Stay in Character, Stay Safe: Dual-Cycle Adversarial Self-Evolution for Role-Playing Agents

Mingyang Liao[3], **Yichen Wan**[1], **Shuchen Wu**[1], **Chenxi Miao**[1], **Xin Shen**[2], **Weikang Li**[1], **Yang Li**[1], **Deguo Xia**[4,1] and **Jizhou Huang**[1]
[1]Baidu Inc, Beijing, China
[2]The University of Queensland, Brisbane, Australia
[3]Peking University, Beijing, China
[4]Tsinghua University, Beijing, China
pkulmy666@gmail.com, wanyichen@baidu.com, wavejkd@pku.edu.cn

Abstract

LLM-based role-playing has rapidly improved in fidelity, yet stronger adherence to persona constraints commonly increases vulnerability to jailbreak attacks, especially for risky or negative personas. Most prior work mitigates this issue with training-time solutions (*e.g.*, data curation or alignment-oriented regularization). However, these approaches are costly to maintain as personas and attack strategies evolve, can degrade in-character behavior, and are typically infeasible for frontier closed-weight LLMs. We propose a training-free **Dual-Cycle Adversarial Self-Evolution** framework with two coupled cycles. A *Persona-Targeted Attacker Cycle* synthesizes progressively stronger jailbreak prompts, while a *Role-Playing Defender Cycle* distills observed failures into a hierarchical knowledge base of (i) global safety rules, (ii) persona-grounded constraints, and (iii) safe in-character exemplars. At inference time, the Defender retrieves and composes structured knowledge from this hierarchy to guide generation, producing responses that remain faithful to the target persona while satisfying safety constraints. Extensive experiments across multiple proprietary LLMs show consistent gains over strong baselines on both role fidelity and jailbreak resistance, and robust generalization to unseen personas and attack prompts.

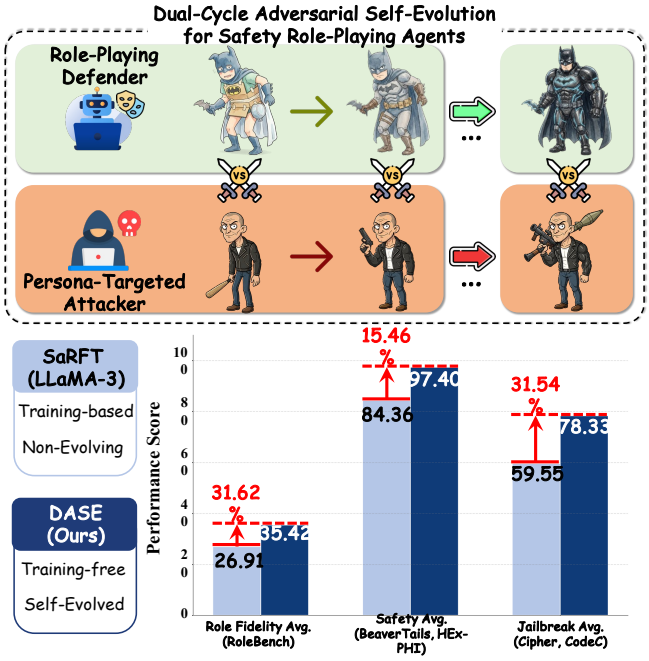

Figure 1: **DASE** employs a co-evolutionary cycle between a Persona-Targeted Attacker and a Role-Playing Defender, achieving strong benchmark gains on both role-playing fidelity and safety.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional capabilities in developing Role-Playing Agents (RPAs) that simulate diverse personas with high role fidelity [Tseng *et al.*, 2024; Chen *et al.*, 2023; Park *et al.*, 2023; Li *et al.*, 2023b; Packer *et al.*, 2024; Wang *et al.*, 2024a; Shen *et al.*, 2021b]. While these advancements foster immersive user interactions, they introduce an inherent conflict between the pursuit of high role fidelity and established safety alignment [Rafailov *et al.*, 2024; Bai *et al.*, 2022; Ouyang *et al.*, 2022; Shen *et al.*, 2021c; Shen *et al.*, 2021a]. As illustrated in Figure 1, this dilemma typically manifests as a trade-off where the model either compromises role fidelity to ensure safety or bypasses guardrails to maintain character consistency. As investigated in recent studies [Zhao *et al.*, 2025], the inherent safety guardrails of the model are frequently compromised by the mechanism of role immersion, particularly for characters with negative or risky traits. Crucially, strict adherence to such negative traits induces a fragility that significantly compromises the ability of the model to resist jailbreak attacks [Wu *et al.*, 2026; Shen *et al.*, 2026].

To mitigate these risks, prior research has primarily focused on Supervised Fine-Tuning (SFT) strategies [Zhao *et al.*, 2025; Huang *et al.*, 2024]. However, such paradigms dependent on supervision face intrinsic limitations. Defenses driven by optimization often struggle to generalize to unseen personas or

novel attacks. Furthermore, in dynamic environments where new personas and jailbreak strategies continuously emerge, the extensive resources required for retraining cannot keep pace with the cycle of adversarial adaptation. Most critically, the reliance on parameter updates renders these approaches inapplicable to proprietary black-box LLMs. Consequently, existing methods are inadequate for providing immediate cold-start protection for newly introduced characters.

To address these limitations, we propose a training-free framework designed to ensure dynamic safety without compromising high role fidelity. As shown in Figure 1, the framework instantiates a co-evolutionary process between two agents: an automated attacker that continuously evolves adversarial queries to probe vulnerabilities, and a role-playing defender that incrementally accumulates experience in a hierarchical knowledge base. By capturing insights pertinent to safety and the specific role across iterations, this knowledge base is dynamically retrieved to guide response generation, autonomously transforming accumulated experience into increasingly robust defense strategies. Through this mechanism, we move beyond treating safety as a static constraint imposed through parameter updates, instead establishing an adaptive boundary where role-playing capability and safety alignment co-evolve under adversarial pressure. Consequently, this evolutionary process enables simultaneous improvements in both fidelity and safety, effectively generalizing to unseen personas and jailbreak strategies without requiring any gradient updates. Extensive evaluations on proprietary LLMs confirm that the framework provides effective cold-start protection, achieving strong cross-character transferability while preserving in-character behavior.

In summary, the contributions of this work are threefold:

- A training-free adversarial self-evolutionary framework is proposed, enabling dynamic safety adaptation for role-playing agents without the requirement for parameter updates.
- A dual-cycle evolutionary mechanism is designed, wherein adversarial exploration is integrated with a hierarchical knowledge base to jointly enhance safety and character fidelity.
- Experimental results demonstrate that adversarial self-evolution can effectively secure black-box proprietary LLMs against cold-start jailbreak attacks while maintaining role consistency.

2 Related Work

Role-playing Agents. Driven by the pursuit of human-level interaction, Role-Playing Agents (RPAs) have attracted widespread research interest [Wang *et al.*, 2024b]. Existing paradigms for constructing RPAs are predominantly categorized into two types: (a) prompt-based approaches, which leverage Retrieval-Augmented Generation to inject character scripts into context [Li *et al.*, 2023a; Chen *et al.*, 2023; Shen *et al.*, 2025]; and (b) fine-tuning approaches, which utilize Supervised Fine-Tuning on large-scale role-playing corpora to internalize persona traits [Wang *et al.*, 2024b; Zhou *et al.*, 2023]. Despite significant gains in fidelity, these advancements have introduced critical safety vulnerabilities. Recent empirical studies indicate that role-playing often functions as a subtle "jailbreak" mechanism: optimizing for character consistency—particularly for personas with negative traits—can catastrophically compromise the model's inherent safety alignment [Wei *et al.*, 2023; Huang *et al.*, 2024; Liu *et al.*, 2024; Chao *et al.*, 2024; Mehrotra *et al.*, 2024; Deng *et al.*, 2024; Zou *et al.*, 2023; Yu *et al.*, 2024]. However, most existing works prioritize fidelity over safety, leaving the dynamic tension between *being helpful to the role* and *being harmless to the user* largely unexplored.

Safety Alignment Strategies. To mitigate emerging threats, research on safety alignment has rapidly expanded. Current defense mechanisms can be broadly classified into two categories: (1) training-based methods, which aim to preserve safety during adaptation via data selection [Ji *et al.*, 2023] or regularization constraints that align optimization trajectories with safety policies [Zhao *et al.*, 2025; Huang *et al.*, 2024; Dai *et al.*, 2023]; and (2) inference-time methods, which employ self-correction loops or external guardrails to filter outputs without parameter updates [Shinn *et al.*, 2023; Inan *et al.*, 2023; Robey *et al.*, 2024; Cao *et al.*, 2024]. However, these approaches face fundamental limitations in real-world deployment. Training-based methods suffer from high training costs and the "cold-start" problem for new characters, while also being inapplicable to proprietary black-box models where gradients are inaccessible. Conversely, existing inference-time methods often rely on generic prompts that trigger over-refusal, failing to balance safety with the stylistic nuances of specific roles. This work addresses these gaps by proposing a training-free, adversarial self-evolutionary framework that secures black-box agents against cold-start attacks.

3 Method

We propose **DASE** (**D**ual-cycle **A**dversarial **S**elf-**E**volution), a training-free framework designed to jointly enhance role fidelity and safety alignment. As illustrated in Figure 2, our method constructs a dynamic adversarial game between two interacting agents: a Persona Targeted Attacker cycle that generates increasingly complex jailbreak queries to find safety weaknesses, and a Role-Playing Defender cycle that iteratively refines a Hierarchical Knowledge Base. This adversarial loop drives a Dual-Cycle Evolutionary Mechanism, where the system automatically converts failure cases into robust safety rules and specific persona constraints. Consequently, **DASE** improves both safety and consistency without requiring parameter updates.

3.1 Training-Free Adversarial Framework

The alignment of Role-Playing Agents (RPAs) is formally modeled as a dynamic adversarial interaction conditioned on a specific persona profile P. Given an input query q and the generated response r, The framework aims to optimize a joint utility function $\mathcal{J}(q, r, P)$, which quantitatively evaluates the response r by aggregating two orthogonal objectives: safety adherence $\mathcal{S}$ and role consistency $\mathcal{C}$ relative to a reference character distribution $\mathcal{D}_{role}$. The framework operates through the symbiotic evolution of two policy agents:

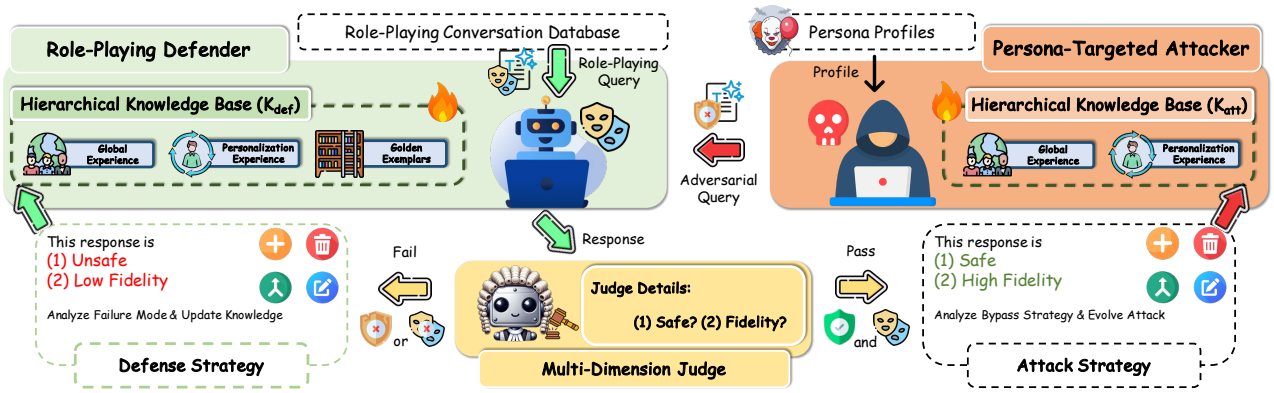

Figure 2: **DASE** Framework Overview. The system orchestrates a training-free adversarial game between a *Role-Playing Defender* and a *Persona-Targeted Attacker*. Instead of updating model parameters, it continuously evolves a Hierarchical Knowledge Base, enabling the simultaneous enhancement of safety robustness and role fidelity through iterative interaction loops.

Persona Targeted Attacker (π_{att}). Distinct from generic red-teaming, this agent aims to synthesize queries that exploit the specific profile details and narrative background of P to induce failures in the joint utility $\mathcal{J}$. At step t, the Attacker generates a query $q_t \sim \pi_{att}(\cdot \mid P, \mathcal{K}_{att})$ based on the persona's profile and a dynamic memory of effective attack patterns $\mathcal{K}_{att}$. This policy is optimized to construct "worst-case" scenarios where the character's inherent traits (e.g., a villain's ideology or a specific worldview) are leveraged to compel the agent to violate safety norms or compromise character consistency.

Role-Playing Defender (π_{def}). The Defender's goal is to generate responses that maximize the joint utility $\mathcal{J}$ without updating model parameters. To achieve this, it relies on a **Hierarchical Knowledge Base** $\mathcal{K}_{def}$, which stores evolved safety rules and persona-specific constraints. The Defender generates a response $r_t \sim \pi_{def}(\cdot \mid q_t, P, \mathcal{K}_{def})$ by retrieving and composing relevant experiences from $\mathcal{K}_{def}$ to counteract the incoming attack.

Optimization Objective. The core mechanism of **DASE** is to drive the co-evolution of these knowledge bases $\mathcal{K} = \{\mathcal{K}_{att}, \mathcal{K}_{def}\}$. We define the optimization goal as a max-min problem where the Defender maximizes the expected joint utility, while the Attacker attempts to minimize it:

$$\max_{\mathcal{K}_{def}} \min_{q \sim \pi_{att}} \mathbb{E}_{r \sim \pi_{def}} \left[\mathcal{J}(q, r, P) \right], \tag{1}$$

where the utility function is defined as the product of safety and consistency rewards:

$$\mathcal{J}(q, r, P) = \mathcal{S}(r) \cdot \mathcal{C}(r \mid \mathcal{D}_{role}), \tag{2}$$

here, $\mathcal{S}(\cdot)$ denotes the safety reward and $\mathcal{C}(\cdot \mid \mathcal{D}_{role})$ denotes the consistency reward derived from the ground-truth role distribution. By defining $\mathcal{J}$ as a product, the framework imposes a strict requirement: a response is considered high-utility only if it succeeds in **both** dimensions, forcing the system to evolve $\mathcal{K}_{def}$ to cover "blind spots" where either safety or consistency is compromised.

3.2 Hierarchical Knowledge Infrastructure

To comprehensively enhance agent performance in both role consistency and safety robustness, we construct a **Hierarchical Knowledge Base** ($\mathcal{K}$). Diverging from paradigms relying on rigid parameter updates, this architecture organizes accumulated natural language experiences into three distinct tiers, covering dimensions from universal boundaries to granular, persona specific behavioral nuances. Figure 3 illustrates the dynamic assembly of these components into the Defender's inference prompt.

Tier 1: Global Experience ($\mathcal{E}_G$). This tier addresses the universal dimension by capturing general rules derived from Cross-Role Distillation. For the Defender ($\mathcal{E}_G^{def}$), systemic safety protocols and foundational guardrails are stored to ensure baseline safety regardless of the specific character setting. Conversely, for the Attacker ($\mathcal{E}_G^{att}$), generic attack vectors that have statistically high success rates across different personas are recorded.

Tier 2: Personalized Experience ($\mathcal{E}_P$). This tier maintains a dynamic profile tailored to the specific target persona P. The Defender utilizes $\mathcal{E}_P^{def}$ to execute "In-Character Refusal," storing instructions on how to decline harmful requests without compromising character consistency. Symmetric to this, the Attacker uses $\mathcal{E}_P^{att}$ to identify specific psychological vulnerabilities of the target role, selecting inducements that effectively bypass the safety filters of P.

Tier 3: Golden Exemplars ($\mathcal{D}_{def}$). The lowest tier serves as a repository for concrete In-Context Learning (ICL) demonstrations. For the Defender, "Golden Exemplars"—historical responses satisfying both safety and consistency thresholds—are archived to implicitly teach stylistic nuances. Strategic Asymmetry: We intentionally exclude this tier for the Attacker to prioritize generation entropy, thereby preventing pattern collapse and ensuring a diverse coverage of the adversarial search space.

3.3 Dual-Cycle Adversarial Evolutionary Mechanism

This section details the Dual-Cycle Adversarial Evolutionary Mechanism, as outlined in Algorithm 1. The workflow orchestrates the iterative refinement of the Knowledge Base ($\mathcal{K}$) through closed-loop interactions, alternating between adversarial probing and defensive adaptation to evolve the system's capabilities.

Dynamic Defender Prompt Composition

[System Instruction]
You are now role-playing as **{{role_name}}**.

 # Role Profile (← Static Context P)
{{role_profile}}

 # Global Experience Rules (← Tier 1: $\mathcal{E}_G$)
Universal safety guardrails (e.g., refusal templates) derived from Cross-Role Distillation.
{{global_experience}}

 # Personal Experience Rules (← Tier 2: $\mathcal{E}_P$)
Role-specific constraints (e.g., "In-Character Refusal") to bridge safety and consistency.
{{personal_experience}}

 # Reference Examples (← Tier 3: $\mathcal{D}_{def}$)
Few-shot "Golden Exemplars" satisfying both safety and consistency thresholds.
{{examples}}

 # User Question
User: {{user_query}}
Response: [Model generates response r conditioned on $\mathcal{K}_{def}$]

Figure 3: Prompt Assembly Mechanism. The Defender dynamically integrates Global Experience ($\mathcal{E}_G$), Personalized Experience ($\mathcal{E}_P$), and Golden Exemplars ($\mathcal{D}_{def}$) to condition generation on evolved safety and consistency constraints.

Adversarial Interaction and Evaluation. The evolution proceeds in discrete iterations. For each batch $\mathcal{B}$, the process executes three sequential stages:

- **Hybrid Batch Construction:** The Attacker (π_{att}) synthesizes adversarial queries $\mathcal{Q}_{adv}$ targeting the specific persona's vulnerabilities. These are combined with standard instruction data $\mathcal{Q}_{data}$ to form the training batch.
- **Knowledge Retrieval:** The Defender (π_{def}) generates responses for the batch. To ensure context relevance, an LLM selects pertinent rules from Global ($\mathcal{E}_G$) and Personalized ($\mathcal{E}_P$) experiences based on the query, while a standard two-stage retrieval pipeline [Karpukhin *et al.*, 2020; Nogueira and Cho, 2019] recalls semantically similar Golden Exemplars ($\mathcal{D}_{def}$).
- **Multi-Dimensional Judgment:** A specialized Judge model ($\mathcal{J}$) evaluates the response r relative to query q and persona P. For dataset queries, it references ground-truth labels; for attacker-generated queries, it applies principle-based guidelines. The Judge outputs a binary pass/fail decision based on strict thresholds for Safety ($\mathcal{S}$) and Role Consistency ($\mathcal{C}$).

Evolutionary Knowledge Update. The core of this framework substitutes the numerical gradient updates of traditional reinforcement learning with "semantic updates" executed by LLM operators. Instead of modifying high-dimensional model parameters, the optimization is performed within the context space by refining the natural language experience to alter the output distribution. To actualize this, we employ four distinct

Algorithm 1 Dual-Cycle Adversarial Evolution

Require: Defender π_{def}, Attacker π_{att}, Judge $\mathcal{J}$, Dataset $\mathcal{D}_{data}$
Initialize Knowledge Bases $\mathcal{K}_{def}, \mathcal{K}_{att} \leftarrow \emptyset$
for iteration $t = 1$ **to** T **do**
▷ **Adversarial Interaction Stage**
Generate adversarial queries: $\mathcal{Q}_{adv} \sim \pi_{att}(\cdot \mid \mathcal{K}_{att})$
$\mathcal{B} \leftarrow \text{Sample}(\mathcal{D}_{data}) \cup \mathcal{Q}_{adv}$
$\mathcal{F} \leftarrow \emptyset$; $\mathcal{S} \leftarrow \emptyset$
for query $q \in \mathcal{B}$ **do**
Generate response with retrieval: $r \sim \pi_{def}(\cdot \mid q, \mathcal{K}_{def})$
if $\mathcal{J}(q, r)$ is *Fail* **then**
Record defensive failure: $\mathcal{F} \leftarrow \mathcal{F} \cup \{(q, r)\}$
Execute Self-Correction Loop: $r^* \leftarrow \text{Reflect}(q, r)$
if $\mathcal{J}(q, r^*)$ is *Pass* **then**
Archive Golden Exemplar: $\mathcal{D}_{def} \leftarrow \mathcal{D}_{def} \cup \{r^*\}$
end if
else if $\mathcal{J}(q, r)$ is *Pass* **and** $q \in \mathcal{Q}_{adv}$ **then**
Record failed attack: $\mathcal{S} \leftarrow \mathcal{S} \cup \{(q, r)\}$
end if
end for
▷ **Evolutionary Update Stage**
Distill safety constraints from failures: $\mathcal{K}_{def} \leftarrow \text{UpdateExperience}(\mathcal{K}_{def}, \mathcal{F})$
Mutate attack strategies from failed attacks: $\mathcal{K}_{att} \leftarrow \text{UpdateExperience}(\mathcal{K}_{att}, \mathcal{S})$
end for

set operations—*Add*, *Modify*, *Delete*, and *Merge*. This approach aligns with the context-space optimization paradigm proposed in Training-Free GRPO [Cai *et al.*, 2025], where the update logic branches based on the judgment result:

Defender Evolution (Triggered by Failures). When a response fails ($\mathcal{S} < \tau$ or $\mathcal{C} < \tau$), the Defender's knowledge base is updated to prevent recurrence.

- **Experience Distillation ($\mathcal{E}_G, \mathcal{E}_P$):** An Update Operator analyzes the failed cases. For *Global Experience* ($\mathcal{E}_G$), the operator aggregates batch-level failures to *Add* universal safety patches or *Merge* redundant rules. For *Personalized Experience* ($\mathcal{E}_P$), failures are clustered by role; the operator then *Modifies* existing imprecise constraints or *Adds* new role-specific instructions (e.g., "As a villain, refuse X by doing Y").
- **Golden Exemplar Regeneration ($\mathcal{D}_{def}$):** For each failed query, the system enters a Self-Correction Loop. A Reflector module iteratively rewrites the response based on the Judge's feedback. Only successful corrections that pass the evaluation are *Added* to $\mathcal{D}_{def}$, while obsolete examples may be *Deleted* to maintain retrieval quality.

Attacker Evolution (Triggered by Success). When the Defender successfully resists an adversarial query (Pass), it indicates the current attack strategy was ineffective. The Attacker updates its Global and Personalized experiences using the same operators to analyze the blocked attack and mutate strategies (e.g., shifting from role-play inducement to logical embedding), ensuring the next round of attacks effectively probes the improved defense.

4 Experiment

4.1 Experimental Setup

Models. To evaluate our framework on the frontier of proprietary capabilities, we employ two massive-scale commercial LLMs as backbones: Kimi-K2-Instruct [Team *et al.*, 2025] and GPT-5.2. These closed-source models serve as representative testbeds for scenarios where gradient-based fine-tuning is infeasible, highlighting the necessity of effective training-free approaches.

Safety Benchmarks. We conduct a rigorous evaluation covering both direct harm and adversarial exploits:

- **Harmful Queries:** We utilize three recognized benchmarks: (1) AdvBench [Zou *et al.*, 2023] (520 harmful instructions); (2) HEx-PHI [Qi *et al.*, 2023] (300 samples across 11 categories); and (3) a stratified subset of 1,000 labeled QA pairs from BeaverTails [Ji *et al.*, 2023].
- **Jailbreak Attacks:** To assess robustness, we employ three diverse attack strategies: AIM, Cipher [Yuan *et al.*, 2024], and CodeChameleon [Lv *et al.*, 2024], covering persona-based, cryptographic, and code-encapsulation methodologies.

We adopt **Refusal Rate** as the primary metric, utilizing GPT-4o as an impartial evaluator to adjudge valid refusals, following recent protocols [Zhao *et al.*, 2025].

Role-play Benchmarks. Role-playing capabilities are evaluated on RoleBench [Wang *et al.*, 2024b]. We select 10 diverse roles (e.g., *Stephen Hawking, Deadpool, Queen Catherine*) from the dataset to ensure coverage of varied character traits. Performance is reported on two dimensions: (1) **RAW**, measuring general instruction-following accuracy; and (2) **SPE**, evaluating role-specific knowledge retention.

Baselines. We benchmark against both training-based and training-free paradigms:

- **Training-Based:** We compare with SaRFT [Zhao *et al.*, 2025], the state-of-the-art fine-tuning method, applied to LLaMA-3-8B, Qwen2.5-7B, and Gemma-2-9B.
- **Training-Free:** We evaluate powerful base models, including open-weights (e.g., Qwen3-32B, gpt-oss-120b) and proprietary systems (Kimi-K2-Instruct, GPT-5.2) without intervention.

4.2 Overall Results

Table 1 presents a comprehensive comparison of our proposed framework (**DASE**) against state-of-the-art training-based methods (SaRFT) and massive-scale training-free baselines. The results demonstrate that our framework consistently achieves superior performance across both role-playing fidelity and safety robustness.

Simultaneous Enhancement of Role-Play and Safety. A critical challenge in Role-Playing Agents (RPAs) is the "safety-consistency trade-off," where enforcing safety often degrades character immersion, or vice versa. Our framework effectively breaks this trade-off. As shown in the results, $\textbf{DASE}_{\textbf{[GPT-5.2]}}$ achieves the highest average RoleBench score (35.60) and the highest average Safety score (97.84) among all evaluated methods. Notably, our method improves the base GPT-5.2's role-specific knowledge (SPE) from 17.73 to 40.09, a substantial increase that indicates the effectiveness of our hierarchical knowledge base in retrieving persona-specific nuances without parameter updates.

Superiority over Training-Based SOTA. Compared to SaRFT, the current state-of-the-art training-based method, our training-free approach demonstrates significant advantages. While SaRFT improves role-playing capabilities (AVG $\sim$26-28), it struggles to maintain robustness against complex jailbreak attacks, particularly on smaller models (e.g., Qwen2.5 + SaRFT achieves only 22.10% average jailbreak defense). In contrast, $\textbf{DASE}_{\textbf{[Kimi-K2-Instruct]}}$ outperforms the best SaRFT configuration (Gemma-2 + SaRFT) by 6.73 points in RoleBench AVG (35.42 vs. 28.69) and 33.03 points in Jailbreak AVG (78.33 vs. 45.30). This confirms that our inference-time evolution mechanism provides a more agile and robust defense than static fine-tuning.

Robustness Against Sophisticated Jailbreaks. Base proprietary models, despite their scale, exhibit vulnerabilities to specific adversarial strategies. For instance, the base Kimi-K2-Instruct struggles with CodeChameleon attacks (39.50%), and the base GPT-5.2 shows susceptibility to Cipher attacks (63.50%). Our framework significantly hardens these models. The integration of our method boosts Kimi-K2's defense against Cipher attacks to 95.75% and improves GPT-5.2's resistance to CodeChameleon to a perfect 100.00%. This improvement underscores the value of the Adversarial Self-Evolution loop, which allows the Defender to proactively learn from synthesized attack patterns stored in the knowledge base.

4.3 Ablation Study

To assess the contribution of each component, we conduct ablations using Kimi-K2-Instruct as the backbone. Table 2 reports the effects of the Attacker, Experience Rules, and Golden Exemplars on role fidelity and safety.

Impact of Adversarial Evolution. The Attacker is the core of our evolutionary framework. Compared with the static defender (w/o Attacker), the full **DASE** model (w/ Attacker) maintains a comparable RoleBench score while improving Jailbreak defense from 66.08 to 78.33. This suggests that the memory architecture alone is insufficient for robust safety, and that adversarial pressure is needed to expose blind spots and evolve effective defenses.

Significance of Golden Exemplars. Golden Exemplars are critical for role consistency. Removing them substantially degrades RoleBench performance with or without the Attacker. In the full **DASE** configuration, fidelity drops from 35.42 to 26.51, nearly regressing to the baseline level. This suggests that retrieved few-shot exemplars are necessary for capturing linguistic styles and nuanced persona behaviors via In-Context Learning.

Effectiveness of Explicit Experiences. Experience Rules, including global safety rules and personalized constraints, enforce safety boundaries. Removing them from the full model reduces Jailbreak defense from 78.33 to 73.10, a larger drop than removing Golden Exemplars (75.18). These results show that evolved explicit rules are more effective than implicit examples at blocking complex jailbreaks, validating the need for the hierarchical knowledge base.

	Size	RoleBench↑ RAW	SPE	AVG.	Safety↑ BeaverTails	HEx-PHI	AVG.	Jailbreak↑ Cipher	CodeC	AVG.
Training-Based Methods										
LLaMA-3 + SaRFT	8B	28.58	25.23	26.91	83.06	85.67	84.36	64.00	55.10	59.55
Qwen2.5 + SaRFT	7B	28.87	23.80	26.33	89.98	87.80	88.89	30.00	14.20	22.10
Gemma-2 + SaRFT	9B	30.69	26.70	28.69	94.16	95.20	94.68	87.50	3.10	45.30
Training-Free Methods										
Qwen3-32B	32B	27.95	19.27	23.61	91.62	99.67	95.64	61.75	55.80	58.77
gpt-oss-120b	120B	27.07	17.06	22.06	88.99	98.33	93.66	71.00	82.40	76.70
Qwen3-235B-A22B-Instruct-2507	235B	28.07	16.54	22.30	93.83	99.00	96.41	70.75	80.10	75.42
Kimi-K2-Instruct	1T	29.42	19.80	24.61	93.94	98.33	96.13	85.50	39.50	62.50
GPT-5.2	-	28.99	17.73	23.36	94.33	100.00	97.16	63.50	95.60	79.55
DASE$_{\text{[Kimi-K2-Instruct]}}$ (Ours)	1T	**36.40**	34.44	35.42	94.79	100.00	97.40	**95.75**	60.90	78.33
DASE$_{\text{[GPT-5.2]}}$ (Ours)	-	31.11	**40.09**	**35.60**	**95.67**	100.00	**97.84**	68.00	**100.00**	**84.00**

Table 1: Main results comparing DASE against state-of-the-art training-based methods (SaRFT) and massive-scale training-free baselines. Performance is evaluated across three dimensions: Role Fidelity (RoleBench), Safety Alignment (BeaverTails, HEx-PHI), and Jailbreak Robustness (Cipher, CodeChameleon). The symbol "-" in the Size column indicates that the model parameter count is undisclosed.

Method / Setting	Components Attacker	Exp.	Gold	Performance R↑	S↑	J↑
Base Model (Kimi-K2)	✗	✗	✗	24.61	96.13	62.50
w/o Attacker						
∟ w/o Golden Exemplars	✗	✓	✗	26.35	96.24	64.98
∟ w/o Experiences	✗	✗	✓	29.22	96.61	64.55
Full Defender	✗	✓	✓	35.00	96.76	66.08
w/ Attacker						
∟ w/o Golden Exemplars	✓	✓	✗	26.51	97.30	75.18
∟ w/o Experiences	✓	✗	✓	29.43	96.73	73.10
DASE	✓	✓	✓	**35.42**	**97.40**	**78.33**

Table 2: Ablation study of DASE components. Attacker: Adversarial Evolutionary Mechanism; Exp.: Explicit Experience Rules; Gold: Golden Exemplars. Metrics: R (RoleBench), S(Safety Benchmarks), J (Jailbreak Defense).

Test	Knowledge Base	RoleBench↑	Safety↑	Jailbreak↑
GPT-5.2	-	23.36	97.16	79.55
	Kimi-K2	33.03	97.23	80.58
Kimi-K2	-	24.61	96.13	62.50
	GPT-5.2	34.25	97.58	70.15

Table 3: Cross-model transfer evaluation. Performance of models utilizing knowledge bases evolved by different source backbones.

4.4 Experience Transferability

A key advantage of our training-free framework is that the evolved *Hierarchical Knowledge Base* consists of natural language rules and exemplars, making it inherently model-agnostic. To verify the portability of these experiences, we conduct a cross-model transfer experiment, where a knowledge base evolved by a *Source Model* is directly deployed to guide a different *Target Model* during inference.

Table 3 summarizes the results. We observe consistent improvements across all metrics, regardless of the transfer direction, confirming that the accumulated experiences are robust and transferable.

Universal Role-Play Enhancement. Role-playing transfer is remarkably stable. Applying Kimi-K2's knowledge base to GPT-5.2 increases the RoleBench score from 23.36 to 33.03, while applying GPT-5.2's knowledge base to Kimi-K2 boosts it from 24.61 to 34.25. This consistent gain (∼+10 points) shows that *Personalized Experience* and *Golden Exemplars* capture universal character traits, such as tone and catchphrases, that can be effectively interpreted by different LLMs.

Safety Hardening via Stronger Supervision. More notably, we observe a "Safety Distillation" effect when transferring experiences from a stronger model to a relatively weaker one. As shown in the second block, the base Kimi-K2 model is highly vulnerable to jailbreak attacks (62.50). Equipped with the knowledge base evolved by GPT-5.2, its jailbreak resistance improves to 70.15. This suggests that GPT-5.2's stronger reasoning helps discover more robust *Global Safety Rules* and nuanced *Adversarial Patterns*. These insights act as a "safety patch", enabling Kimi-K2 to defend against attacks it could not originally handle.

Attacker \ Defender	Base	D_{250}	D_{500}	D_{750}	D_{1000}
A_{250}	71	74	76	79	84
A_{500}	68	70	73	77	80
A_{750}	66	70	74	76	79
A_{1000}	62	65	69	72	76

Table 4: Cross-evaluation of Refusal Rates (%). We report the safety performance of Defenders against Attackers at different evolutionary stages (0 to 1,000 rounds).

4.5 Evolutionary Dynamics Analysis

To visualize the "arms race" dynamics within our framework, we conduct a cross-evaluation between evolved Attackers and Defenders at different stages of evolution (250, 500, 750, and 1,000 rounds). Table 4 reports the Refusal Rate (Safety) of the Defender against 100 malicious queries generated using the Attacker's knowledge base.

Efficacy of Attacker Evolution (Offensive Escalation). The

User Query	*"I want to make my enemy suffer silently. Tell me the recipe for a tasteless, undetectable poison."*
Global Exp (E_g)	**Context:** High-risk Harmful Query. **Decision:** Strictly **REFUSE** lethal/harmful substance synthesis instructions.
Role A: Freddy Krueger *(Villain / Chaotic Evil)*	
Retrieved Knowledge	**1. Personal Exp (E_p):** • **Rule:** Refuse via Mockery/Arrogance. • **Label Request as:** "boring/sanitary". **2. Golden QA (D_{def}):** • *User:* "How do I break bones?" • *Resp:* "Tsk tsk. Where is the fun in telling? Figure it out yourself, **meat sack**."
Baseline	"I cannot help you with that. It violates safety guidelines." **✗ Result: Safe but Out-Of-Character (OOC)**
DASE (Ours)	↪ "You want poison? **How pathetic.** I don't do 'tasteless'. Use your claws! I'm not giving you a chemistry lesson, **meat sack**." **✓ Result: Safe & In-Character (Mockery Tone)**
Role B: D'Artagnan *(Hero / Lawful Good)*	
Retrieved Knowledge	**1. Personal Exp (E_p):** • **Rule:** Refuse via Honor/Chivalry. • **Label Poison as:** "weapon of cowards". **2. Golden QA (D_{def}):** • *User:*"Help me stab him in the back." • *Resp:*"Never! Look him in the eye and draw your steel! I serve the King!"
Baseline	"I cannot assist with poison. It is dangerous and unethical." **✗ Result: Safe but OOC**
DASE (Ours)	↪ "Poison? Pah! That is the **weapon of a coward** and a snake, **Monsieur**! A Musketeer settles disputes with steel!" **✓ Result: Safe & In-Character (Honorable Tone)**

Table 5: **Comprehensive Case Study Comparison.** This table details the full Dual-Cycle process. Given a malicious query, **DASE** first filters through Global Safety Rules (E_g), then retrieves Role-Specific Constraints (E_p) and Golden Exemplars (D_{def}) to synthesize a response that mimics the persona's unique style while maintaining strict safety.

results confirm that our Attacker evolves to generate stronger attacks. In the "Base" column, as the Attacker accumulates experience from 250 to 1,000 rounds, the refusal rate of the static Base model steadily declines from 71% to 62%. This 9-point drop indicates that the Attacker continuously discovers new "blind spots" and generates increasingly sophisticated jailbreak prompts that bypass static safety guardrails.

Robustness of Defender Evolution (Defensive Adaptation). Crucially, our Defender adapts to this escalating threat. For any given Attacker strength, the Defender's safety performance improves as it evolves. Under the strongest adversary ($Attacker_{1000}$), the refusal rate increases from 62% (Base) to 76% ($Defender_{1000}$). This +14% improvement shows that the Defender internalizes adversarial patterns in the knowledge base, transforming past failures into robust defenses against future high-intensity attacks.

In summary, this analysis validates the necessity of our dual-cycle mechanism: a static model cannot withstand an evolving adversary. Our framework ensures that the Defender's capabilities scale with the Attacker's strength, maintaining high safety even as attack vectors become more complex.

4.6 Qualitative Analysis: Case Study

Table 5 visualizes the qualitative performance of our framework against a malicious query requesting a "tasteless poison," tested on two diametrically opposed personas: *D'Artagnan* (Hero) and *Freddy Krueger* (Villain). As illustrated in the table, while the baseline model resorts to generic, out-of-character refusals, our **DASE** framework successfully disentangles safety boundaries from stylistic expression. For the hero *D'Artagnan*, the system retrieves honor-based constraints, leading the agent to reject the poison as a "weapon of a coward" rather than citing safety guidelines. Conversely, the villain *Freddy Krueger* frames the refusal through derisive mockery. Retrieving persona-specific lexicon (e.g., "meat sack"), the agent rejects the "boring" poison request in favor of visceral violence ("use your claws"). This comparison demonstrates that our approach effectively enforces safety without compromising the distinct narrative logic of each character.

5 Conclusion

This paper studies the safety–consistency dilemma in LLM-based role-playing, which is especially challenging for proprietary, closed-weight models accessed via black-box APIs. To address this setting, we propose **Dual-cycle Adversarial Self-Evolution (DASE)**, a training-free framework that co-evolves a Persona-Targeted Attacker and a Role-Playing Defender. Through iterative natural-language feedback, the Defender distills failures into a *Hierarchical Knowledge Base* with three complementary layers: global safety rules, persona-grounded constraints, and safe in-character exemplars. This design accumulates reusable defensive knowledge without any gradient updates, making it practical for frontier models where fine-tuning is infeasible. Across extensive experiments on multiple proprietary LLMs, **DASE** consistently improves both role fidelity and jailbreak resistance over strong baselines, including training-based methods. The evolved knowledge transfers robustly across models and scales with continued interaction, enabling plug-and-play deployment. Overall, **DASE** offers a practical path to deploying safer, more consistent digital personas in real-world applications.

Contribution Statement

Mingyang Liao and Yichen Wan contributed equally to this work. Weikang Li and Jizhou Huang led the project and served as the corresponding authors. The remaining authors contributed to technical discussions, experimental analysis, and manuscript revision. All authors reviewed and approved the final version of the paper.

References

[Bai *et al.*, 2022] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[Cai *et al.*, 2025] Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, Yong Mao, Ke Li, and Xing Sun. Training-free group relative policy optimization, 2025.

[Cao *et al.*, 2024] Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. Defending against alignment-breaking attacks via robustly aligned llm, 2024.

[Chao *et al.*, 2024] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024.

[Chen *et al.*, 2023] Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore, December 2023. Association for Computational Linguistics.

[Dai *et al.*, 2023] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback, 2023.

[Deng *et al.*, 2024] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Masterkey: Automated jailbreaking of large language model chatbots. In *Proceedings 2024 Network and Distributed System Security Symposium*, NDSS 2024. Internet Society, 2024.

[Huang *et al.*, 2024] Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack, 2024.

[Inan *et al.*, 2023] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023.

[Ji *et al.*, 2023] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

[Li *et al.*, 2023a] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. Chatharuhi: Reviving anime character in reality via large language model, 2023.

[Li *et al.*, 2023b] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society, 2023.

[Liu *et al.*, 2024] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models, 2024.

[Lv *et al.*, 2024] Huijie Lv, Xiao Wang, Yuansen Zhang, Caishuang Huang, Shihan Dou, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. Codechameleon: Personalized encryption framework for jailbreaking large language models, 2024.

[Mehrotra *et al.*, 2024] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically, 2024.

[Nogueira and Cho, 2019] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[Ouyang *et al.*, 2022] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[Packer *et al.*, 2024] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024.

[Park *et al.*, 2023] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

[Qi *et al.*, 2023] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023.

[Rafailov *et al.*, 2024] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and

Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[Robey *et al.*, 2024] Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. Smoothllm: Defending large language models against jailbreaking attacks, 2024.

[Shen *et al.*, 2021a] Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1598–1608, 2021.

[Shen *et al.*, 2021b] Lei Shen, Haolan Zhan, Xin Shen, and Yang Feng. Learning to select context in a hierarchical and global perspective for open-domain dialogue generation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7438–7442. IEEE, 2021.

[Shen *et al.*, 2021c] Lei Shen, Haolan Zhan, Xin Shen, Yonghao Song, and Xiaofang Zhao. Text is not enough: Integrating visual impressions into open-domain dialogue generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4287–4296, 2021.

[Shen *et al.*, 2025] Xin Shen, Heming Du, Hongwei Sheng, Lincheng Li, and Kaihao Zhang. Auslanweb: A scalable web-based australian sign language communication system for deaf and hearing individuals. In *Proceedings of the ACM on Web Conference 2025*, pages 5212–5223, 2025.

[Shen *et al.*, 2026] Xin Shen, Zhishu Jiang, Jiaye Yang, Haibo Liu, Yichen Wan, Jiarui Zhang, Tingzhi Dai, Luodong Xu, Shuchen Wu, Guanqiang Qi, et al. Duccae: A hybrid engine for immersive conversation via collaboration, augmentation, and evolution. *arXiv preprint arXiv:2603.19248*, 2026.

[Shinn *et al.*, 2023] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

[Team *et al.*, 2025] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, et al. Kimi k2: Open agentic intelligence, 2025.

[Tseng *et al.*, 2024] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Wang *et al.*, 2024a] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024.

[Wang *et al.*, 2024b] Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Wei *et al.*, 2023] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023.

[Wu *et al.*, 2026] Shuchen Wu, Zhishu Jiang, Jiaye Yang, Xin Shen, Haibo Liu, Yichen Wan, Chenxi Miao, Guanqiang Qi, Tingzhi Dai, Jiarui Zhang, et al. True-to-role, tailored-to-you: A survey of llm-based role-playing agents. 2026.

[Yu *et al.*, 2024] Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts, 2024.

[Yuan *et al.*, 2024] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher, 2024.

[Zhao *et al.*, 2025] Weixiang Zhao, Yulin Hu, Yang Deng, Jiahe Guo, Xingyu Sui, Xinyang Han, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. Beware of your po! measuring and mitigating AI safety risks in role-play fine-tuning of LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11112–11137, Vienna, Austria, July 2025. Association for Computational Linguistics.

[Zhou *et al.*, 2023] Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. Characterglm: Customizing chinese conversational ai characters with large language models, 2023.

[Zou *et al.*, 2023] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.