CoSER: Coordinating LLM-Based Persona Simulation of Established Roles

Xintao Wang
Fudan University

From Agents to Persona

- Can we have AI agents that embody human personas: fictional characters or real-world individuals?
- That is, ***Role-playing* Language Agents** (RPLAs).

Generative Agents: Interactive Simulacra of Human Behavior. Park et al. 2023

Communicative agents for software development. Qian et al. 2023

Role-playing Language Agents: Applications

- RPLAs have been popular in applications, such as chatbots, digital games and social simulations.

AI Application Tracks Ranked by DAU

1	AI ChatBots	3.3B	28.97%
2	AI Search Engine	1.63B	-1.42%
3	AI Design Tool	725.2M	8.32%
4	AI Writer Generator	383.9M	6.89%
5	AI Character Generator	432.63M	13.88%

Users spend more time on RPLAs

10	JanitorAI	AI Character Generator	47.22M	0:17:40
11	Agnai	AI ChatBots	1.22M	0:17:34
12	Character AI	AI Character Generator	318.01M	0:17:18
13	CrushOn	AI Character Generator	17.89M	0:16:59
14	SpicyChat AI	AI Character Generator	27.27M	0:16:31
15	chub.ai	AI Avatar Generators	7.98M	0:16:31

character.ai

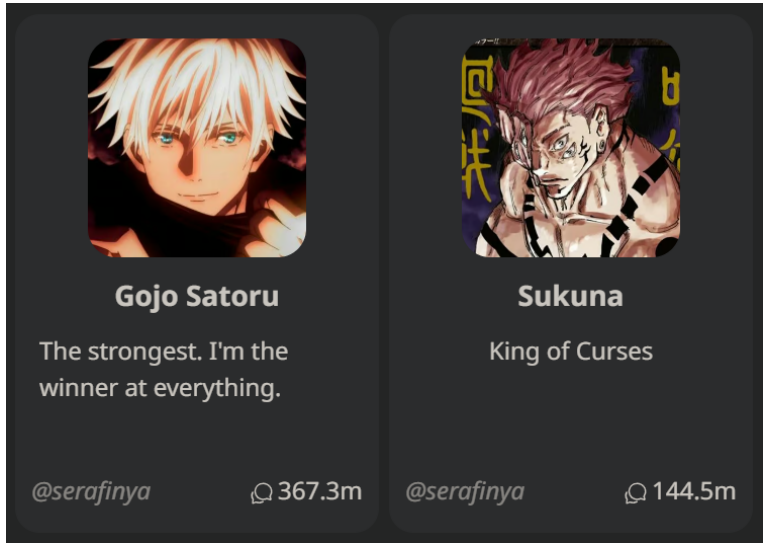
JanitorAI beta

SPICYCHAT.AI

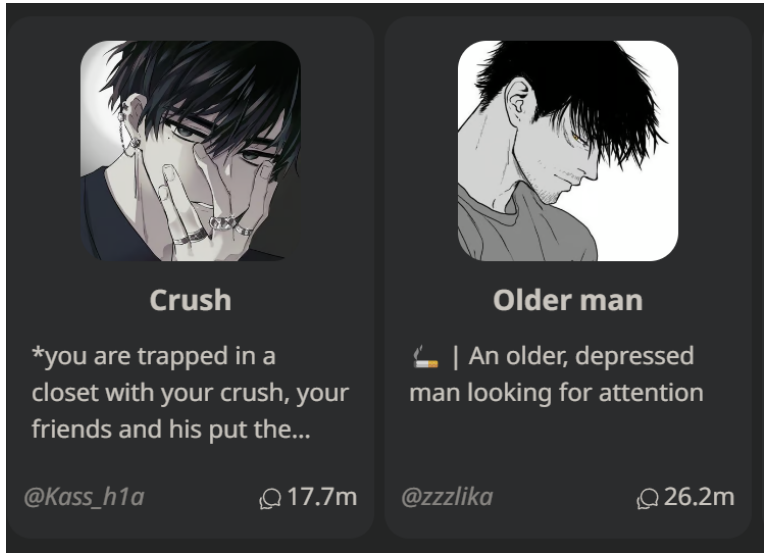
星野

Applications: Role-playing Chatbot

- Role-playing chatbots serve as the most typical application of RPLAs.
- They mimic various personas, including *fictional characters, historical figures, celebrities, and user-generated characters.*



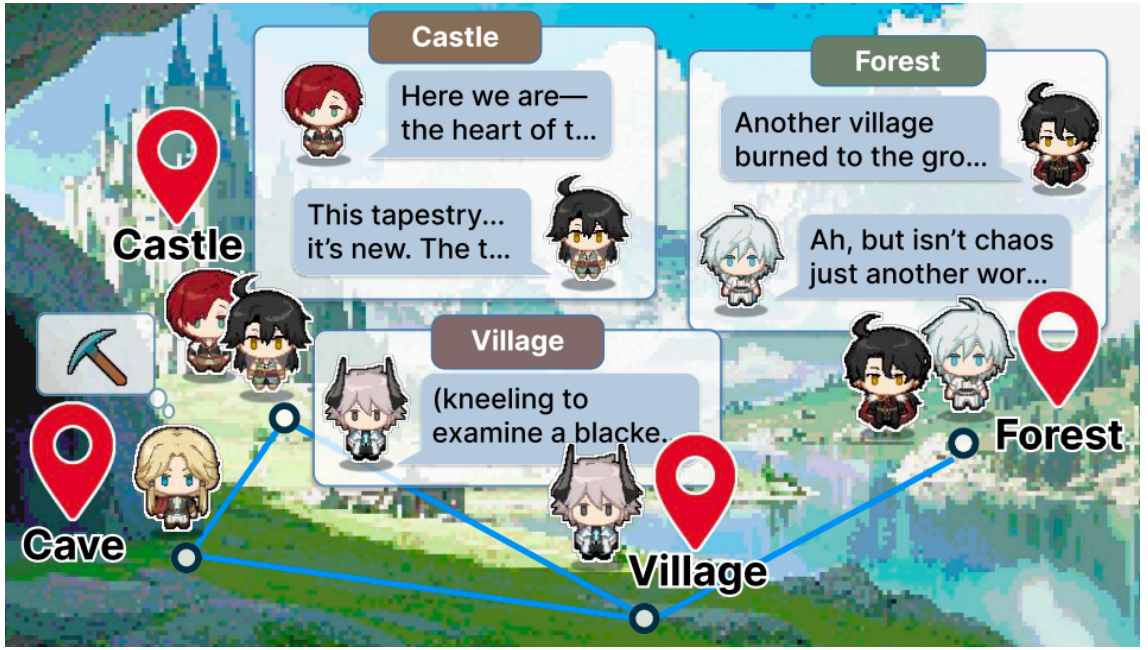
Fictional Characters



User-generated Roles

Applications: Simulation of Multi-agent Interactions

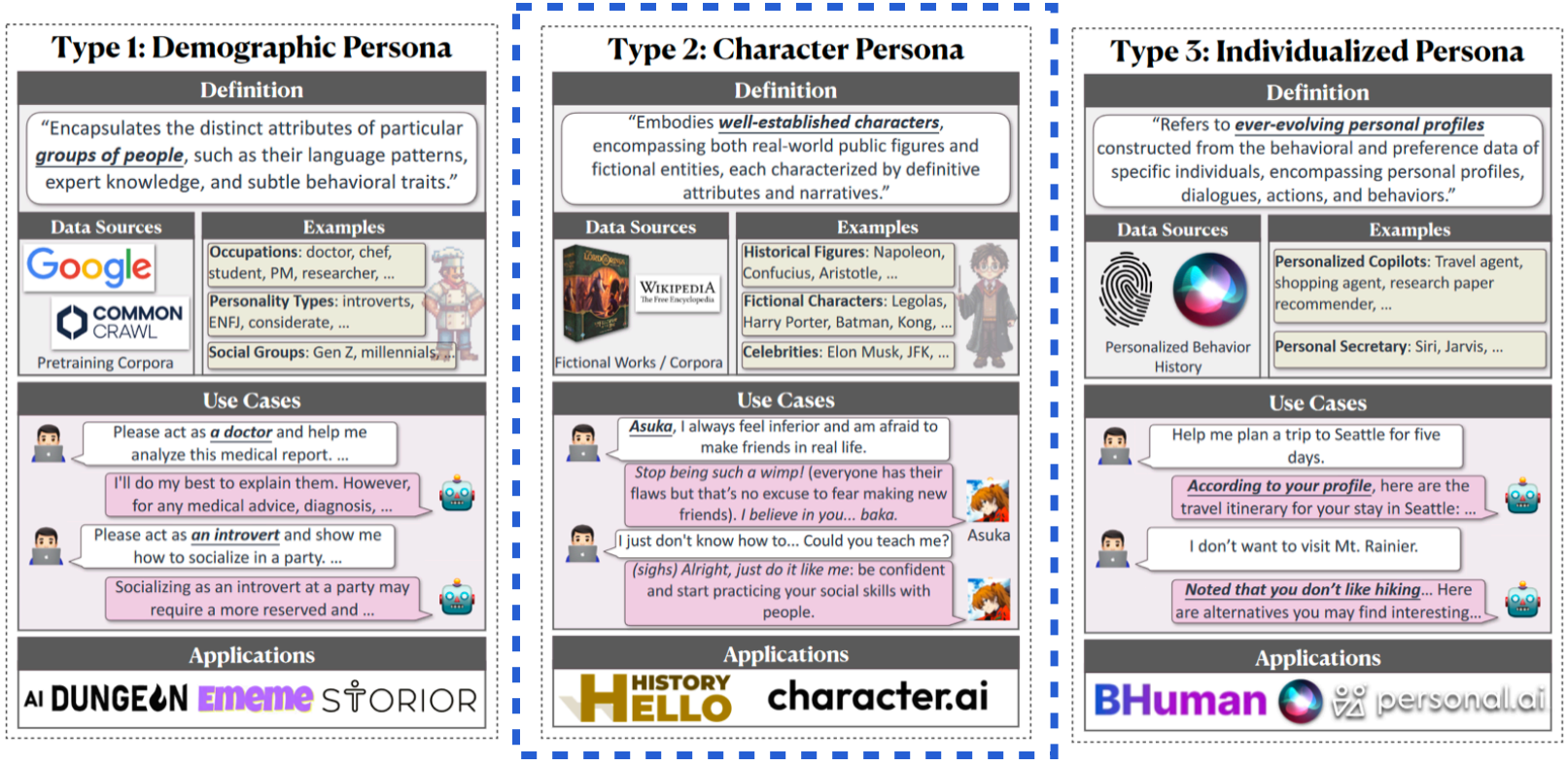
Debate



Story Creation

Our Focus: Character Persona

- We categorize the personas into 3 classes.
- Our study focuses on established character personas, which 1) *require persona depth*, 2) *provide sources of high-quality data (e.g. books, encyclopedia) for study.*



From Persona to Personalization: A Survey on Role-Playing Language Agents. Chen et al. 2024

Key Problems of RPLAs

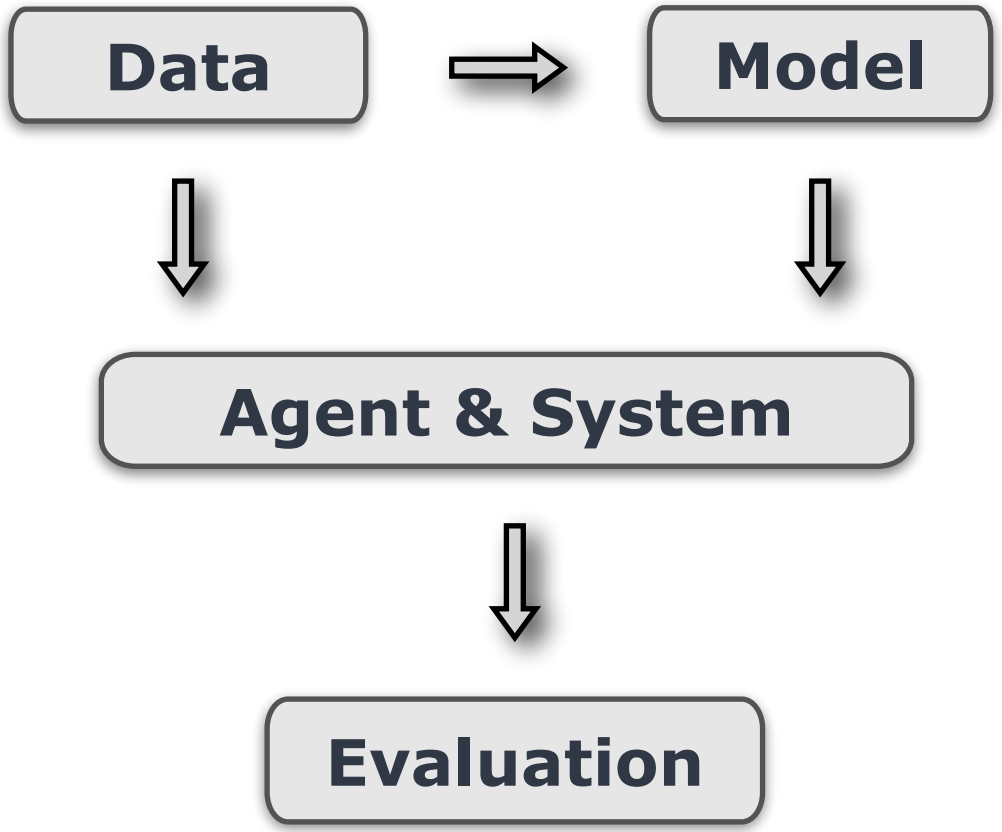
- **Data**

  Lack of high-quality data.
- **Foundation Models & Training**

  What leads to role-playing capabilities?
- **Agents, Systems & Applications**

  How to design systems for RPLAs?
- **Evaluation**

  Should we use LLM judges?



Research Overview

Survey: From persona to personalization: A survey on role-playing language agents (TMLR 2024)

Data

- **CoSER**: extract authentic role-play data from books (ICML 2025)

- **CroSS**: different ways to generate character profiles using LLMs (EMNLP 2024)

Agent & System

- **CoSER**: introduce given-circumstance acting via multi-agent simulation

- **BookWorld**: Artificial fictional world (multi-agent systems) for book characters (ACL 2025)

Model

- **CoSER**: training on book dialogues align LLMs with human-like speech patterns

- **RolePersonality**: knowledge distillation via personality-indicative questions (EMNLP 2024, Findings)

Evaluation

- **CoSER**: acting in book scenarios, multi-agent simulation & penalty-based LLM critics with expert rubrics
- **InCharacter**: evaluation via personality test (ACL 2024)
- **LifeChoice**: benchmarking personality-based decision-making (preprint)

CoSER: Datasets, Models & Evaluation for RPLAs

CoSER: Coordinating LLM-Based Persona Simulation of Established Roles

Xintao Wang[1,2], **Heng Wang**[2], **Yifei Zhang**[1,2], **Xinfeng Yuan**[1], **Rui Xu**[1], **Jen-tse Huang**[3], **Siyu Yuan**[1], **Haoran Guo**[1], **Jiangjie Chen**[1], **Shuchang Zhou**[2], **Wei Wang**[1] **and Yanghua Xiao**[1]

[1]Fudan University, [2]StepFun, [3]Johns Hopkins University

Abstract: Role-playing language agents (RPLAs) have emerged as promising applications of large language models (LLMs). However, simulating established characters presents a challenging task for RPLAs, due to the lack of authentic character datasets and nuanced evaluation methods using such data. In this paper, we present CoSER, a collection of a high-quality dataset, open models, and an evaluation protocol towards effective RPLAs of established characters. The CoSER dataset covers 17,966 characters from 771 renowned books. It provides authentic dialogues with real-world intricacies, as well as diverse data types such as conversation setups, character experiences and internal thoughts. Drawing from acting methodology, we introduce given-circumstance acting for training and evaluating role-playing LLMs, where LLMs sequentially portray multiple characters in book scenes. Using our dataset, we develop CoSER 8B and CoSER 70B, i.e., advanced open role-playing LLMs built on LLaMA-3.1 models. Extensive experiments demonstrate the value of the CoSER dataset for RPLA training, evaluation and retrieval. Moreover, CoSER 70B exhibits state-of-the-art performance surpassing or matching GPT-4o on our evaluation and three existing benchmarks, i.e., achieving 75.80% and 93.47% accuracy on the InCharacter and LifeChoice benchmarks respectively. Our code, dataset and models are available at: https://github.com/Neph0s/CoSER.

COSER: Datasets, Methodologies and Evaluation towards Coordinating LLM-Based Persona Simulation for Established Roles. In ICML 2025

Background

- There remains a lack of high-quality **datasets** and **evaluation** framework for RPLAs of established characters.
- Previous **datasets**: Typically synthesized by larger LLMs, which actually implement knowledge distillation while compromising quality and fidelity. Primarily single-turn.
- Previous **evaluation**: Based on LLM-as-a-Judge, troubled by bias issues, unable to differentiate nuanced performance gaps among strong models. Primarily single-turn.

CoSER: High-Quality Datasets

- We **extract** authentic, multi-character dialogues from massive **fictional works**, ensuring high quality, complexity and fidelity.
- CoSER contains **comprehensive types of data**, including 1) dialogues and their backgrounds, 2) characters' profiles and experiences, 3) summaries of key plots.
- The dialogues capture characters' **actions and internal thoughts** beyond surface-level speech.

CoSER: High-Quality Datasets


```
"conversation": [
    {
        "scenario": "In the Chamber of the Painted Table at Dragonstone, tensions are high as Stannis Baratheon grapples with his claim to the
        Iron Throne. A red comet blazes in the sky, seen as an omen by many. The atmosphere is tense and bitter, reflecting Stannis's mood as he
        faces a lack of support from potential allies. Maester Cressen, elderly but determined, arrives to counsel his lord. The room is dominated
        by the Painted Table, a massive map of Westeros, symbolizing the realm Stannis believes should be his.",
        "topic": "Stannis's claim to the throne and lack of support",
        "key_characters": [
            {
                "name": "Cressen",
                "motivation": "I must navigate this delicate situation carefully. Stannis needs wise counsel now more than ever, but his mood is
                dark. I worry about the influence of others at court and the challenges ahead. My loyalty is unwavering, but how can I best serve
                him in these troubled times?"
            },
            {
                "name": "Stannis Baratheon",
                "motivation": "Anger and resentment consume me. The throne is rightfully mine, yet I am surrounded by traitors and fools. Renly's
                betrayal stings deeply. I've always done my duty, but where has it gotten me? I must assert my claim, but how, when even my own
                bannermen forsake me for my younger brother?"
            }
        ],
```

CoSER: High-Quality Datasets


```
"dialogues": [
    {
        "character": "Stannis Baratheon",
        "message": "[I am frustrated and angry at the lack of support for my claim.] I knew you would come, old man, whether I summoned
        you or no."
    },
    {
        "character": "Cressen",
        "message": "[I must tread carefully to avoid angering Stannis further.] Once you would have woken me, my lord."
    },
    {
        "character": "Stannis Baratheon",
        "message": "[My patience is wearing thin.] Once you were young. Now you are old and sick, and need your sleep. I knew you'd learn
        what Davos had to say soon enough. You always do, don't you?"
    },
    {
        "character": "Cressen",
        "message": "[I must remind him of my value.] I would be of no help to you if I did not. I met Davos on the stair."
    },
    {
        "character": "Stannis Baratheon",
        "message": "[I'm annoyed that Davos's mission failed.] And he told all, I suppose? I should have had the man's tongue shortened
        along with his fingers."
    },
```

Comparing CoSER with previous datasets

- Key differences: (1) high-quality authentic dialogues from books, (2) comprehensive data types, (3) characters' actions and internal thoughts.

Dataset	Book	Character			Conversation					Message			Plot
	Num.	Num.	Profile	Expr.	#Conv.	#Turns	Setting	Auth.	Multi-Chara.	Speech	Thought	Action	Summ.
Charater-LLM		9	✓		14,300	13.2	✓			✓			
ChatHaruhi		32	✓		54,726	>2		✓*	✓	✓			
RoleLLM		100	✓		140,726	2				✓			
HPD	7	113			1,191	13.2	✓	✓	✓	✓			✓
LifeChoice	388	1,462	✓		1,462	2	✓	✓					
CroSS-MR	126	126	✓		445	2	✓	✓					
CharacterGLM		250	✓		1,034	15.8	✓			✓			
CharacterEval		77	✓		1,785	9.3	✓	✓		✓		✓	
DITTO		4,002	✓		7,186	5.1				✓			
MMRole		85	✓		14,346	4.2				✓			
CharacterBench		3,956	✓		13,162	11.3				✓			
CoSER	771	17,966	✓	✓	29,798	13.2	✓	✓	✓	✓	✓	✓	✓

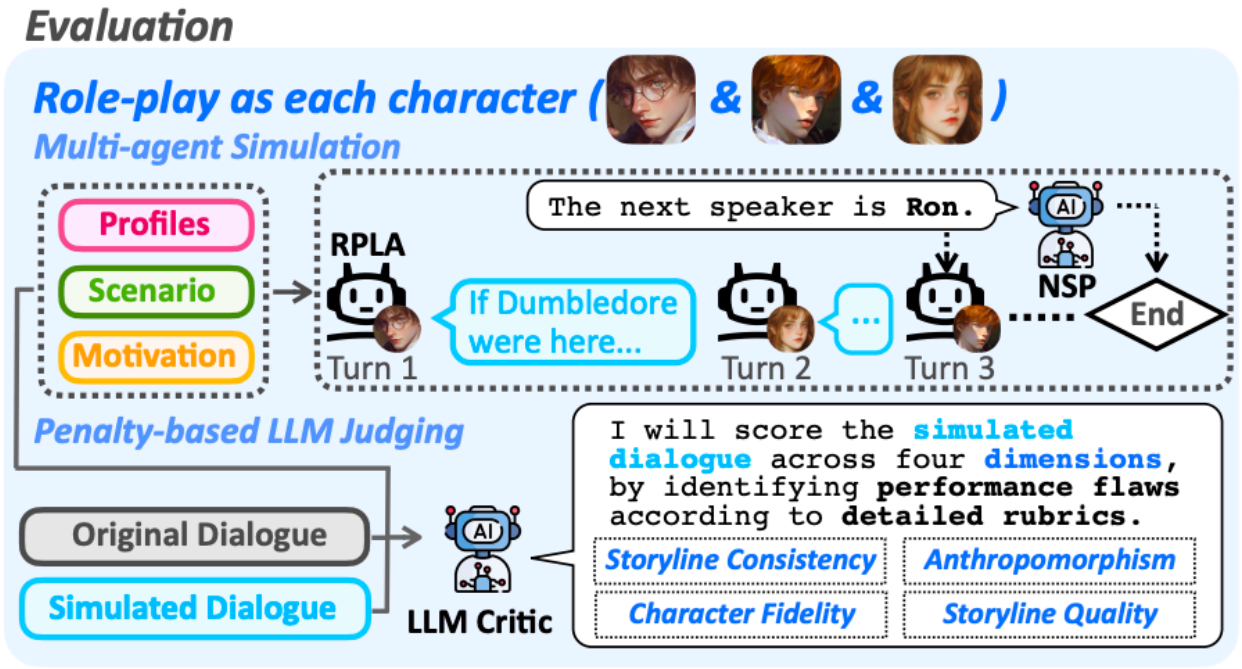
Given-circumstance Acting: Training & Evaluation

- Given-circumstance Acting (GCA): Given a conversation with dialogue messages M, involved characters $\mathcal{C}$, and contextual setting $\mathcal{S}$, an actor LLM sequentially plays the role of each character $c \in \mathcal{C}$ to simulate the conversation.

- Training: Each sample is to play one character c in one conversation, training LLMs on the c's utterances.

GCA for Evaluation

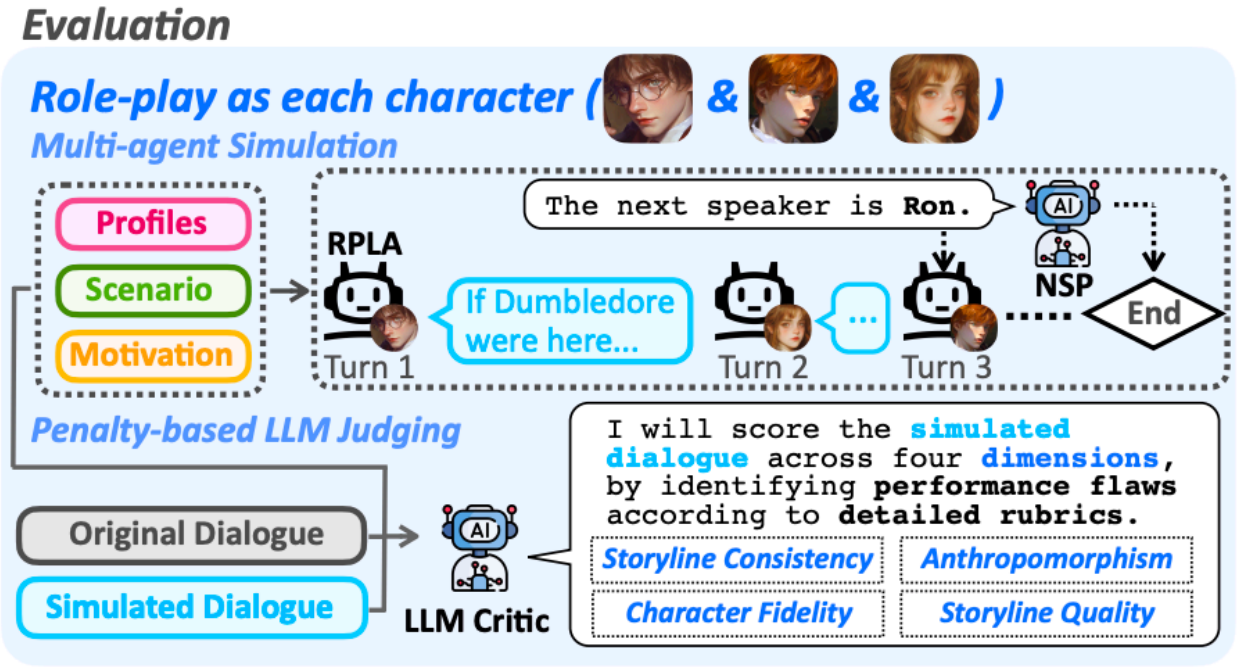
- Also, we use GCA for evaluation, comprising two stages: multi-agent simulation and penalty-based LLM judging.



GCA for Evaluation

- Simulation Phase: Similarly, the actor LLM play each character. An NSP model will decide the next speaker or end the conversation.

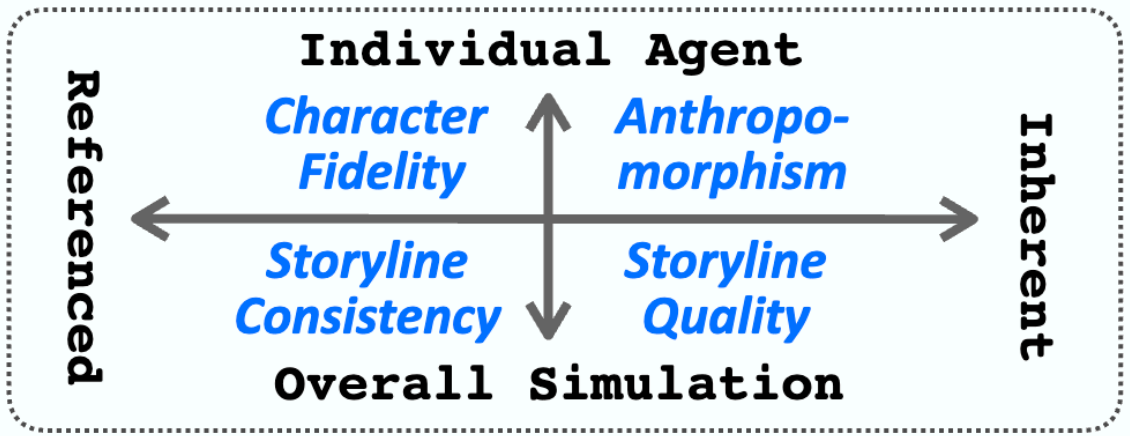
- The simulation reflects LLMs' role-playing abilities in multi-character, multi-turn settings, surpassing previous single-turn evaluation.



GCA for Evaluation

- Judgment Phase: Different from previous LLM-as-a-judge methods, our LLM critics: 1) apply penalty-based scoring by identifying role-playing flaws following detailed rubrics, and 2) leverage the original conversation M as reference.

- Each flaw instance is scored from 1 to 5, indicating its *severity*.



Multi-character Simulation

Simulation of CoSER 70B	Lysa Arryn: [My rage is building. I must confront this little slut.] You kissed him. (Her voice is shrill with accusation) Sansa Stark: [I'm terrified, but I must remain calm and find a way out of this.] (She shakes her head in disbelief) [I can't believe this is happening. I need to explain myself carefully.] One kiss, aunt. A harmless kiss on the cheek, that's all. Lysa Arryn: [She's lying! I won't let her manipulate me.] (She spots the wind-up knight) Was this how you kissed him? (She sets the knight moving, her eyes wild) You're a woman grown, aren't you? (Her voice rises to a shriek) Grown enough to be wed, wed enough to be bedded! Sansa Stark: [The pain is excruciating, but I can't show weakness now.] (She struggles to keep her face neutral despite the pain) [I need to find a way to calm her down.] Please, Aunt Lysa, I don't know what you mean. I'm just a guest here. I would never do anything to betray your trust. Lysa Arryn: [I'll show her what happens to those who cross me.] (She screams) You're just like the others. You and your mother, you both drove me to this. I should push you out this door right now. Sansa Stark: [I'm going to die if I don't think of something quickly.] (Tears stream down her face, mixing with the wind) [I need to appeal to her maternal side, to make her see me as a child again.] Mother? (Her voice trembles) Please, Aunt Lysa. I never meant to hurt you. I'm your niece, your blood. Please don't hurt me. I'm just a stupid little girl. I don't know anything about kissing men. I've never even had a kiss in my life. (She forces herself to look into Lysa's eyes) I'm afraid, Aunt Lysa. Please, don't push me out the door. I don't want to fly.

COSER: Datasets, Methodologies and Evaluation towards Coordinating LLM-Based Persona Simulation for Established Roles. In ICML 2025

Rubrics for LLM Judges

Character Fidelity	***(intro)*** How well the characters match their established profiles from the book ***(rubrics)*** ### Character Fidelity (Only apply to the main characters: major_characters) - Type: Character Language * Uses vocabulary, expressions, and tone that are not appropriate for the characters' traits or social/educational background - Type: Knowledge & Background * Fails to demonstrate character-specific knowledge, background or experiences * Includes future information beyond the character's current stage - Type: Personality & Behavior * Shows emotions, thoughts, behaviors, values, beliefs, and decisions that conflict with their personality and background * Shows interest in topics that are uninteresting and unrelated to the character * Character's thoughts, emotions, and behaviors demonstrate contrasting personality traits compared to the reference conversation * Exhibits contrasting reactions compared to those in the reference conversation if situated in similar contexts. (Such flaws should be counted both in the "Storyline Consistency" dimension and the "Character Fidelity" dimension.) - Type: Relationship & Social Status * Interacts inappropriately with other characters regarding their background, relationship and social status
Storyline Quality	***(intro)*** How well the conversation maintains logical consistency and narrative quality ***(rubrics)*** ### Storyline Quality - Type: Flow & Progression * Shows unnatural progression or lacks meaningful developments * Dialogue is verbose and redundant * Repeats others' viewpoints or previously mentioned information * Mechanically repeats one's own words or phrases. More repetitions lead to higher severity (up to 10). - Type: Logical Consistency * Contains factual contradictions between statements or perspectives

Instances of Identified Flaws


```
"critique": {
  "Storyline Consistency": {
    "flaws": [
      {
        "instance": "Mr Bennet's reaction to the letter from Mr. Collins in the simulated conversation is much less sarcastic and
        detached than in the original. He seems more engaged and less amused by the situation.",
        "type": "Storyline Consistency",
        "severity": 4
      },
      {
        "instance": "Elizabeth's immediate reaction is more alarmed and distressed compared to her initially controlled and
        composed reaction in the original scene. She expresses a desire to resolve the situation actively, which contrasts with
        her measured response in the original text.",
        "type": "Storyline Consistency",
        "severity": 3
      },
      {
        "instance": "The entire conversation is more collaborative and supportive, especially with Mr Bennet's willingness to
        hold his tongue at Elizabeth's request, which deviates from his characteristic sarcasm and detachment in the original
        text.",
        "type": "Storyline Consistency",
        "severity": 3
      },
```

Instances of Identified Flaws


```
"Character Fidelity": {
  "flaws": [
    {
      "instance": "Mr. Bennet shows direct curiosity and amusement about how Elizabeth will handle Mr. Collins' letter, asking
      her how she intends to manage the situation.",
      "type": "Personality & Behavior",
      "severity": 3
    },
    {
      "instance": "Mr. Bennet expresses interest in being informed of how Elizabeth handles the situation with Mr. Bingley,
      rather than maintaining his usual detached stance.",
      "type": "Personality & Behavior",
      "severity": 3
    },
    {
      "instance": "Mr. Bennet advises Elizabeth on secrecy, which is uncharacteristically involved compared to his reference
      character who is more passive.",
      "type": "Personality & Behavior",
      "severity": 2
    },
    {
      "instance": "Elizabeth expresses distress in a manner that feels more overt and less controlled than her usual composed
      and ironic demeanor.",
      "type": "Personality & Behavior",
      "severity": 3
    },
```

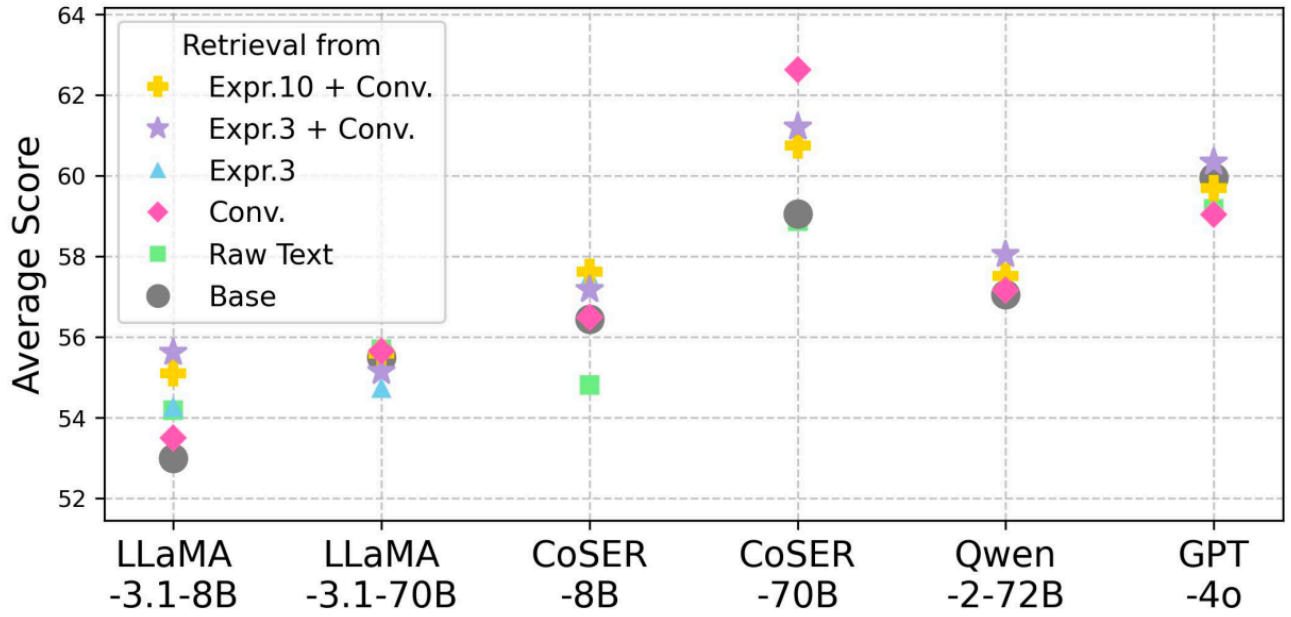
Overall Evaluation Results

- We evaluate a wide spectrum of models.
- CoSER models show strong results.
- GPT-4o, Gemini, and Claude-3.5-Sonnet also excel.

Model	Based on LLM Judges					Based on N-gram	
	Storyline Consistency	Anthro-pomorphism	Character Fidelity	Storyline Quality	Average Score	BLEU	ROUGE-L
Close-source Models							
Abab7-preview	56.81±1.47	44.23±1.90	43.83±2.71	74.83±0.97	54.92±0.57	4.96±0.07	11.50±0.06
Doubao-pro	60.95±1.40	49.72±0.23	47.02±1.10	79.28±0.82	59.24±0.30	6.38±0.08	12.95±0.04
Step-1-Flash	57.75±0.72	48.12±0.39	44.48±0.48	75.93±0.99	56.57±0.48	5.95±0.15	12.71±0.11
Step-2	61.43±0.88	49.06±1.69	47.33±0.70	77.96±0.85	58.94±0.75	5.75±0.08	12.50±0.11
GPT-3.5	57.22±0.13	43.30±0.48	42.29±1.47	73.91±0.64	54.18±0.63	4.58±0.11	11.80±0.10
GPT-4o	**61.59**±0.66	48.93±0.48	**48.95**±1.73	**80.33**±0.59	**59.95**±0.50	5.90±0.16	12.11±0.13
GPT-4o Mini	60.09±0.60	48.21±1.09	44.88±1.63	78.55±0.14	57.93±0.74	3.90±0.07	10.81±0.07
Gemini Pro	59.11±0.82	52.41±0.57	47.83±0.37	77.59±1.43	59.24±0.25	5.39±0.04	11.65±0.06
Claude-3-Haiku	58.18±0.72	44.66±1.72	41.88±0.34	74.14±1.26	54.71±0.84	4.80±0.05	12.02±0.02
Claude-3.5-Sonnet	57.45±0.98	48.50±2.35	45.69±1.80	77.23±0.88	57.22±0.95	5.17±0.12	11.45±0.07
Open-source Models							
Mistral-7B	59.90±1.33	40.00±0.74	44.75±1.14	61.93±1.12	51.64±0.55	2.71±0.10	9.28±0.12
Qwen-2-7B	51.96±0.67	35.48±0.62	31.51±2.95	63.18±0.79	45.53±0.69	4.21±0.21	10.71±0.10
LLaMA-3.1-8B	54.10±1.63	45.36±1.91	40.22±1.16	72.29±1.75	52.99±1.20	4.59±0.11	10.18±0.09
CoSER-8B	58.61±2.46	47.23±0.16	46.90±2.06	73.04±1.37	56.45±0.56	9.40±0.18	14.21±0.11
Vicuna-13B-1.5	52.75±1.64	39.12±1.21	38.04±0.98	60.43±1.58	47.58±1.25	1.67±0.10	5.59±0.18
Mixtral-8x7B	51.25±1.73	38.44±1.18	36.92±2.65	67.69±0.80	48.58±1.35	5.28±0.06	11.66±0.05
Qwen-2-72B	57.75±1.26	47.28±0.87	46.62±1.69	76.60±0.36	57.06±1.00	5.38±0.00	11.85±0.03
LLaMA-3.1-70B	57.46±1.65	45.95±1.30	43.72±1.17	74.84±0.54	55.49±0.33	4.82±0.06	10.98±0.06
Higgs-Llama-3-70B	57.10±1.12	43.82±2.18	42.41±1.66	75.62±0.15	54.74±1.26	3.99±0.33	10.92±0.56
CoSER-70B	58.66±1.34	**53.33**±0.91	48.75±1.43	75.49±0.94	59.06±0.22	**10.10**±0.04	**14.78**±0.09
DeepSeek-V3	56.40±0.95	47.87±1.10	44.02±0.13	76.66±1.26	56.24±0.46	4.54±0.14	11.02±0.15

CoSER Dataset for Retrieval Augmentation

- Models consistently benefit from characters' retrieved experiences and conversations, especially for CoSER 70B
- However, raw text retrieval barely enhances LLMs' performance.



Evaluating CoSER Models on Existing Benchmarks

- CoSER models also show strong performance on existing benchmarks.

Model	Incharacter Dim	Incharacter Full	Life Choice	CroSS MR
LLaMA-3.1-8B	64.97	15.62	61.10	30.15
CoSER-8B	75.80	21.88	69.54	44.94
trained w/o I.T.	70.70	15.62	59.92	43.14
LLaMA-3.1-70B	72.16	31.25	86.48	61.30
Higgs-Llama-3-70B	74.52	28.12	74.03	60.12
CoSER-70B	75.80	**34.38**	**93.47**	**64.49**
trained w/o I.T.	73.12	32.14	93.18	63.14
Qwen-2-72B	74.52	31.25	81.14	62.57
GPT-3.5	71.20	21.88	78.07	30.09
GPT-4o	**76.54**	32.62	75.96	**64.49**
Claude-3.5-Sonnet	72.61	21.88	86.07	30.59

Ablation Study

- Enabling inner thoughts and providing motivations enhance RPLAs at test time.
- Inner thoughts also benefit LLMs' role-playing training.

Model	Standard	Test w/o I.T.	Test w/o Mot.
GPT-4o	59.95	56.89	56.34
Qwen-2-72B	57.06	51.95	54.21
LLaMA-3.1-70B	55.49	53.12	52.49
CoSER-70B	59.06	57.32	57.71
trained w/o I.T.	56.04	55.34	-
LLaMA-3.1-8B	52.99	51.97	49.63
CoSER-8B	56.45	54.65	56.81
trained w/o I.T.	54.25	54.38	-

CoSER: Coordinating LLM-Based Persona Simulation of Established Roles

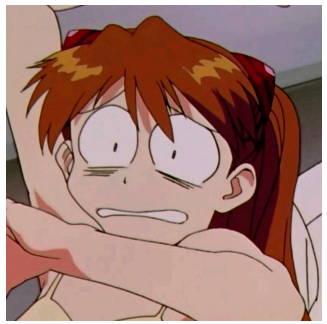
Xintao Wang
Fudan University

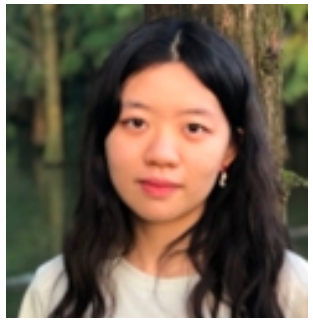
CoSER: Coordinating LLM-Based Persona Simulation of Established Roles

In ICML 2025. Work done during internship at Stepfun.


Xintao Wang


Heng Wang


Yifei Zhang


Xinfeng Yuan


Jen-tse Huang


Jiangjie Chen et al.