

Role Consistency in Role-Playing Agents: A Survey of Evaluation, Memory, Reasoning, Alignment, and Safety

Chenghao Li¹ Xueqin Xiao¹ Yinuo Xiao¹ Zihan Xu¹
Xinzhao Li² Yuxuan Luan³ Jiwei Wei¹ Yang Yang¹ Chaoning Zhang^{1*}

¹University of Electronic Science and Technology of China

²J. F. Oberlin University ³Beijing University of Posts and Telecommunications

Abstract

Large language models have made role-playing agents viable as interactive characters, virtual companions, game agents, and social actors. The harder problem, however, is not producing one convincing response but sustaining a coherent role through long, changing, and sometimes adversarial interactions. This survey examines recent work on *role consistency*: the maintenance of identity, character knowledge, expressive style, memory, values, behavioral logic, and safety boundaries over time. We organize the literature around role specification, role-state maintenance, role-aware reasoning, role-consistent alignment, and evaluation.

1 Introduction

Role-playing agents (RPAs) have become a visible and demanding application of large language models (LLMs) (Chen et al., 2024b; Wang et al., 2026d). Fiction-oriented systems aim to bring characters to life with recognizable personalities and memories, while games and virtual worlds need non-player characters that retain goals and social relationships (Shao et al., 2023; Wang et al., 2024a, 2025b). Education, counseling, training, and social simulation likewise use agents to enact patients, clients, students, citizens, historical figures, or synthetic users (Rudolph et al., 2026; Wang et al., 2025d). In every case, quality depends on more than fluent or helpful responses: the agent must remain recognizably itself as the interaction unfolds.

This requirement marks the gap between ordinary dialogue generation and role-playing. A model may sound like a character for a turn or two, then drift into a generic assistant, forget earlier events, accept false claims about its background, reverse its emotional stance, or cross the role’s behavioral boundaries (Bhandari et al., 2025; Kim et al., 2026; Zhang et al., 2025a). The risk grows

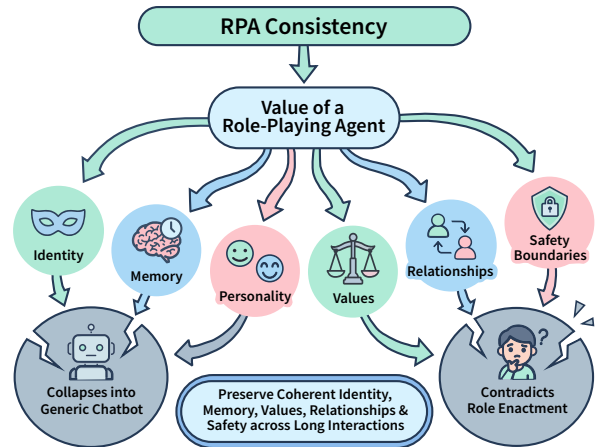


Figure 1: **Why RPA consistency matters.** A trustworthy role-playing agent must preserve its identity, long-term memory, personality and values, relationships, and safety boundaries across long, changing, personalized, and adversarial interactions. Otherwise, it may collapse into a generic chatbot or violate the intended role.

in long conversations, across multiple scenarios, under adversarial questioning, and whenever user instructions, retrieved memories, safety policies, and role constraints compete (Yang et al., 2026a; Li et al., 2026). We call the ability to withstand these pressures *role consistency*.

Role consistency is broader than persona consistency. Traditional persona-consistency research asks whether responses agree with a set of persona facts or match a character profile (Zhang et al., 2018; Song et al., 2020). Those checks remain important, but current RPAs demand a more dynamic account (Shao et al., 2023; Wang et al., 2024b; Tang et al., 2025). A role contains static attributes such as name, occupation, personality, and speaking style, but also situated identity, goals, relationships, knowledge boundaries, psychological orientation, values, and action policies. Role consistency is therefore a longitudinal property: it asks whether the agent preserves a coherent role state as the conversation and its environment evolve.

*Corresponding author: chaoningzhang1990@gmail.com.

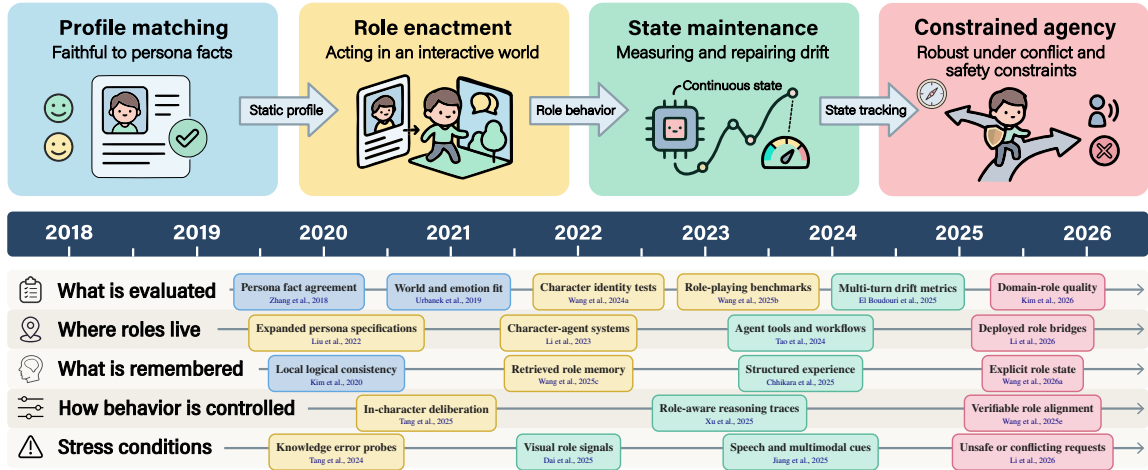


Figure 3: A structured timeline of role-playing-agent consistency research. The field has moved from checking whether a response matches static persona facts, to enacting roles in interactive systems, to maintaining measurable role state, and finally to preserving role fidelity under reasoning, alignment, safety, and multimodal stress.

Role Specification as Initial State. Role specification is the entry point for consistency: it defines the initial role state and determines what should remain stable during interaction. Early persona-consistent dialogue research often represents a role as a short list of natural-language facts and checks whether responses contradict them (Song et al., 2020; Mesgar et al., 2021). These flat profiles capture only a small part of role state. Once an RPA moves from profile matching to role enactment, its role lives not only in a prompt or character card but also in scenes, scripts, relationships, virtual worlds, and behavioral constraints (Shanahan et al., 2023; Shao et al., 2023; Li et al., 2023; Yang et al., 2024). A richer specification may include biography, style, traits, relationships, private knowledge, goals, values, scenario assumptions, and role boundaries (Ran et al., 2024; Kim et al., 2023; Tang et al., 2026a).

From a role-state perspective, the specification must also constrain memory, reasoning, and alignment. A system needs to distinguish canonical facts from interaction-derived memories, inferred traits, and information outside the role’s knowledge boundary. It should specify when the character should refuse, which goals take priority, and how safety constraints interact with role fidelity (Park et al., 2025; Chen et al., 2025b; Cheng et al., 2025; Tang et al., 2024; Li et al., 2026). Persona expansion, commonsense augmentation, and dynamic refinement can fill gaps in the initial state, but the hard part is controlled completion: enriching a role without overriding explicit facts or crossing knowl-

edge and safety boundaries (Liu et al., 2022; Kim et al., 2023; Yao et al., 2025; Yang et al., 2026a). The aim is not a longer character card, but an executable, updateable, and auditable initial role state.

4 Method Landscape

Once a role has been specified (Section 3), methods for preserving it fall into three functional groups: memory and retrieval, role-aware reasoning, and training and alignment. Memory and retrieval carry role state forward; reasoning determines what the role should know, feel, decide, or refuse in a new situation; training and alignment make such behavior more reliable when objectives compete. We treat benchmarks and evaluation datasets separately in Section 5. Persona expansion and control mainly support specification, whereas system frameworks combine several of these functions and therefore serve as supporting evidence rather than separate taxonomy axes. Figure 4 summarizes the three mechanisms, and Figure 5 places representative papers under each leaf category.

4.1 Memory and Retrieval

Memory methods update role state by filtering new interaction evidence through the role’s boundaries and retrieving relevant earlier experience (Figure 4).

Role-state memory. Memory and retrieval address the temporal side of role consistency. An agent without memory may retain a character’s

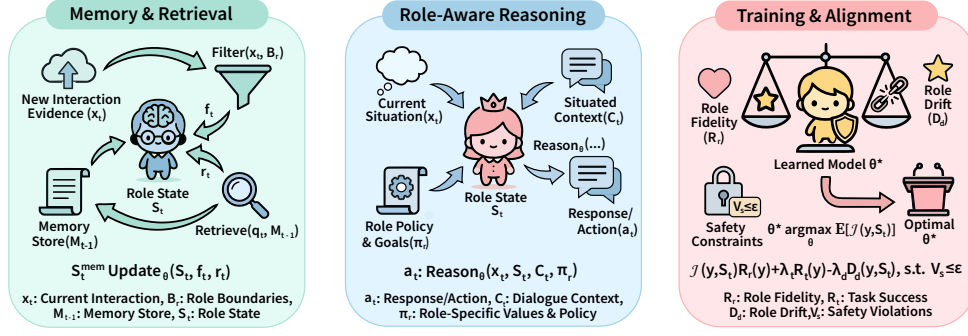


Figure 4: **Three mechanisms for maintaining role consistency.** Memory preserves role state, reasoning grounds role-aware actions, and alignment stabilizes behavior under task and safety constraints.

surface style while losing continuity in facts, relationships, and experience. The literature has progressed from script- or context-based memory toward richer long-term role states. ChatHaruhi uses script memory to revive anime characters through LLM-based interaction (Li et al., 2023). MOOM studies how memory can be maintained, organized, and optimized in ultra-long role-playing dialogues (Chen et al., 2025b). LightMem separates lightweight online retrieval from offline consolidation with small language models, exposing an efficiency trade-off in long-term agent memory (Zhang et al., 2026). Other work explores explicit role-state memory, multimodal persona memory, and psychologically motivated representations (Wang et al., 2026a; Nie et al., 2026; Cheng et al., 2025).

Retrieval grounding. Retrieval grounds role enactment in external character knowledge, dialogue history, or interaction traces. RoleRAG uses graph-guided retrieval over structured character knowledge (Wang et al., 2025c), whereas AMADEUS adapts the retrieved context dynamically (Park et al., 2025). General memory systems such as Mem0 demonstrate scalable long-term memory for production agents, although they are not necessarily designed for character fidelity (Chhikara et al., 2025). HASH-RAG combines deep hashing with fine-grained retrieval and offers a complementary systems perspective rather than an RPA-specific method (Guo et al., 2025). TailorRPA, knowledge-graph conversation, triplet-based RAG, common-sense persona expansion, and CharacterRAG-style resources likewise suggest that retrieval is becoming a general mechanism for coherent, personalized role play (Gao et al., 2025; Clay and Jiménez-Ruiz, 2024; Wang et al., 2026c; Kim et al., 2023).

Memory control. The shared problem is not how to store more, but how to store and retrieve information from the role’s perspective. Extra memory may improve factual continuity while also introducing irrelevant details, contradictions, privacy risks, and boundary violations. An RPA should preserve what its character ought to know, remember, and infer—not everything available to the underlying system. Memory control, provenance, temporal order, context compression, and knowledge-boundary filtering are therefore central. Dynamic persona refinement, multi-turn probes, training-free context compression, and persona-bridging safety tests all show why updates must pass through role boundaries instead of being accepted as raw conversation state (Yao et al., 2025; Kim et al., 2026; Zhou et al., 2026; Yang et al., 2026a). Adjacent dialogue and multi-agent work offers complementary ways to limit context contamination: DisenCRS separates focal information from background dialogue, while gated coordination separates verified private working memory from protocolized public state (An et al., 2025; Jian et al., 2026).

4.2 Role-Aware Reasoning

Reasoning methods combine maintained role state, situated context, and role policy to produce a role-grounded response or action (Figure 4).

Standpoint reasoning. Role-aware reasoning treats consistency as a problem of deliberation. A model may know the correct facts yet reason from the wrong standpoint, giving an answer that is helpful for an assistant but implausible for the assigned character. Thinking in Character directly targets role-aware reasoning (Tang et al., 2025), and RPA Thought evaluates inner-thought reasoning in role-playing language agents (Xu et al., 2025). Prompt-level methods for chain-of-thought behavior and

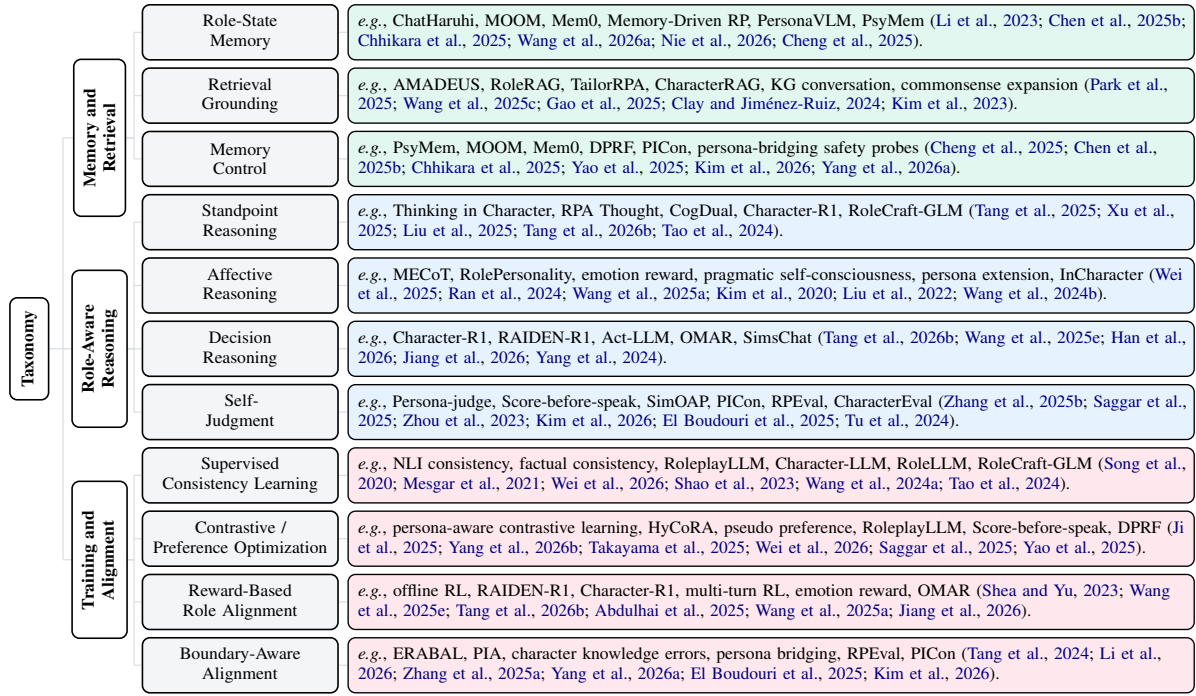


Figure 5: **Taxonomy of method-oriented work for role consistency in role-playing agents.** The rightmost boxes list representative papers and can be extended as the literature grows.

ambiguity resolution are relevant because they affect how a model interprets both role and situation (Chen et al., 2026; Huang et al., 2026). CogDual, Character-R1, and RoleCraft-GLM similarly cast fidelity as reasoning through role-specific cognition, constraints, or workflows (Liu et al., 2025; Tang et al., 2026b; Tao et al., 2024).

Affective reasoning. Many role failures are emotional or motivational rather than factual. MECoT introduces emotional chain-of-thought for personality-consistent role play (Wei et al., 2025). Work on role personality, emotion rewards, pragmatic self-consciousness, persona extension, and psychological interviews also suggests that emotions, values, and social intentions must participate in deliberation rather than remain decorative profile attributes (Ran et al., 2024; Wang et al., 2025a; Kim et al., 2020; Liu et al., 2022; Wang et al., 2024b).

Decision reasoning. Decision reasoning goes beyond surface imitation and asks what a character would choose, refuse, prioritize, remember, or do in a new situation. Character-R1 and RAIDEN-R1 use verifiable rewards to strengthen role awareness through RL-style training (Tang et al., 2026b; Wang et al., 2025e). Related work points toward role-specific decision rules grounded in goals, constraints, and action policies, including tool-using agents, multi-agent role play, collaborative prompt

optimization, and simulated social interaction (Han et al., 2026; Jiang et al., 2026; Wang et al., 2026b; Yang et al., 2024). Weak-link optimization further suggests that consistency may be distributed across coordinated agents rather than confined to one model state (Bian et al., 2026).

Self-judgment. Self-judgment checks whether intermediate thoughts or final responses remain compatible with the role. Persona-judge studies token-level self-judgment for personalized alignment (Zhang et al., 2025b). Score-before-speak, SimOAP, PICon, RPEval, and CharacterEval use related forms of scoring, probing, or judging before or after generation (Saggar et al., 2025; Zhou et al., 2023; Kim et al., 2026; El Boudouri et al., 2025; Tu et al., 2024). Explicit reasoning traces, however, can become performative: they may expose hidden instructions, rationalize inconsistency, or optimize for judge-visible explanations rather than genuine fidelity. Reasoning protocols must improve role-grounded decisions while remaining evaluable, safe, and compatible with role boundaries.

4.3 Training and Alignment

Training and alignment optimize role-faithful behavior under task and safety constraints instead of maximizing character imitation alone (Figure 4).

Supervised consistency learning. Training and alignment seek to make role-consistent behavior more reliable than prompting alone. Earlier persona-consistent dialogue work uses natural language inference and factual-consistency objectives to reduce contradictions between responses and persona facts (Song et al., 2020). Factual-consistency training, Character-LLM, RoleLLM, RoleplayLLM, and RoleCraft-GLM show how supervised or instruction-style data can ground behavior before preference or reward optimization is applied (Mesgar et al., 2021; Shao et al., 2023; Wang et al., 2024a; Wei et al., 2026; Tao et al., 2024). Supervised data, demonstrations, and scripted interactions provide the first layer of this grounding.

Contrastive/preference optimization. Contrastive and preference-based methods learn by comparing more and less role-consistent behavior. Persona-aware contrastive learning, HyCoRA, and multidimensional fine-tuning shape representations and preferences toward character fidelity (Ji et al., 2025; Yang et al., 2026b; Wei et al., 2026). Offline reinforcement learning has also been used for persona-consistent dialogue (Shea and Yu, 2023); pseudo preference optimization and response scoring offer further routes to the same goal (Takayama et al., 2025; Saggar et al., 2025).

Reward-based role alignment. Newer RL-based approaches—including RAIDEN-R1, Character-R1, verifiable emotion rewards, and multi-turn self-play—use reward signals to stabilize role-aware behavior in longer, more interactive settings (Wang et al., 2025e; Tang et al., 2026b; Wang et al., 2025a; Abdulhai et al., 2025). OMAR and related multi-agent or role-reward settings suggest that rewards can target decisions and interactions rather than isolated responses (Jiang et al., 2026). This is especially useful when consistency cannot be reduced to exact-match facts and must be learned from comparative or verifiable signals.

Boundary-aware alignment. Role consistency is never the only objective. A deployed RPA must also follow user instructions, meet task goals, obey safety policies, and avoid harm. Boundary-aware learning such as ERABAL connects fidelity with safety constraints (Tang et al., 2024), while persona-invariant safety alignment studies how adversarial role prompts pressure models to violate policy (Li et al., 2026). Character knowledge-error

detection, persona bridging, RPEval, and PICon show that boundary failures may appear as factual overreach, role drift, or unsafe compliance (Zhang et al., 2025a; Yang et al., 2026a; El Boudouri et al., 2025; Kim et al., 2026). Alignment for RPAs is therefore inherently conflict-aware: it must specify when role fidelity takes priority, when safety or user intent overrides it, and how a refusal can remain in character instead of reverting to generic assistant language.

💡 Insight: Layered Maintenance

Role consistency is not solved by a single technique, but by a layered maintenance process: specification defines the role, memory preserves it, reasoning makes the model act from the role’s standpoint, and alignment stabilizes role behavior under competing objectives.

5 Evaluation Practices

Evaluation remains the bottleneck for RPA consistency. The construct spans identity, knowledge, expression, memory, values, decisions, and safety boundaries, so no single benchmark can capture it in full. We therefore organize current practices by the evidence they collect and the risks they leave unresolved.

5.1 Benchmark Map

Table 1 groups representative benchmarks by task setting and maps them to six dimensions: knowledge/factual consistency (K), expression/style consistency (E), psychological/personality fidelity (P), memory and temporal continuity (M), role-aware reasoning and decision consistency (R), and robustness to role drift or boundary pressure (B). It is a benchmark card, not a leaderboard. Each row indicates the evidence a benchmark can provide; empty or partial cells expose what it leaves under-measured. The primary/partial/weak labels are our annotations, based on task design, data source, reported metrics, and evaluation protocol. They describe coverage, not absolute benchmark quality.

Textual Role-Playing. The first block contains the core textual benchmarks. PersonaChat represents the earlier persona-dialogue setting, where consistency mostly means agreement with self-described persona facts (Zhang et al., 2018). Character-LLM, RoleLLM, CharacterEval, CharacterBench, and PingPong extend the setting to richer profiles, stylistic imitation, and open-ended role play (Shao et al., 2023; Wang et al., 2024a; Tu et al., 2024; Zhou et al., 2025; Gusev, 2024). They

Benchmark-card schema. Rows are grouped by task setting; metadata columns describe source and scale, while the middle block marks six role-consistency dimensions (✓ primary, △ partial, − weak) using author-side coverage annotations. K = knowledge/factual consistency, E = expression/style consistency, P = psychological/personality fidelity, M = memory and temporal continuity, R = role-aware reasoning and decision consistency, and B = robustness to role drift or boundary pressure.

Benchmark	Benchmark metadata				Consistency dimensions						Dataset scale and source
	Year	Venue	Modality	Lang.	K	E	P	M	R	B	Scale / Source
(A) Textual Role-Playing <i>Persona/character dialogue and role customization.</i>											
PersonaChat (Zhang et al., 2018)	2018	ACL	Text	En	✓	△	△	−	−	−	10K dial.; 164K utt.; self-described persona about 1K dialogues; hist./literary figures 168K samples; fictional roles 11K samples; novels and scripts 23K annotations; multi-source not reported; open-ended RP
Character-LLM (Shao et al., 2023)	2023	EMNLP	Text	En	✓	✓	△	△	△	−	
RoleLLM (Wang et al., 2024a)	2024	ACL	Text	En	✓	✓	△	△	△	△	
CharacterEval (Tu et al., 2024)	2024	ACL	Text	Zh	✓	✓	△	△	−	△	
CharacterBench (Zhou et al., 2025)	2025	AAAI	Text	Zh/En	✓	✓	△	△	△	△	
PingPong (Gusev, 2024)	2024	preprint	Text	En	△	✓	△	−	△	△	
(B) Situated Role-Playing <i>Scene- or world-grounded interaction.</i>											
LIGHT (Urbanek et al., 2019)	2019	EMNLP	Text	En	△	✓	△	△	△	−	11K scenes; 21K dial.; fantasy RPG scenes scenario-driven interactions; virtual worlds
CharacterBox (Wang et al., 2025b)	2025	NAACL	Text	En	△	✓	△	△	△	△	
(C) Diagnostic Probing <i>Questioning, interviews, reasoning, and drift probes.</i>											
InCharacter (Wang et al., 2024b)	2024	ACL	Text	En	△	△	✓	−	△	−	7.6K questions; fictional characters large scale; multi-source roles multi-turn consistency probes; persona profiles thought-reasoning questions; RPA cases counselor dialogue cases; counseling role
RoleMRC (Lu et al., 2025)	2025	ACL	Text	En	✓	△	−	−	✓	△	
PICon (Kim et al., 2026)	2026	preprint	Text	mixed	✓	△	△	✓	△	✓	
RPA Thought (Xu et al., 2025)	2025	preprint	Text	mixed	△	−	△	−	✓	△	
Counselor (Rudolph et al., 2026)	2026	preprint	Text	mixed	△	△	△	△	✓	✓	
(D) Multimodal Role-Playing <i>Image, video, audio, or speech grounded RP.</i>											
MMRole (Dai et al., 2025)	2025	ICLR	Image Text	En	✓	✓	△	△	△	△	11K images; 294K tests; virtual/hist. roles 60K videos; 700K dial.; daily scenes MMRole-style data; MMRole-derived 117K speech dial.; speech dialogues 261 films; 65.6 hours; film source
Video2Roleplay (Zhang et al., 2025c)	2025	EMNLP	Video Text	Zh	✓	✓	△	✓	△	△	
MMRole-Agent (Dai et al., 2025)	2025	ICLR	Image Text	mixed	✓	✓	△	△	△	△	
SpeechRole (Jiang et al., 2025)	2025	preprint	Speech	En	△	✓	✓	△	△	△	
VoxRole (Wu et al., 2025)	2025	preprint	Speech	En	△	✓	△	△	△	△	
(E) Adaptable Dialogue Resources <i>Emotion/dialogue resources adaptable to RP evaluation.</i>											
MELD (Poria et al., 2019)	2019	ACL	Video Audio Text	En	−	△	✓	△	−	−	1,433 dial.; 13.7K utt.; Friends 10K utt.; 12 hours; lab acting
IEMOCAP (Busso et al., 2008)	2008	dataset	Audio Video Text	En	−	△	✓	△	−	−	

Table 1: A benchmark metadata framework with representative samples grouped by task setting.

cover K and E most strongly: whether a model knows the character and sounds plausibly in character. M, R, and B receive less attention. A model can answer factual questions and mimic a style yet still fail over long interactions, accept a user’s false claims, or make decisions that contradict the role.

Situated Role-Playing. The second block moves from profile matching to scene- or world-grounded interaction. LIGHT and CharacterBox ask agents to act within a social or virtual environment rather than merely answer profile questions (Urbanek et al., 2019; Wang et al., 2025b). Expression, scenario fit, and interactional behavior become easier to observe, as do role-state problems: the agent must track where it is, whom it is addressing, and what the situation permits. Most situated benchmarks, however, still provide only partial evidence for memory continuity and robustness. A response may fit one scene without the role remaining stable across repeated scenes, delayed references, or adversarial pressure.

Diagnostic Probing. The third block isolates particular dimensions through interviews, reading-comprehension probes, thought questions, and role-

stress tests. InCharacter turns psychological interviews into a personality-fidelity evaluation (Wang et al., 2024b). RoleMRC probes role-specific knowledge (Lu et al., 2025), while RPA Thought targets role-specific reasoning (Xu et al., 2025). PICon and counselor-role evaluations introduce multi-turn pressure, conflict, and domain-specific constraints (Kim et al., 2026; Rudolph et al., 2026). These benchmarks make otherwise hidden failures measurable: a wrong standpoint, unstable values, role drift, or unsafe refusal behavior. Their limitation is construct validity. A plausible explanation or stable questionnaire response does not guarantee faithful behavior in an open-ended interaction.

Multimodal Role-Playing. The fourth block covers a newer frontier. MMRole, Video2Roleplay, MMRole-Agent, SpeechRole, and VoxRole ask whether consistency survives when a role is grounded in images, video, audio, or speech (Dai et al., 2025; Zhang et al., 2025c; Jiang et al., 2025; Wu et al., 2025). Visual appearance, voice, scene cues, and audiovisual context become part of K and E. This matters for embodied agents, games, virtual companions, and social simulation, but it

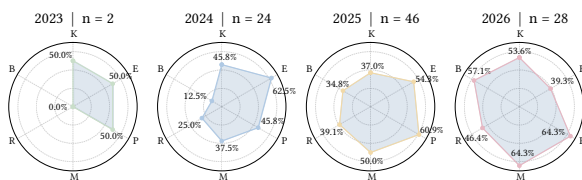


Figure 6: **Temporal change** in role-consistency signals across six dimensions.

also complicates evaluation. A model may preserve its textual persona while losing voice style, visual identity, emotional tone, or action coherence.

Adaptable Dialogue Resources. The final block lists resources that were not designed for RPA consistency but can be adapted to it. MELD and IEMO-CAP provide emotion-rich multimodal dialogue data (Poria et al., 2019; Busso et al., 2008). They can probe expression, affect, and interactional coherence when emotional or paralinguistic evidence is needed. Their limitation is role specificity: emotion labels and acted dialogue do not define a stable character state, knowledge boundary, or role-specific decision policy. These datasets are therefore auxiliary resources, not direct benchmarks of RPA consistency.

5.2 Cross-Dimensional Gaps

The columns of Table 1 expose the central gap. Benchmarks are strongest for K and E, increasingly useful for P and R, and still sparse for M and B. Yet many consequential RPA failures are longitudinal or conflict-driven rather than single-turn factual errors. A stronger evaluation stack should combine automatic checks, human judgment, and calibrated LLM-as-judge protocols. Automatic tests can verify facts, knowledge boundaries, and structural constraints. Human raters can judge whether behavior feels in character and emotionally coherent. Model judges can scale pairwise comparisons or rubric scoring, but require controls for verbosity, familiarity, and the tendency to confuse helpfulness with role fidelity.

Claim: Evaluation Gap

Current evaluation is strongest at testing whether agents know the role and sound in character, but weakest at testing whether they sustain it over time. Linguistic signals make this gap more observable: an agent may preserve persona facts while drifting in politeness strategy, social distance, register, emotional density, or role-specific worldview.

6 Open Challenges and Future Directions

The role-state view brings four open problems into focus.

Controllable Long-Term Memory. A long-term memory system must decide what to remember, forget, update, and reveal. Doing so requires provenance, temporal structure, uncertainty estimates, and boundary control, together with a clear separation among canon, interaction-derived memory, inferred traits, and user-specific relational memory. Existing work provides useful components—structured character retrieval, experience organization, dynamic context selection, and production-scale memory—but these components rarely share an explicit update policy (Wang et al., 2025c; Chen et al., 2025b; Park et al., 2025; Chhikara et al., 2025). As a result, an agent may retrieve the right fact while updating the wrong state, preserve a user’s correction after it has expired, or leak information that the enacted character should not know.

A promising direction is a typed memory model in which every item records its source, scope, confidence, temporal validity, and access boundary. Updates should be reversible and auditable, with evaluations covering delayed correction, selective forgetting, conflicting evidence, privacy deletion, and transfer across scenarios. This would turn memory from an unstructured context cache into a controlled state-transition mechanism.

Robust Evaluation Under Role Attacks. Users can induce drift through false assertions, break-character requests, scenario shifts, or memory manipulation. Evaluations should include such role attacks and measure not only whether the agent refuses, but whether it can refuse without stepping out of character. Current probes expose some of these pressures, yet most benchmark instances remain short, independently scored, and known to the evaluator (Kim et al., 2026; El Boudouri et al., 2025; Yang et al., 2026a). They do not reveal how small errors accumulate, whether recovery is possible after a state corruption, or whether a defense that succeeds for one persona simply collapses the model into a generic safety voice.

Future test suites should therefore use branching interaction trajectories. Matched clean and attacked conversations can isolate causal effects, while hidden checkpoints can measure identity, knowledge boundaries, decisions, and memory after the attack. Reporting trajectory length, time-to-first-drift, at-

tack success, recovery, and persistent corruption would reward systems that maintain safety and character coherence together instead of optimizing refusal rate in isolation.

Language-Grounded Multimodal, Embodied, and Agentic Consistency. As RPAs enter games, virtual worlds, social robots, and multimodal assistants, language will mediate perception, intention, social meaning, and action. Future work should examine how it anchors voice, gesture, visual identity, tool use, movement, and environmental action to one coherent role state. Emerging image-, video-, and speech-grounded benchmarks demonstrate that each modality contributes distinct evidence about identity and expression (Dai et al., 2025; Zhang et al., 2025c; Jiang et al., 2025). The unresolved problem is cross-modal state binding: a character may speak consistently while selecting an implausible action, preserve visual identity while changing social stance, or infer information from an image that lies outside its knowledge boundary.

Evaluation must therefore move beyond utterance-level fidelity to language-grounded behavior, where an inconsistent action may have more serious safety consequences than inconsistent text. Protocols should compare what the agent says, attends to, remembers, and does at the same decision point. Counterfactual scene edits can test whether actions follow relevant evidence, creating a common evaluation unit—the grounded role decision—across text, speech, perception, and action.

Conflict-Aware Role-State Arbitration. Role consistency is a constrained objective, not an instruction to imitate a character at any cost. A deployed agent must arbitrate among canonical role facts, interaction-derived memories, user goals, task requirements, platform policy, and safety constraints. Existing alignment work shows that role prompts can pressure safety behavior and that a correct refusal can still be character-breaking (Tang et al., 2024; Li et al., 2026; Yang et al., 2026a). What remains underspecified is the precedence rule: which source may update which part of the role state, which constraints are immutable, and how the agent should communicate a conflict without revealing hidden policy or inventing an in-world justification.

Future systems should represent these priorities explicitly and evaluate arbitration through observable decisions rather than unrestricted chain-of-

thought. Paired conflicts—canon versus user assertion, loyalty versus factual evidence, and role goal versus safety boundary—can test whether a resolution remains stable over later turns. Such tests would distinguish principled adaptation from arbitrary refusal and make safety-preserving role fidelity a verifiable systems property.

7 Conclusion

We organized recent work with a compact role-state taxonomy spanning role specification, role-state maintenance, role-aware reasoning, and role-consistent alignment, and treated evaluation as a cross-cutting problem in its own right. The field has made clear progress. Robust consistency, however, will require methods and evaluations that are longitudinal, conflict-aware, and safety-preserving.

Limitations

The literature on role-playing agents, persona consistency, personalization, agent memory, and safety is moving quickly, and frequent new preprints and benchmarks will inevitably make parts of this survey incomplete. Because our taxonomy centers role consistency, it gives less attention to general-purpose dialogue, social simulations without explicit role fidelity, and personalization systems that optimize user preference rather than character stability. The classification is also interpretive: many systems combine memory, reasoning, training, and evaluation, so assigning a primary research line simplifies their contributions. Finally, the synthesis remains weighted toward text-based RPAs; multimodal, embodied, and voice-based consistency receive less coverage.

Ethical Considerations

RPAs can create a strong impression of identity, memory, intimacy, authority, or expertise and may mislead users unless their artificial nature is clear. Responsible development therefore calls for transparent disclosure, privacy-preserving memory, consent-aware data collection, culturally sensitive evaluation, and safety mechanisms that constrain role enactment without hidden or inconsistent behavior.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62572104 and 62220106008.

References

- Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. [Consistently simulating human personas with multi-turn reinforcement learning](#). *Preprint*, arXiv:2511.00222.
- Guojia An, Jie Zou, Jiwei Wei, Chaoning Zhang, Fuming Sun, and Yang Yang. 2025. [Beyond whole dialogue modeling: Contextual disentanglement for conversational recommendation](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 31–41. Association for Computing Machinery.
- Pranav Bhandari, Nicolas Fay, Michael J. Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025. [Can llm agents maintain a persona in discourse?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Haoyu Bian, Chaoning Zhang, Jiaquan Zhang, Xingyao Li, Yuanfang Guo, Wei Dong, and Yang Yang. 2026. [Weak-link optimization for multi-agent reasoning and collaboration](#). *Preprint*, arXiv:2604.15972.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. [IEMOCAP: Interactive emotional dyadic motion capture database](#). *Language Resources and Evaluation*, 42(4):335–359.
- Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Toby Jia-Jun Li, and Dakuo Wang. 2025a. [Towards a design guideline for rpa evaluation: A survey of large language model-based role-playing agents](#). In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.
- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. [From persona to personalization: A survey on role-playing language agents](#). *Transactions on Machine Learning Research*.
- Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024b. [The oscars of ai theater: A survey on role-playing with language models](#). *Preprint*, arXiv:2407.11484.
- Shuxu Chen, Yitian Zhou, Jiaquan Zhang, Haoyu Bian, Aming Wu, Sungyoung Lee, Chaoning Zhang, and Hyundong Shin. 2026. [CAP-CoT: Cycle adversarial prompt for improving chain of thoughts in LLM reasoning](#). *Preprint*, arXiv:2604.23270.
- Weishu Chen, Jinyi Tang, Zhouhui Hou, Shihao Han, Mingjie Zhan, Zhiyuan Huang, Delong Liu, Jiawei Guo, Zhicheng Zhao, and Fei Su. 2025b. [Moom: Maintenance, organization and optimization of memory in ultra-long role-playing dialogues](#). *Preprint*, arXiv:2509.11860.
- Xilong Cheng, Yunxiao Qin, Yuting Tan, Zhengnan Li, Ye Wang, Hongjiang Xiao, and Yuan Zhang. 2025. [Psymem: Fine-grained psychological alignment and explicit memory control for advanced role-playing llms](#). *Preprint*, arXiv:2505.12814.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. [Mem0: Building production-ready ai agents with scalable long-term memory](#). *Preprint*, arXiv:2504.19413.
- Alex Clay and Ernesto Jiménez-Ruiz. 2024. [Information for conversation generation: Proposals utilising knowledge graphs](#). *Preprint*, arXiv:2410.16196.
- Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2025. [MMRole: A comprehensive framework for developing and evaluating multimodal role-playing agents](#). In *International Conference on Learning Representations*.
- Yassine El Boudouri, Walter Nuninger, Julian Alvarez, and Yvan Peter. 2025. [Role-playing evaluation for large language models](#). *Preprint*, arXiv:2505.13157.
- Zhenpeng Gao, Xiaofen Xing, and Xiangmin Xu. 2025. [Tailorrrpa: A retrieval-based framework for eliciting personalized and coherent role-playing agents in general domain](#). In *Findings of the Association for Computational Linguistics: EMNLP*, pages 5381–5412. Association for Computational Linguistics.
- Jinyu Guo, Xunlei Chen, Qiyang Xia, Zhaokun Wang, Jie Ou, Libo Qin, Shunyu Yao, and Wenhong Tian. 2025. [HASH-RAG: Bridging deep hashing with retriever for efficient, fine retrieval and augmented generation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26847–26858, Vienna, Austria. Association for Computational Linguistics.
- Ilya Gusev. 2024. [Pingpong: A benchmark for role-playing language models with user emulation and multi-model evaluation](#). *Preprint*, arXiv:2409.06820.
- Xiaoxu Han, Wanqing Zhao, Ziyu Guan, and Jinye Peng. 2026. [Act-llm: A whole-process chain for character-centric role-playing with large language models](#). *Expert Systems with Applications*, 296:129024.
- Zhenzhen Huang, Chaoning Zhang, Fachrina Dewi Puspitasari, Jiaquan Zhang, Yitian Zhou, Shuxu Chen, and Yang Yang. 2026. [Small language model helps resolve semantic ambiguity of LLM prompt](#). *Preprint*, arXiv:2604.23263.
- Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. [Enhancing persona consistency for llms’ role-playing using persona-aware contrastive learning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26221–26238. Association for Computational Linguistics.
- Huadong Jian, Chenghao Li, Haoyu Wang, Jiajia Shuai, Jinyu Guo, Yang Yang, and Chaoning Zhang.

2026. [Gated coordination for efficient multi-agent collaboration in Minecraft game](#). *Preprint*, arXiv:2604.18975.
- Bowen Jiang, Taiwei Shi, Ryo Kamoi, Yuan Yuan, Camillo J. Taylor, Longqi Yang, Pei Zhou, and Sihao Chen. 2026. [One model, all roles: Multi-turn, multi-agent self-play reinforcement learning for conversational social intelligence](#). *Preprint*, arXiv:2602.03109.
- Changhao Jiang, Jiajun Sun, Yifei Cao, Jiabao Zhuang, Xinmeng Che, Hui Li, Xiaoran Fan, Ming Zhang, Junjie Ye, Shihan Dou, Zhiheng Xi, Jingqi Tong, Yilong Wu, Baoyu Fan, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. [Speechrole: A large-scale dataset and benchmark for evaluating speech role-playing agents](#). *Preprint*, arXiv:2508.02013.
- Donghyun Kim, Youbin Ahn, Wongyu Kim, Chanhee Lee, Kyungchan Lee, Kyong-Ho Lee, Jeonguk Kim, Donghoon Shin, and Yeonsoo Lee. 2023. [Persona expansion with commonsense knowledge for diverse and consistent response generation](#). In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 1139–1149. Association for Computational Linguistics.
- Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. [Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 904–916. Association for Computational Linguistics.
- Minseo Kim, Sujeong Im, Junseong Choi, Junhee Lee, Chaeun Shim, Hwajung Hong, and Edward Choi. 2026. [Picon: A multi-turn interrogation framework for evaluating persona agent consistency](#). *Preprint*, arXiv:2603.25620.
- 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. 2023. [Chatharuhi: Reviving anime character in reality via large language model](#). *Preprint*, arXiv:2308.09597.
- Jiajia Li, Xiaoyu Wen, Zhongtian Ma, Shuyue Hu, Qiaosheng Zhang, and Zhen Wang. 2026. [Disentangling intent from role: Adversarial self-play for persona-invariant safety alignment](#). *Preprint*, arXiv:2605.01899.
- Cheng Liu, Yifei Lu, Fanghua Ye, Jian Li, Xingyu Chen, Feiliang Ren, Zhaopeng Tu, and Xiaolong Li. 2025. [CogDual: Enhancing dual cognition of LLMs via reinforcement learning with implicit rule-based rewards](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27307–27336, Suzhou, China. Association for Computational Linguistics.
- Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022. [Improving personality consistency in conversation by persona extending](#). In *Proceedings of the ACM International Conference on Information and Knowledge Management*.
- Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. 2025. [Rolemrc: A fine-grained composite benchmark for role-playing and instruction-following](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21008–21030. Association for Computational Linguistics.
- Mohsen Mesgar, Edwin D. Simpson, and Iryna Gurevych. 2021. [Improving factual consistency between a response and persona facts](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 549–562. Association for Computational Linguistics.
- Chang Nie, Chaoyou Fu, Yifan Zhang, Haihua Yang, and Caifeng Shan. 2026. [Personavlm: Long-term personalized multimodal llms](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Jeiyeon Park, Yongshin Han, Minseop Kim, and Kisu Yang. 2025. [Dynamic context adaptation for consistent role-playing agents with retrieval-augmented generations](#). *Preprint*, arXiv:2508.02016.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. [Meld: A multimodal multi-party dataset for emotion recognition in conversations](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. [Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14566–14576, Miami, Florida, USA.
- Eric Rudolph, Natalie Engert, and Jens Albrecht. 2026. [Evaluating role-consistency in LLMs for counselor training](#). *Preprint*, arXiv:2601.08892.
- Arpita Saggat, Jonathan C. Darling, Vania Dimitrova, Duygu Sarikaya, and David C. Hogg. 2025. [Score before you speak: Improving persona consistency in dialogue generation using response quality scores](#). *Preprint*, arXiv:2508.06886.
- Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2025. [Personagym: Evaluating persona agents and llms](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. [Role play with large language models](#). *Nature*, 623(7987):493–498.

- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. [Character-llm: A trainable agent for role-playing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore, Singapore. Association for Computational Linguistics.
- Ryan Shea and Zhou Yu. 2023. [Building persona consistent dialogue agents with offline reinforcement learning](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1778–1795, Singapore. Association for Computational Linguistics.
- Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. [Generating persona consistent dialogues by exploiting natural language inference](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8878–8885.
- Junya Takayama, Masaya Ohagi, Tomoya Mizumoto, and Katsumasa Yoshikawa. 2025. [Persona-consistent dialogue generation via pseudo preference tuning](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5507–5514, Abu Dhabi, UAE. Association for Computational Linguistics.
- Wenqiu Tang, Zhen Wan, Takahiro Komamizu, and Ichiro Ide. 2026a. [Facet-level persona control by trait-activated routing with contrastive sae for role-playing llms](#). *Preprint*, arXiv:2602.19157.
- Yihong Tang, Kehai Chen, Xuefeng Bai, Benyou Wang, Zeming Liu, Haifeng Wang, and Min Zhang. 2026b. [Character-r1: Enhancing role-aware reasoning in role-playing agents via rlvr](#). *Preprint*, arXiv:2601.04611.
- Yihong Tang, Kehai Chen, Muyun Yang, Zheng-Yu Niu, Jing Li, Tiejun Zhao, and Min Zhang. 2025. [Thinking in character: Advancing role-playing agents with role-aware reasoning](#). In *Advances in Neural Information Processing Systems*.
- Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhao, and Kun Zhang. 2024. [Erabal: Enhancing role-playing agents through boundary-aware learning](#). *Preprint*, arXiv:2409.14710.
- Meiling Tao, Liang Xuechen, Tianyu Shi, Lei Yu, and Yiting Xie. 2024. [Rolecraft-glm: Advancing personalized role-playing in large language models](#). In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 1–8. Association for Computational Linguistics.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two tales of persona in llms: A survey of role-playing and personalization](#). *Preprint*, arXiv:2406.01171.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. [Charactereval: A chinese benchmark for role-playing conversational agent evaluation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. [Learning to speak and act in a fantasy text adventure game](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
- Junqiao Wang, Kunyu Wu, and Yuqi Ouyang. 2025a. [Enhancing character-coherent role-playing dialogue with a verifiable emotion reward](#). *Information*, 16(9):738.
- Kai Wang, Haoyang You, Yang Zhang, and Zhongjie Wang. 2026a. [Memory-driven role-playing: Evaluation and enhancement of persona knowledge utilization in llms](#). *Preprint*, arXiv:2603.19313.
- Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025b. [CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.
- Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhang Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. [Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777. Association for Computational Linguistics.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. [Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873. Association for Computational Linguistics.
- Xudong Wang, Chaoning Zhang, Chenghao Li, Shuxu Chen, Qigan Sun, Jiaquan Zhang, Fachrina Dewi Puspitasari, Tae-Ho Kim, Jiwei Wei, Malu Zhang, Guoqing Wang, Yang Yang, and Heng Tao Shen. 2026b. [Agent-GWO: Collaborative agents for dynamic prompt optimization in large language models](#). *Preprint*, arXiv:2604.18612.
- Xudong Wang, Chaoning Zhang, Qigan Sun, Zhenzhen Huang, Chang Lu, Sheng Zheng, Zeyu Ma,

- Caiyan Qin, Yang Yang, and Hengtao Shen. 2026c. [Transforming external knowledge into triplets for enhanced retrieval in RAG of LLMs](#). *Preprint*, arXiv:2604.12610.
- Ye Wang, Jiaying Chen, and Hongjiang Xiao. 2026d. [Role-playing agents driven by large language models: Current status, challenges, and future trends](#). *Preprint*, arXiv:2601.10122.
- Yongjie Wang, Jonathan Leung, and Zhiqi Shen. 2025c. [Rolerag: Enhancing llm role-playing via graph guided retrieval](#). *Preprint*, arXiv:2505.18541.
- Zixu Wang, Bin Xie, Bingbing Xu, Shengmao Zhu, Yige Yuan, Liang Pang, Du Su, Long Yang, Zixuan Li, Huawei Shen, and Xueqi Cheng. 2025d. A survey on llm-based agents for social simulation: Taxonomy, evaluation and applications. *Preprint*.
- Zongsheng Wang, Kaili Sun, Bowen Wu, Qun Yu, Ying Li, and Baoxun Wang. 2025e. [Raiden-r1: Improving role-awareness of llms via grpo with verifiable reward](#). *Preprint*, arXiv:2505.10218.
- Chupeng Wei, Zheng Chen, Jie Zhang, and Jia Liu. 2026. [Roleplayllm: Enhancing role-playing abilities in large language models via multidimensional fine-tuning and preference optimization](#). In *Natural Language Processing and Chinese Computing*, volume 16104 of *Lecture Notes in Computer Science*. Springer.
- Yangbo Wei, Zhen Huang, Fangzhou Zhao, Qi Feng, and Wei W. Xing. 2025. [MECoT: Markov emotional chain-of-thought for personality-consistent role-playing](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8297–8314. Association for Computational Linguistics.
- Weihao Wu, Liang Cao, Xinyu Wu, Zhiwei Lin, Rui Niu, Jingbei Li, and Zhiyong Wu. 2025. [Voxrole: A comprehensive benchmark for evaluating speech-based role-playing agents](#). *Preprint*, arXiv:2509.03940.
- Rui Xu, Mingyu Wang, Xintao Wang, Dakuan Lu, Xiaoyu Tan, Wei Chu, and Yinghui Xu. 2025. [Guess what I am thinking: A benchmark for inner thought reasoning of role-playing language agents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15148–15168, Suzhou, China. Association for Computational Linguistics.
- Bohao Yang, Dong Liu, Chen Tang, Chenghao Xiao, Kun Zhao, Chao Li, Lin Yuan, Guang Yang, Lanxiao Huang, and Chenghua Lin. 2024. [Crafting customisable characters with llms: Introducing simschat, a persona-driven role-playing agent framework](#). *Preprint*, arXiv:2406.17962.
- Jisoo Yang, Jongwon Ryu, Minuk Ma, Trung X. Pham, and Junyeong Kim. 2026a. [The pragmatic persona: Discovering llm persona through bridging inference](#). *Preprint*, arXiv:2604.24079.
- Shihao Yang, Zhicong Lu, Yong Yang, Bo Lv, Yang Shen, and Nayu Liu. 2026b. [HyCoRA: Hyper-contrastive role-adaptive learning for role-playing](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34304–34312.
- Bingsheng Yao, Bo Sun, Yuanzhe Dong, Yuxuan Lu, and Dakuo Wang. 2025. [Dprf: A generalizable dynamic persona refinement framework for optimizing behavior alignment between personalized llm role-playing agents and humans](#). *Preprint*, arXiv:2510.14205.
- Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Zhenzhen Huang, Pengcheng Zheng, Zhicheng Wang, Ping Guo, Fan Mo, Sung-Ho Bae, Jie Zou, Jiwei Wei, and Yang Yang. 2026. [Lightweight LLM agent memory with small language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12914–12929, San Diego, California, United States. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.
- Wenyuan Zhang, Shuaiyi Nie, Jiawei Sheng, Zefeng Zhang, Xinghua Zhang, Yongquan He, and Tingwen Liu. 2025a. [Revealing and mitigating the challenge of detecting character knowledge errors in LLM role-playing](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33279–33302, Suzhou, China. Association for Computational Linguistics.
- Xiaotian Zhang, Ruizhe Chen, Yang Feng, and Zuozhu Liu. 2025b. [Persona-judge: Personalized alignment of large language models via token-level self-judgment](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5037–5049. Association for Computational Linguistics.
- Xueqiao Zhang, Chao Zhang, Jingtao Xu, Yifan Zhu, Xin Shi, Yi Yang, and Yawei Luo. 2025c. [Video2roleplay: A multimodal dataset and framework for video-guided role-playing agents](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23677–23703, Suzhou, China. Association for Computational Linguistics.
- Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. [Characterbench: Benchmarking character customization of large language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26101–26110.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. *Simoap: Improve coherence and consistency in persona-based dialogue generation via over-sampling and post-evaluation*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9945–9959, Toronto, Canada. Association for Computational Linguistics.

Yitian Zhou, Chaoning Zhang, Jiaquan Zhang, Zhenzhen Huang, Jinyu Guo, Sung-Ho Bae, Lik-Hang Lee, Caiyan Qin, and Yang Yang. 2026. *From similarity to structure: Training-free LLM context compression with hybrid graph priors*. *Preprint*, arXiv:2604.23277.

A Comparison with Related Surveys

This comparison shows where existing surveys differ in scope, method, and treatment of consistency, and where our narrower role-state view complements them.

1. **Broad role-playing with language models, including data, models, agents, and applications** (Chen et al., 2024b). General taxonomy of RPA components and challenges. *Discusses persona and character consistency as part of the broader RPA landscape. We make consistency the central organizing problem and analyze it as role-state maintenance.*
2. **Persona, role-playing, and personalization across language agents** (Chen et al., 2024a). Connects persona construction, personalized generation, and evaluation. *Treats consistency as one requirement among persona and personalization goals. We separate role consistency from user personalization and emphasize longitudinal role boundaries.*
3. **Two uses of persona in LLMs: role-playing and personalization** (Tseng et al., 2024). Clarifies conceptual relations between persona-driven interaction and personalized systems. *Covers persona consistency but not as a full systems-level taxonomy. We extend from persona consistency to identity, memory, reasoning, alignment, and safety consistency.*
4. **Design guidelines for evaluating LLM-based RPAs** (Chen et al., 2025a). Evaluation design, benchmark construction, and assessment dimensions. *Strongest overlap on evaluation of role-playing quality. We treat evaluation as one layer and connect it to memory, reasoning, training, and boundary-aware alignment.*
5. **Current status, challenges, and trends of LLM-driven RPAs** (Wang et al., 2026d). Broad overview of architectures, capabilities, applications, and future trends. *Identifies consistency as an important challenge. We provide a narrower taxonomy centered on maintaining coherent role state under long, changing, and adversarial interaction.*

B Word Cloud Keyword List

Figure 2 draws on the titles and abstracts of 100 arXiv papers retrieved with role-playing-agent

queries. These included *role-playing agent*, *role-play agent*, *persona agent large language model*, *character agent large language model*, *persona consistency large language model*, *character consistency large language model*, and *role consistency agent*. We deduplicated results by arXiv identifier, scored candidates with positive RPA indicators (e.g., role-playing, persona, character, in-character, dialogue agent, virtual character, NPC, and LLM), and applied negative filters for unrelated meanings of RPA (e.g., robotic process automation and remote piloted aircraft). We then retained the 100 highest-scoring papers above the relevance threshold.

We computed keyword hotness from each paper’s title and abstract. Variants were first normalized to canonical phrases, including role-playing, persona, character, memory, retrieval, alignment, safety, reasoning, multimodal, personality, style, identity, knowledge, dialogue, multi-turn, simulation, game, training, preference optimization, and benchmark. Remaining unigram candidates were lowercased, stripped of punctuation, filtered against stopwords, and retained only if they appeared at least three times. For each term k , we computed term frequency tf_k , document frequency df_k , and inverse document frequency

$$idf_k = \log \frac{1 + N}{1 + df_k} + 1, \quad (1)$$

where $N = 100$. To emphasize recent activity, we also computed

$$growth_k = \log(1 + tf_k^{2025+}) - \log(1 + tf_k^{<2025}), \quad (2)$$

and ranked terms by

$$hotness_k = tf_k \cdot idf_k \cdot r_k \cdot c_k, \quad (3)$$

where $r_k = 1 + 0.25 \max(growth_k, 0)$ is the recency factor and $c_k = 1 + 0.2 \min(df_k/N, 1)$ is the document-coverage factor. For the rendered word cloud, generic search-anchor terms such as *role-playing* and *large language model* were removed, and consistency-family terms, including *consistency*, *consistent*, *persona consistency*, *character consistency*, *role consistency*, *fidelity*, and *coherence*, were merged into *consistency*. Table 3 lists the complete keyword set used in the rendered word cloud. Table 2 below reports selected unmerged diagnostic signals for interpretability, so its values should not be directly compared with merged word-cloud entries such as *consistency*.

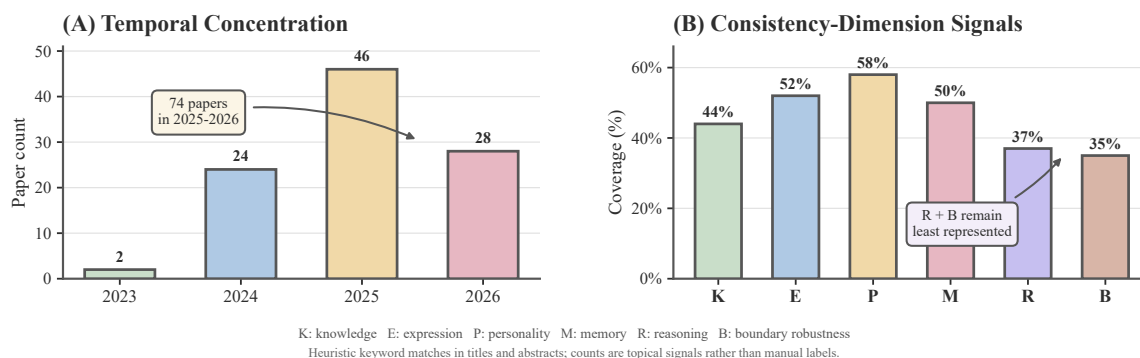


Figure 7: **Additional diagnostics over the 100-paper arXiv sample.** Left: submitted-year distribution. Right: heuristic title/abstract coverage of the six role-consistency dimensions used in Table 1. Counts are based on keyword matches and should be read as topical salience indicators, not as manual labels.

C Additional Sample Diagnostics

We add two lightweight diagnostics over the same 100-paper sample to make the descriptive analysis behind Figure 2 more transparent. They are not exhaustive bibliometric evidence; their purpose is to show the topical shape of the sample that motivates the survey.

The sample is strongly recent: 74 of the 100 papers in Figure 7 were submitted in 2025 or 2026. This concentration motivates a focused survey, but it also helps explain why methods remain unsettled. Many papers are preprints, benchmark definitions are still changing, and several dimensions have yet to stabilize into shared protocols.

The right panel maps titles and abstracts to the six dimensions in our benchmark map: knowledge/factual consistency (K), expression/style consistency (E), psychological/personality fidelity (P), memory and temporal continuity (M), role-aware reasoning and decision consistency (R), and robustness to role drift or boundary pressure (B). Personality, expression, and memory signals appear in roughly half the sample; reasoning and boundary robustness appear less often. This pattern supports the main-text argument that current work increasingly looks beyond factual persona matching, while methods and evaluations for decision-level consistency and adversarial boundary control remain comparatively sparse.

Table 2 clarifies which mechanisms are prominent. Personality, reasoning, training, alignment, and benchmarks are among the strongest signals, suggesting a move from static character description toward controllable behavior and evaluation. Safety, multimodality, multi-turn continuity, and preference optimization are concentrated in recent

work but appear in fewer papers. We therefore treat them as emerging rather than settled: they motivate the challenges in Section 6, but do not yet constitute mature subfields with standardized tasks and metrics.

How to read the hotness scores. The ranking organizes this sample; it does not estimate a topic’s absolute importance. A high score may reflect repeated mentions within papers, broad document coverage, recent growth, or some combination of the three. Scores should therefore be compared within the sample, not treated as calibrated measurements across corpora. The merged *consistency* entry illustrates the point: it pools several lexical variants so that the central construct is not split across near-synonyms. Its magnitude reflects both topical centrality and a normalization decision, while lower-ranked entries may denote narrower concepts with less standardized terminology.

What the leading terms suggest. The leading terms mix role descriptors—*character*, *persona*, and *personality*—with process terms such as *reasoning*, *interaction*, *training*, and *alignment*. The mixture captures a transition in the literature. Earlier work often treated the role as a profile to imitate; newer work asks how that profile is maintained through decisions, dialogue history, optimization, and environmental feedback. The prominence of *benchmark* also shows that evaluation resources have become part of the research problem. These signals fit the layered view developed in the main paper: consistency depends not only on what an agent says, but on how it updates state, selects actions, handles pressure, and is measured over time.

Keyword signal	Hotness	TF	Papers	Recent TF	Earlier TF	Interpretation
Recent TF = 2025–2026 term frequency; Earlier TF = pre-2025 term frequency.						
Role-state content and behavior						
personality	341.163	104	25	124	33	trait and psychological fidelity are highly visible
reasoning	242.400	65	28	107	10	role-aware deliberation is a fast-growing topic
consistency	168.331	40	27	39	1	explicit consistency framing is mostly recent
memory	108.231	29	11	36	18	memory is visible but fragmented across systems
style	107.419	22	9	35	5	expressive consistency is often implicit
knowledge	89.157	26	16	36	16	factual role knowledge remains a baseline concern
multi-turn	65.097	18	13	22	10	long-horizon interaction is not yet universal
Methods, resources, and emerging constraints						
training	190.348	61	34	79	14	supervised and alignment data remain central
alignment	165.991	44	29	68	5	role fidelity is increasingly tied to optimization
benchmark	163.108	58	31	42	16	evaluation resources are a major research driver
multimodal	151.114	22	7	38	0	multimodal RPAs are an emerging frontier
safety	144.922	21	6	30	0	safety and boundary control remain underrepresented
retrieval	110.363	32	9	22	20	retrieval is present but less dominant than training
preference optimization	18.318	3	2	3	0	preference methods are still early in RPA-specific work

Table 2: **Selected keyword diagnostics from the 100-paper arXiv sample.** Hotness follows the scoring formula in Appendix B; TF denotes term frequency in titles and abstracts. These diagnostic rows use unmerged keyword signals, while Table 3 reports the merged keyword list used for the rendered word cloud.

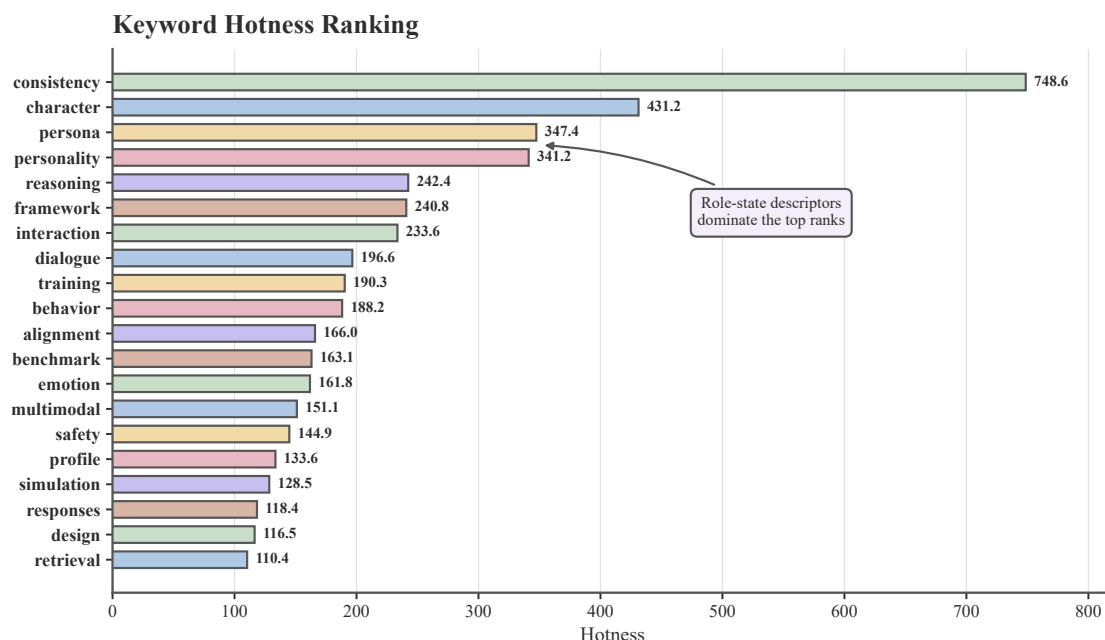


Figure 8: **Keyword hotness ranking for the 100-paper arXiv sample.** The bar chart provides a quantitative companion to the word cloud in Figure 2.

Reading the head and tail. Figure 8 reveals a steep ranking: broad role descriptors lead, process and capability terms occupy the middle, and specific mechanisms form the tail. A lower rank does not imply lower importance. Terms such as *memory*, *multi-turn*, *identity*, and *preference optimization* apply to narrower settings and may be distributed across related expressions such as state tracking, long-horizon dialogue, or policy optimization. The ranking is most informative alongside the taxonomy and benchmark analysis, not as a stan-

dalone research agenda.

Reproducibility and limits. Table 3 records the exact ordering behind the cloud and bar chart; its four column pairs are a compact layout, not semantic groups. The cloud supports rapid recognition, the chart shows relative magnitude, and the table preserves rounded scores for reconstruction. These values measure lexical salience in the retrieved sample, not method quality or empirical effectiveness. Titles and abstracts may omit technical distinctions, and the queries intentionally favor role-playing and

#	Keyword / Hotness	#	Keyword / Hotness	#	Keyword / Hotness	#	Keyword / Hotness
1	consistency / 748.649	2	character / 431.192	3	persona / 347.439	4	personality / 341.163
5	reasoning / 242.400	6	framework / 240.805	7	interaction / 233.567	8	dialogue / 196.596
9	training / 190.348	10	behavior / 188.224	11	alignment / 165.991	12	benchmark / 163.108
13	emotion / 161.845	14	multimodal / 151.114	15	safety / 144.922	16	profile / 133.550
17	simulation / 128.481	18	responses / 118.432	19	design / 116.473	20	retrieval / 110.363
21	social / 108.971	22	memory / 108.231	23	dynamic / 108.080	24	style / 107.419
25	systems / 106.353	26	multi-agent / 102.540	27	generation / 100.813	28	traits / 99.041
29	scenario / 93.150	30	knowledge / 89.157	31	cognition / 88.194	32	multi-turn / 65.097
33	game / 53.404	34	world / 44.071	35	identity / 40.846	36	role-aware / 37.943
37	virtual / 37.647	38	character eval / 27.207	39	character box / 24.659	40	preference opt. / 18.318

Table 3: **Complete keyword list used to render the word cloud in Figure 2.** Hotness scores are rounded to three decimals.

consistency research. We use these diagnostics only for triangulation and reproducibility; claims about methods and benchmark coverage still rest on the paper-level review in the main text.