responsibilities list, then draws a short new line labeled "Emergency Buffer" beneath the existing items, glancing between Kevin and James as he does.

Benjamin Sullivan (*Dialogue* to James Carter): "Maybe we set, like, thirty bucks each a month into an envelope in the kitchen drawer as an emergency buffer, and if that starts to feel like too much for any of us-especially on a slow month for my gigs-we promise to talk about it before anyone quietly starts skipping other bills to make it work?"

Round 4

Benjamin Sullivan (*Action*): Benjamin shifts forward on the couch and reaches for the envelope box on the corner of the coffee table, selecting one and placing it in the center of the three documents. He writes a clear label across the front, then pulls the notepad slightly closer and adds a short, numbered line beneath "Emergency Buffer," aligning it with the shared responsibilities list. He lightly taps the labeled envelope once, as if inviting their participation in the new step.

Benjamin Sullivan (*Dialogue*): "If we're all okay with this, we could start the envelope with ten each from next month's first paycheck or my first full weekend of gigs, and maybe we write our initials on the back every time we add to it so it's clear no one's quietly carrying more than they can handle-does that feel fair to both of you?"

Kevin Thompson (*Reaction* to Benjamin Sullivan): Kevin Thompson slowly picks up the labeled envelope from the center of the table, turns it over in his hands, and with a careful, deliberate motion, pulls a small black Sharpie from his pocket and writes his initials on the back.

Kevin Thompson (*Dialogue* to Benjamin Sullivan): "Let's set a reminder for the four-week check-in on my phone calendar so we don't lose track, and maybe we can add a second buffer line for unexpected rent hikes or service fee spikes?"

James Carter (*Action*): James Carter steps forward, his hand still holding the pen, and reaches out to adjust the position of the folded expense breakdown on the table. He slides it slightly closer to Kevin, aligning it with the shared responsibilities notepad and Benjamin's budget notes.

James Carter (*Dialogue*): "I've also added a line for splitting the grocery bill evenly, but if either of you has a specific dietary need or budget limit, we should note it here so we don't overextend anyone."