Novelty Assessment Report

Paper: Enhancing Persona Following at Decoding Time via Dynamic Importance Estimation for Role-Playing Agents
PDF URL: https://openreview.net/pdf?id=lVE8H8QNcx
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

The utility of Role-Playing Language Agents in sociological research is growing alongside the adoption of Large Language Models. For realism in social simulation, these agents must adhere to their personas defined by character profiles, yet existing strategies—static prompt engineering or costly fine-tuning—fail to adapt personas to dynamic scenarios. Psychological theories, such as the Cognitive-Affective Personality Systems, provide a crucial explanation for this failure: a persona's influence on behavior is not static but varies with the scenarios. This context-dependence highlights the critical need for adaptive persona management. To address this gap, we propose a novel, theory-driven method that dynamically estimates context-dependent persona importance and integrates it into weighted reward-guided decoding, enabling inference-time persona following. Specifically, we introduce Persona Dynamic Decoding (PDD) framework that consists of two key components: (1) Persona Importance Estimation (PIE) module, which dynamically quantifies the contextual importance of persona attributes without requiring ground-truth supervision; and (2) Persona-Guided Inference-Time Alignment (PIA) paradigm, which leverages these importance scores to construct weighted multi-objective rewards and modulate generation probabilities during inference. Extensive experiments show the effectiveness of our method in utterance consistency and behavioral fidelity.

Core Task Landscape

This paper addresses: **Dynamic Persona Following in Role-Playing Language Agents**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Persona Representation and Modeling**
- **Dialogue Generation and Behavioral Alignment**
- **Training and Alignment Paradigms**
- **Evaluation and Benchmarking**
- **Multimodal Role-Playing**
- **Application-Specific Role-Playing Systems**
- **Prompt Engineering and Interaction Patterns**

Complete Taxonomy Tree

- Dynamic Persona Following in Role-Playing Language Agents Survey Taxonomy
- Persona Representation and Modeling
  - Personality-Driven Persona Construction (4 papers)
  - [1] Psyplay: Personality-infused role-playing conversational agents (Yang Tao, 2025) View paper
  - [16] InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews (Chen, 2023) View paper
  - [22] Orca: Enhancing role-playing abilities of large language models by integrating personality traits (Huang Yuxuan, 2024) View paper
  - [37] Illuminating the black box: A psychometric investigation into the multifaceted nature of large language models (Lu Yang, 2023) View paper
  - Profile-Based Persona Representation (5 papers)
  - [4] Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model (Yeyong Yu, 2025) View paper
  - [11] Crafting customisable characters with llms: Introducing simschat, a persona-driven role-playing agent framework (B Yang, 2024) View paper
  - [13] RoleCraft-GLM: Advancing Personalized Role-Playing in Large Language Models (Liang Xuechen, 2023) View paper
  - [39] CharacterGPT: A Persona Reconstruction Framework for Role-Playing Agents (Jeiyoon Park, 2024) View paper
  - [50] Crafting Customisable Characters with LLMs: A Persona-Driven Role-Playing Agent Framework (Bohao Yang, 2025) View paper
  - Knowledge-Enhanced Persona Systems (3 papers)
  - [5] Personalized quest and dialogue generation in role-playing games: A knowledge graph-and language model-based approach (Trevor Ashby, 2023) View paper
  - [7] Personalized Non-Player Characters: A Framework for Character-Consistent Dialogue Generation (Xiao Liu, 2025) View paper
  - [29] PersonaAgent with GraphRAG: Community-Aware Knowledge Graphs for Personalized LLM (Siqi Liang, 2025) View paper
  - User-Centric and Adaptive Personas (4 papers)
  - [17] One chatbot per person: Creating personalized chatbots based on implicit user profiles (Zheng-Yi Ma, 2021) View paper
  - [19] Persona-L has Entered the Chat: Leveraging LLMs and Ability-based Framework for Personas of People with Complex Needs (Lipeipei Sun, 2025) View paper
  - [20] When Personas Talk to You: Evaluating the Evolution of User Personas from Static Profiles to Conversational User Interfaces (Ilkka Kaate, 2025) View paper

- [21] Simulating before planning: Constructing intrinsic user world model for user-tailored dialogue policy planning (He Tao, 2025) View paper
- Dialogue Generation and Behavioral Alignment
  - Dynamic Persona Adaptation ★ (3 papers)
  - [0] Enhancing Persona Following at Decoding Time via Dynamic Importance Estimation for Role-Playing Agents (Anon et al., 2026) View paper
  - [25] HonkaiChat: Companions from Anime that feel alive! (Liu Yueze, 2025) View paper
  - [46] DPRF: A Generalizable Dynamic Persona Refinement Framework for Optimizing Behavior Alignment Between Personalized LLM Role-Playing Agents and Humans (Yao, 2025) View paper
  - Emotion-Aware Role-Playing (2 papers)
  - [15] Enhancing Character-Coherent Role-Playing Dialogue with a Verifiable Emotion Reward (Junqiao Wang, 2025) View paper
  - [43] Emotional RAG: Enhancing Role-Playing Agents through Emotional Retrieval (Le Huang, 2024) View paper
  - Character Consistency Mechanisms (4 papers)
  - [8] Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds (Lei Wang, 2025) View paper
  - [34] TimeChara: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models (Ahn Jae-Woo, 2024) View paper
  - [42] Role play-based question-answering by real users for building chatbots with consistent personalities (Ryuichiro Higashinaka, 2018) View paper
  - [48] Identity Models for Role-Play Dialogue Characters (P Chaffey, 2021) View paper
  - Retrieval-Augmented Role-Playing (2 papers)
  - [28] A quest for information: Enhancing game-based learning with LLM-Driven NPCs (T TÃ³dovÃ¡, 2025) View paper
  - [44] Dynamic Context Adaptation for Consistent Role-Playing Agents with Retrieval-Augmented Generations (Park, 2025) View paper
  - Reasoning-Enhanced Role-Playing (1 papers)
  - [9] Thinking in Character: Advancing Role-Playing Agents with Role-Aware Reasoning (Tang Yihong, 2025) View paper
- Training and Alignment Paradigms
  - Self-Alignment and Bootstrapping (1 papers)
  - [10] Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment (Keming Lu, 2024) View paper
  - Supervised Role-Playing Training (2 papers)
  - [38] Towards immersive computational storytelling: Card-framework for enhanced persona-driven dialogues (BINGLI LIAO, 2024) View paper
  - [41] DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications (Sathya Krishnan Suresh, 2024) View paper
  - Reward Modeling for Role-Playing (2 papers)
  - [32] ChARM: Character-based Act-adaptive Reward Modeling for Advanced Role-Playing Language Agents (Lin, 2025) View paper
  - [45] RoleRMBench & RoleRM: Towards Reward Modeling for Profile-Based Role Play in Dialogue Systems (Hang Ding, 2025) View paper
  - Boundary-Aware and Robustness Training (2 papers)
  - [33] ERABAL: Enhancing Role-Playing Agents through Boundary-Aware Learning (Tang Yihong, 2024) View paper
  - [49] Moral Susceptibility and Robustness under Persona Role-Play in Large Language Models (Davi Bastos Costa, 2025) View paper
- Evaluation and Benchmarking
  - Comprehensive Role-Playing Benchmarks (3 papers)
  - [2] The oscars of ai theater: A survey on role-playing with language models (Chen, 2024) View paper
  - [12] DMT-RoleBench: A Dynamic Multi-Turn Dialogue Based Benchmark for Role-Playing Evaluation of Large Language Model and Agent (Dingbo Yuan, 2025) View paper
  - [27] From Persona to Personalization: A Survey on Role-Playing Language Agents (Chen, 2024) View paper
  - Specialized Evaluation Dimensions (2 papers)
  - [14] VoxRole: A Comprehensive Benchmark for Evaluating Speech-Based Role-Playing Agents (Wu, 2025) View paper
  - [35] Character is Destiny: Can Role-Playing Language Agents Make Persona-Driven Decisions? (Xu Rui, 2024) View paper
- Multimodal Role-Playing (3 papers)
  - [6] OmniCharacter: Towards Immersive Role-Playing Agents with Seamless Speech-Language Personality Interaction (Zhang Haonan, 2025) View paper
  - [24] Towards Embedding Dynamic Personas in Interactive Robots: Masquerading Animated Social Kinematic (MASK) (Jeongeun Park, 2024) View paper
  - [26] Voila: Voice-Language Foundation Models for Real-Time Autonomous Interaction and Voice Role-Play (Shi Yemin, 2025) View paper
- Application-Specific Role-Playing Systems
  - Game-Based Role-Playing Agents (2 papers)
  - [3] Generating dynamic and lifelike NPC dialogs in role-playing games using large language model (Huang, 2024) View paper
  - [36] Larp: Language-agent role play for open-world games (Yan Ming, 2023) View paper
  - Educational and Training Role-Playing (4 papers)
  - [18] Simulating Professional Workplaces: A Pedagogical Framework for Generative AI-Powered Role-Play for Competency-Based Education (L Liu, 2025) View paper
  - [23] An AI-Based Virtual Client for Educational Role-Playing in the Training of Online Counselors (Eric Rudolph, 2024) View paper
  - [30] LLM-Powered AI Tutors with Personas for d/Deaf and Hard-of-Hearing Online Learners (Chen Si, 2024) View paper
  - [47] A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skill Training (Julia Othlinghaus-Wulhorst, 2020) View paper
  - Simulation and Agent Architectures (1 papers)
  - [40] Evolving agents: Interactive simulation of dynamic and diverse human personalities (Li Jiale, 2024) View paper
- Prompt Engineering and Interaction Patterns (1 papers)
  - [31] Toward a Pattern Language for Persona-Based Interactions with LLMs (Will Schreiber, 2025) View paper

Narrative

Core task: dynamic persona following in role-playing language agents. The field centers on enabling language models to adopt and maintain consistent character identities during interaction, spanning several interconnected branches. Persona Representation and Modeling addresses how character traits, memories, and psychological profiles are encoded and retrieved, with works like Characterbox[8] and PersonaAgent GraphRAG[29] exploring structured knowledge bases and graph-based retrieval. Dialogue Generation and Behavioral Alignment focuses on producing utterances that reflect these personas authentically, including dynamic adaptation mechanisms that adjust character behavior in response to conversational context. Training and Alignment Paradigms investigates methods such as reinforcement learning and profile-dialogue alignment (Profile Dialogue Alignment[4]) to improve persona consistency. Evaluation and Benchmarking provides datasets and metrics—exemplified by RoleRMBench[45] and DMT RoleBench[12]—to measure fidelity and coherence. Multimodal Role-Playing extends these ideas to voice and visual modalities (VoxRole[14]), while Application-Specific Role-Playing Systems targets domains like gaming (Dynamic NPC Dialogs[3], Personalized Quest Generation[5]) and professional simulation (Simulating Professional Workplaces[18]). Prompt Engineering and Interaction Patterns examines how carefully designed prompts and conversational scaffolds guide agents toward desired role behaviors.

A particularly active line of work explores how agents can dynamically adjust persona emphasis as dialogue unfolds, balancing static character profiles with real-time contextual cues. Dynamic Importance Estimation[0] sits squarely in this space, proposing mechanisms to weigh different persona attributes adaptively rather than treating all traits uniformly. This contrasts with more static approaches like InCharacter[16] or RoleCraft GLM[13], which rely on fixed persona encodings, and complements recent efforts in dynamic context adaptation (Dynamic Context Adaptation[44]) and flexible persona frameworks (DPRF[46]). Neighboring works such as HonkaiChat[25] and Psyplay[1] also address persona consistency in interactive settings, yet Dynamic Importance Estimation[0] distinguishes itself by explicitly modeling the salience of persona elements over time. This emphasis on adaptive weighting reflects a broader trend toward more nuanced, context-sensitive role-playing agents that can navigate the trade-off between character fidelity and conversational fluidity.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. HonkaiChat: Companions from Anime that feel alive!

Authors: Liu Yueze, Zhang, Yichi, Yueze Liu, Yichi Zhang, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Modern conversational agents, including anime-themed chatbots, are frequently reactive and personality-driven but fail to capture the dynamic nature of human interactions. We propose an event-driven dialogue framework to address these limitations by embedding dynamic events in conversation prompts and fine-tuning models on character-specific data. Evaluations on GPT-4 and comparisons with industry-leading baselines demonstrate that event-driven prompts significantly improve conversational engage...

Relationship Analysis

Both papers belong to the Dynamic Persona Adaptation category, focusing on methods that adjust persona influence during generation. They overlap in addressing persona consistency in role-playing agents through dynamic mechanisms—the original paper proposes inference-time importance estimation and reward-guided decoding (PDD), while the candidate paper introduces event-driven dialogue frameworks that embed dynamic events in prompts. The key difference is that the original paper focuses on computational methods for dynamically weighting persona attributes at decoding time without fine-tuning, whereas the candidate paper emphasizes event-based prompt engineering combined with character-specific fine-tuning to achieve dynamic interactions.

2. DPRF: A Generalizable Dynamic Persona Refinement Framework for Optimizing Behavior Alignment Between Personalized LLM Role-Playing Agents and Humans

Authors: Yao, Bingsheng, Sun Bo, Bingsheng Yao, Dong Yuanzhe, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The emerging large language model role-playing agents (LLM RPAs) aim to simulate individual human behaviors, but the persona fidelity is often undermined by manually-created profiles (e.g., cherry-picked information and personality characteristics) without validating the alignment with the target individuals. To address this limitation, our work introduces the Dynamic Persona Refinement Framework (DPRF). DPRF aims to optimize the alignment of LLM RPAs' behaviors with those of target individuals ...

Relationship Analysis

Both papers belong to the Dynamic Persona Adaptation category, focusing on methods that adjust persona influence based on context during generation. They overlap in addressing the challenge of aligning LLM role-playing agents with target personas across varying scenarios. However, the original paper (PDD) operates at inference-time decoding by dynamically estimating persona importance and modulating token-level generation probabilities without training, while the candidate paper (DPRF) takes an iterative refinement approach that optimizes persona profiles themselves by identifying and correcting cognitive divergences between generated behaviors and human ground truth across multiple iterations.

Contributions Analysis

Overall novelty summary. The paper introduces a Persona Dynamic Decoding (PDD) framework that dynamically adjusts persona attribute importance during inference, grounded in Cognitive-Affective Personality Systems theory. It resides in the Dynamic Persona Adaptation leaf, which contains only three papers total, indicating a relatively sparse research direction within the broader taxonomy of fifty papers. This leaf focuses specifically on methods that adjust persona influence contextually during generation, distinguishing it from static prompt engineering or offline training approaches that dominate neighboring branches.

The taxonomy reveals that Dynamic Persona Adaptation sits within Dialogue Generation and Behavioral Alignment, adjacent to leaves addressing emotion-aware role-playing, character consistency mechanisms, and retrieval-augmented approaches. While neighboring branches like Profile-Based Persona Representation (five papers) and Personality-Driven Persona Construction (four papers) focus on static persona encoding, and Training and Alignment Paradigms addresses offline optimization, this leaf uniquely targets inference-time adaptation. The scope note explicitly excludes static methods, positioning the work at the intersection of real-time behavioral adjustment and persona fidelity maintenance.

Among eighteen candidates examined, none clearly refute the three core contributions: the PDD framework (two candidates examined), the PIE module for unsupervised importance estimation (ten candidates), and the PIA inference-time alignment paradigm (six candidates). The PIE module received the most scrutiny, yet no overlapping prior work emerged from this limited search. This suggests that within the top semantic matches and their citations, the specific combination of dynamic importance quantification without ground-truth supervision and weighted reward-guided decoding appears relatively unexplored, though the modest search scope leaves open the possibility of relevant work beyond these eighteen papers.

Given the sparse population of the Dynamic Persona Adaptation leaf and the absence of refuting candidates among eighteen examined papers, the work appears to occupy a less-crowded niche within role-playing agent research. However, the limited search scale—eighteen candidates from semantic retrieval—means this assessment reflects only a narrow slice of the literature. The taxonomy structure

indicates growing interest in dynamic adaptation mechanisms, yet the specific theory-driven approach to context-dependent persona weighting may represent a novel angle within this emerging direction.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Persona Dynamic Decoding (PDD) framework

Description: A framework that dynamically adapts persona importance to varying scenarios and guides generation without fine-tuning. It consists of two components: Persona Importance Estimation (PIE) for quantifying contextual importance of persona attributes, and Persona-Guided Inference-Time Alignment (PIA) for modulating generation probabilities during inference.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Persona-Infused Dynamic Collaborative Decoding

URL: View paper

Brief Assessment

Dynamic Collaborative Decoding[62] focuses on a real-time weighting mechanism to fuse multiple persona perspectives, while the original paper's PDD framework specifically addresses dynamic persona importance estimation (PIE) and persona-guided inference-time alignment (PIA) for role-playing agents. The candidate's approach to collaborative decoding differs from the original's theory-driven importance quantification method.

---

2. A pre-training based personalized dialogue generation model with persona-sparse data

URL: View paper

Brief Assessment

Persona Sparse Data[61] focuses on pre-training methods for persona-sparse dialogue data with static attribute embeddings and attention routing, not dynamic persona importance estimation for inference-time guided decoding in role-playing agents.

Contribution 2: Persona Importance Estimation (PIE) module

Description: A module that quantifies the influence of each persona attribute by assessing Conditional Mutual Information using only inference-time log probabilities, eliminating reliance on ground-truth supervision. The authors theoretically show that model-generated outputs provide a reliable basis for deriving importance rankings.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Self-supervised alignment with mutual information: Learning to follow principles without preference labels

URL: View paper

Brief Assessment

Self Supervised Alignment[52] focuses on aligning language models to behavioral principles (constitutions) using mutual information between principles and responses, without requiring persona attribute importance estimation or role-playing scenarios. The technical approach and application domain differ fundamentally from PIE's persona-specific importance quantification.

---

2. Conditional contrastive learning for improving fairness in self-supervised learning

URL: View paper

Brief Assessment

Conditional Contrastive Learning[53] focuses on fairness in self-supervised learning by conditioning on sensitive attributes (gender/race), not on quantifying persona attribute importance for role-playing agents using conditional mutual information.

---

3. BANGS: Game-Theoretic Node Selection for Graph Self-Training

URL: View paper

Brief Assessment

BANGS[57] focuses on graph self-training for semi-supervised learning using conditional mutual information for node selection in graph neural networks. This is fundamentally different from estimating persona attribute importance in role-playing language agents.

---

4. An Unsupervised Mutual Information Feature Selection Method Based on SVM for Main Transformer Condition Diagnosis in Nuclear Power Plants

URL: View paper

Brief Assessment

Transformer Condition Diagnosis[58] addresses feature selection for transformer fault diagnosis using mutual information in a completely different domain (power systems engineering). It does not relate to persona attributes, language models, or role-playing agents.

---

5. Mutual Information-Based Unsupervised Feature Transformation for Heterogeneous Feature Subset Selection

URL: View paper

Brief Assessment

Heterogeneous Feature Selection[59] focuses on mutual information-based feature transformation for heterogeneous datasets, not on quantifying persona attribute importance in role-playing agents or language models.

---

6. Testing (Conditional) Mutual Information

URL: View paper

Brief Assessment

Testing Mutual Information[54] focuses on statistical hypothesis testing for (conditional) mutual information with sample complexity bounds. The original paper's PIE module applies CMI estimation to persona attribute importance in role-playing agents using model log probabilities—a distinct application domain with different technical objectives.

---

7. A novel unsupervised approach to heterogeneous feature selection based on fuzzy mutual information

URL: View paper

Brief Assessment

Fuzzy Mutual Information[51] addresses unsupervised feature selection from heterogeneous data using fuzzy mutual information to measure feature relevance and redundancy. The ORIGINAL paper's PIE module quantifies persona attribute importance in role-playing agents using conditional mutual information from model log probabilities. These are fundamentally different applications: feature selection in machine learning versus persona attribute weighting in conversational AI.

---

8. Alignment via Mutual Information

URL: View paper

Brief Assessment

Alignment Mutual Information[56] focuses on computing pointwise mutual information between source and target spans in translation and grounded reference tasks, not on quantifying persona attribute importance in role-playing agents.

---

9. Language Model Based Unsupervised Dependency Parsing with Conditional Mutual Information and Grammatical Constraints

URL: View paper

Brief Assessment

Unsupervised Dependency Parsing[55] focuses on syntactic parsing using conditional mutual information to measure bi-lexical dependencies between words, not persona attribute importance in role-playing agents. The technical domains are entirely different.

---

10. Enhancing Attribute-Factorized Representations in Variational Autoencoder by Regularizing Multiple Mutual Information Elements

URL: View paper

Brief Assessment

Attribute Factorized Representations[60] focuses on unsupervised attribute factorization in variational autoencoders using mutual information regularization. This is fundamentally different from PIE's task of quantifying persona attribute importance in role-playing agents using conditional mutual information from inference-time log probabilities.

Contribution 3: Persona-Guided Inference-Time Alignment (PIA) paradigm

Description: A paradigm that uses importance scores from PIE to construct weighted multi-objective rewards and modulate token-level generation probabilities during decoding. It formulates multi-persona alignment as a normalized reward function that preserves hierarchical structure of persona attributes without requiring training.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards

URL: View paper

Brief Assessment

Rewarded Soups[68] focuses on weight interpolation of models fine-tuned on diverse rewards to achieve Pareto-optimal alignment, not on inference-time token-level probability modulation using persona importance scores for role-playing agents.

---

2. PARM: Multi-Objective Test-Time Alignment via Preference-Aware Autoregressive Reward Model

URL: View paper

Brief Assessment

PARM[63] focuses on multi-objective test-time alignment using preference-aware autoregressive reward models for diverse user preferences, not persona-guided alignment with character attributes in role-playing contexts.

---

3. Guided task planning under complex constraints

URL: View paper

Brief Assessment

Guided Task Planning[66] addresses task planning with weighted reinforcement learning for constraint satisfaction in course/trip planning domains. It does not focus on persona-guided inference-time alignment or token-level generation probability modulation for role-playing agents.

---

4. Aligning LLMs on a Budget: Inference-Time Alignment with Heuristic Reward Models

URL: View paper

Brief Assessment

Inference Time Alignment[65] focuses on general user preference alignment using heuristic reward models and filtering, not on persona-specific multi-objective rewards with hierarchical attribute weighting for role-playing agents.

---

5. Learning to Optimize Multi-Objective Alignment Through Dynamic Reward Weighting

URL: View paper

Brief Assessment

Dynamic Reward Weighting[67] focuses on multi-objective reinforcement learning for online training with adaptive weight adjustment during the RL process, not inference-time alignment with persona attributes for role-playing agents.

---

6. MAVIS: Multi-Objective Alignment via Value-Guided Inference-Time Search

URL: View paper

Brief Assessment

MAVIS[64] focuses on multi-objective alignment using value models for general preference trade-offs (helpfulness, harmlessness, humor), not persona-specific attributes or role-playing contexts. The original work addresses dynamic persona importance estimation for character consistency in role-playing agents, which is a distinct application domain.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Enhancing Persona Following at Decoding Time via Dynamic Importance Estimation for Role-Playing Agents View paper

- [1] Psyplay: Personality-infused role-playing conversational agents View paper
- [2] The oscars of ai theater: A survey on role-playing with language models View paper
- [3] Generating dynamic and lifelike NPC dialogs in role-playing games using large language model View paper
- [4] Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model View paper
- [5] Personalized quest and dialogue generation in role-playing games: A knowledge graph-and language model-based approach View paper
- [6] OmniCharacter: Towards Immersive Role-Playing Agents with Seamless Speech-Language Personality Interaction View paper
- [7] Personalized Non-Player Characters: A Framework for Character-Consistent Dialogue Generation View paper
- [8] Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds View paper
- [9] Thinking in Character: Advancing Role-Playing Agents with Role-Aware Reasoning View paper
- [10] Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment View paper
- [11] Crafting customisable characters with llms: Introducing simschat, a persona-driven role-playing agent framework View paper
- [12] DMT-RoleBench: A Dynamic Multi-Turn Dialogue Based Benchmark for Role-Playing Evaluation of Large Language Model and Agent View paper
- [13] RoleCraft-GLM: Advancing Personalized Role-Playing in Large Language Models View paper
- [14] VoxRole: A Comprehensive Benchmark for Evaluating Speech-Based Role-Playing Agents View paper
- [15] Enhancing Character-Coherent Role-Playing Dialogue with a Verifiable Emotion Reward View paper
- [16] InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews View paper
- [17] One chatbot per person: Creating personalized chatbots based on implicit user profiles View paper
- [18] Simulating Professional Workplaces: A Pedagogical Framework for Generative AI-Powered Role-Play for Competency-Based Education View paper
- [19] Persona-L has Entered the Chat: Leveraging LLMs and Ability-based Framework for Personas of People with Complex Needs View paper
- [20] When Personas Talk to You: Evaluating the Evolution of User Personas from Static Profiles to Conversational User Interfaces View paper
- [21] Simulating before planning: Constructing intrinsic user world model for user-tailored dialogue policy planning View paper
- [22] Orca: Enhancing role-playing abilities of large language models by integrating personality traits View paper
- [23] An AI-Based Virtual Client for Educational Role-Playing in the Training of Online Counselors View paper
- [24] Towards Embedding Dynamic Personas in Interactive Robots: Masquerading Animated Social Kinematic (MASK) View paper
- [25] HonkaiChat: Companions from Anime that feel alive! View paper
- [26] Voila: Voice-Language Foundation Models for Real-Time Autonomous Interaction and Voice Role-Play View paper
- [27] From Persona to Personalization: A Survey on Role-Playing Language Agents View paper
- [28] A quest for information: Enhancing game-based learning with LLM-Driven NPCs View paper
- [29] PersonaAgent with GraphRAG: Community-Aware Knowledge Graphs for Personalized LLM View paper
- [30] LLM-Powered AI Tutors with Personas for d/Deaf and Hard-of-Hearing Online Learners View paper
- [31] Toward a Pattern Language for Persona-Based Interactions with LLMs View paper
- [32] ChARM: Character-based Act-adaptive Reward Modeling for Advanced Role-Playing Language Agents View paper
- [33] ERABAL: Enhancing Role-Playing Agents through Boundary-Aware Learning View paper
- [34] TimeChara: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models View paper
- [35] Character is Destiny: Can Role-Playing Language Agents Make Persona-Driven Decisions? View paper
- [36] Larp: Language-agent role play for open-world games View paper
- [37] Illuminating the black box: A psychometric investigation into the multifaceted nature of large language models View paper
- [38] Towards immersive computational storytelling: Card-framework for enhanced persona-driven dialogues View paper
- [39] CharacterGPT: A Persona Reconstruction Framework for Role-Playing Agents View paper
- [40] Evolving agents: Interactive simulation of dynamic and diverse human personalities View paper
- [41] DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications View paper
- [42] Role play-based question-answering by real users for building chatbots with consistent personalities View paper
- [43] Emotional RAG: Enhancing Role-Playing Agents through Emotional Retrieval View paper
- [44] Dynamic Context Adaptation for Consistent Role-Playing Agents with Retrieval-Augmented Generations View paper
- [45] RoleRMBench & RoleRM: Towards Reward Modeling for Profile-Based Role Play in Dialogue Systems View paper
- [46] DPRF: A Generalizable Dynamic Persona Refinement Framework for Optimizing Behavior Alignment Between Personalized LLM Role-Playing Agents and Humans View paper
- [47] A Technical and Conceptual Framework for Serious Role-Playing Games in the Area of Social Skill Training View paper
- [48] Identity Models for Role-Play Dialogue Characters View paper
- [49] Moral Susceptibility and Robustness under Persona Role-Play in Large Language Models View paper
- [50] Crafting Customisable Characters with LLMs: A Persona-Driven Role-Playing Agent Framework View paper
- [51] A novel unsupervised approach to heterogeneous feature selection based on fuzzy mutual information View paper
- [52] Self-supervised alignment with mutual information: Learning to follow principles without preference labels View paper
- [53] Conditional contrastive learning for improving fairness in self-supervised learning View paper
- [54] Testing (Conditional) Mutual Information View paper
- [55] Language Model Based Unsupervised Dependency Parsing with Conditional Mutual Information and Grammatical Constraints View paper
- [56] Alignment via Mutual Information View paper
- [57] BANGS: Game-Theoretic Node Selection for Graph Self-Training View paper
- [58] An Unsupervised Mutual Information Feature Selection Method Based on SVM for Main Transformer Condition Diagnosis in Nuclear Power Plants View paper
- [59] Mutual Information-Based Unsupervised Feature Transformation for Heterogeneous Feature Subset Selection View paper
- [60] Enhancing Attribute-Factorized Representations in Variational Autoencoder by Regularizing Multiple Mutual Information Elements View paper
- [61] A pre-training based personalized dialogue generation model with persona-sparse data View paper
- [62] Persona-Infused Dynamic Collaborative Decoding View paper
- [63] PARM: Multi-Objective Test-Time Alignment via Preference-Aware Autoregressive Reward Model View paper
- [64] MAVIS: Multi-Objective Alignment via Value-Guided Inference-Time Search View paper

- [65] Aligning LLMs on a Budget: Inference-Time Alignment with Heuristic Reward Models View paper
- [66] Guided task planning under complex constraints View paper
- [67] Learning to Optimize Multi-Objective Alignment Through Dynamic Reward Weighting View paper
- [68] Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards View paper