Towards a Design Guideline for RPA Evaluation: A Survey of Large Language Model-Based Role-Playing Agents

Chaoran Chen[†]
University of Notre Dame

Bingsheng Yao[†]
Northeastern University

Ruishi Zou
University of California, San Diego

Wenyue Hua
University of California, Santa Barbara

Weimin Lyu
Stony Brook University

Toby Jia-Jun Li
University of Notre Dame

Dakuo Wang [*]
Northeastern University

Abstract

Role-Playing Agent (RPA) is an increasingly popular type of LLM Agent that simulates human-like behaviors in a variety of tasks. However, evaluating RPAs is challenging due to diverse task requirements and agent designs. This paper proposes an evidence-based, actionable, and generalizable evaluation design guideline for LLM-based RPA by systematically reviewing 1,676 papers published between Jan. 2021 and Dec. 2024. Our analysis identifies six agent attributes, seven task attributes, and seven evaluation metrics from existing literature. Based on these findings, we present an RPA evaluation design guideline to help researchers develop more systematic and consistent evaluation methods.

1 Introduction

LLMs have yielded human-like performance in various cognitive tasks (e.g., memorization (Schwarzschild et al., 2025), reasoning (Wang et al., 2023a; Plaat et al., 2024), and planning (Song et al., 2023; Huang et al., 2024)). These emergent capabilities have fueled growing research interest on **Role-Playing Agent** (RPA) (Chen et al., 2024d; Tseng et al., 2024): RPAs are digital intelligent agent systems powered by LLMs, where users provide human-like **agent attributes** (e.g., personas) and **task attributes** (e.g., task descriptions) as input, and prompt the LLM to generate human-like behaviors and the reasoning process. The potential of RPAs is promising and far-reaching, as illustrated by the early results of the massive interdisciplinary studies in social science (Park et al., 2022, 2023; Hua et al., 2023), network science (Chen et al., 2024b), psychology(Jiang et al., 2024) and juridical science (He et al., 2024b).

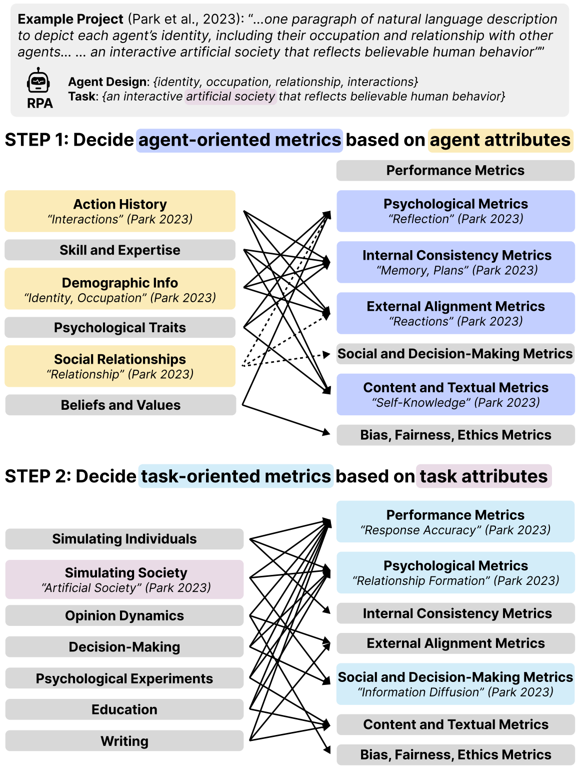

Figure 1: RPA evaluation design guideline. To illustrate how to use it in practice, we pretended we were selecting the evaluation metrics for the "Stanford Agent Village" (Park et al., 2023) given agent attributes (yellow) and task attributes (pink). The original authors' selection of evaluation metrics (purple and blue) perfectly aligns with our RPA design guideline, which echoes their work's robustness. More details in Sec 5.1 and a bad example in Sec 5.2.

Despite growing interest in RPAs, a fundamental question remains: **how can we systematically and consistently evaluate an RPA?** How should we select the evaluation metrics, so that the evaluation results can be comparable or generalizable from one task to another task? Addressing these challenges is difficult (Dai et al., 2024; Tu et al., 2024; Wang et al., 2024c). due to the vast diversity of tasks (e.g., simulating an individual's online

[*]Corresponding author: d.wang@northeastern.edu
[†] Equal contribution.
Github repository: https://github.com/CRChenND/LLM_roleplay_agent_eval_survey
Searchable webpage: https://agentsurvey.hailab.io/

Findings of the Association for Computational Linguistics: ACL 2025, pages 18229–18268
July 27 - August 1, 2025

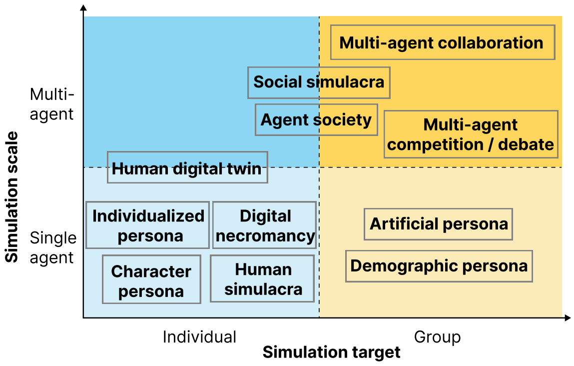

Figure 2: Taxonomy of RPAs.

browser behavior (Chen et al., 2024b) or simulating a hospital (Li et al., 2024c)), and the high flexibility in RPA design (e.g., an agent persona can be one sentence or 2-hours of interview log (Park et al., 2024)). Another challenge is the inconsistent and often arbitrary selection of evaluation methods and metrics for RPAs, raising concerns about the validity and reliability of evaluation results (Wang et al., 2025; Zhang et al., 2025). As a result, the research community finds it difficult to compare the performance across multiple RPAs in similar tasks reliably and systematically.

To address this gap, we propose an evidence-based, actionable, and generalizable design guideline for evaluating LLM-based RPAs. We conducted **a systematic literature review** of $1,676$ papers on the LLM Agent topic and identified 122 papers describing its evaluation details. Through expert coding, we found that agent attribute design interacts with task characteristics (e.g., simulating an individual or simulating a society requires a diverse set of agent attributes). Furthermore, we synthesized common patterns in how prior research successfully (or unsuccessfully) designed their evaluation metrics to correspond to the RPA's agent attributes and task attributes. Building on these insights, we propose an RPA evaluation design guideline (Fig. 1) and illustrate its generalizability through two case studies.

2 Related Work

2.1 Taxonomy of RPAs

Existing literature (Chen et al., 2024d; Tseng et al., 2024; Chen et al., 2024e; Mou et al., 2024a) classifies RPAs along two independent dimensions: Simulation Target and Simulation Scale. The Simulation Target dimension differentiates between agents that simulate specific individuals (e.g., historical figures, fictional characters, or individualized personas) and those that simulate group characteristics (e.g., artificial personas) (Chen et al., 2024d; Tseng et al., 2024; Chen et al., 2024e). The Simulation Scale dimension categorizes agents by the complexity of their interactions, ranging from single-agent simulations with no social interaction to multi-agent systems that replicate structured or emergent societal behaviors (Mou et al., 2024a).

To unify these perspectives, we introduce an integrated taxonomy for RPAs (Fig.2). The *Simulation Target* axis distinguishes between individual-focused and group-focused agents. Examples of individual-focused agents include digital twins, which model an individual's decision-making process (Rossetti et al., 2024), and personas, which emulate specific human-like characteristics (Chen et al., 2024b). Group-focused agents include social simulacra, which model interactions between specific individuals within a group (e.g., the relationship dynamics in Detective Conan) (Wu et al., 2024a), and synthetic societies, which replicate large-scale social structures and emergent group behaviors (Park et al., 2023). The *Simulation Scale* axis differentiates between single-agent and multi-agent systems. Single-agent RPAs operate at an individual level, such as digital twins used for personalized recommendations or personas that generalize group characteristics for interaction. Multi-agent RPAs involve more complex interactions, with social simulacra capturing interpersonal dynamics within small, predefined groups, and synthetic societies modeling large-scale collective decision-making and societal structures.

2.2 Evaluation of RPAs

Existing surveys on the evaluation of RPAs (Gao et al., 2024; Chen et al., 2024d; Tseng et al., 2024; Chen et al., 2024e; Mou et al., 2024a) provide a unified classification of RPA evaluation metrics from the perspective of evaluation approaches. However, they lack a comprehensive and consistent taxonomy for versatile evaluation metrics, leading to arbitrary metrics selection in practices.

Prior works (Gao et al., 2024; Mou et al., 2024a) categorize RPA evaluations into three types: automatic evaluations, human-based evaluations, and LLM-based assessments. Automatic evaluations are efficient and objective, but lack context sensitivity, failing to capture nuances like persona consistency. Human-based evaluations provide deep insight into character alignment and engagement, but they are costly, less scalable, and prone to subjectivity. LLM-based evaluations are automatic and offer scalability and speed, but may not always align with human judgments.

The classification of evaluation metrics in prior works varies significantly, leading to inconsistency and ambiguity. For instance, Gao et al. (2024) focuses on realness validation and ethics evaluation, whereas Chen et al. (2024d) differentiates between character persona and individualized persona. Furthermore, Chen et al. (2024e) classifies evaluation into conversation ability, role-persona consistency, role-behavior consistency, and role-playing attractiveness, which partially overlap with Mou et al. (2024a)'s individual simulation and scenario evaluation. These discrepancies indicate a lack of standardized taxonomy, making it difficult to compare results across studies and select appropriate evaluation metrics for specific applications.

While existing surveys offer different taxonomies of RPA evaluation, they do not provide concrete evaluation design guidelines. Our work addresses this gap by proposing a structured framework that systematically links evaluation metrics to RPA attributes and real-world applications.

3 Method

We conduct a systematic literature review to address our research question. Following prior method (Nightingale, 2009), we aim to identify relevant research papers on RPAs and provide a comprehensive summary of the literature. We selected four widely used academic databases: Google Scholar, ACM Digital Library, IEEE Xplore, and ACL Anthology. These databases encompass a broad spectrum of research across AI, human-computer interaction, and computational linguistics. Given the rapid advancements in LLM research, we included both peer-reviewed and preprint studies (e.g., from arXiv) to capture the latest developments. Below, we detail our paper selection and annotation process.

3.1 Literature Search and Screening Method

Our literature review focuses on LLM agents that role-play human behaviors, such as decision-making, reasoning, and deliberate actions. We specifically focus on studies where LLM agents demonstrate the ability to simulate human-like cognitive processes in their objectives, methodologies, or evaluation techniques. To ensure methodological rigor, we define explicit inclusion and exclusion criteria (Tab. 6 in Appendix A).

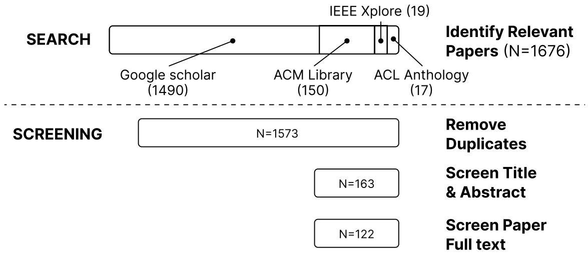

Figure 3: Screening process of literature review. We initially retrieved 1,676 papers published between 2021 and 2024, and narrowed down to 122 final selections.

The inclusion criteria require that an LLM agent in the study exhibits human-like behavior, engages in cognitive activities such as decision-making or reasoning, and operates in an open-ended task environment. We excluded studies where LLM agents primarily serve as chatbots, task-specific assistants, evaluators, or agents operating within predefined and finite action spaces. Additionally, studies focusing solely on perception-based tasks (e.g., computer vision or sensor-based autonomous driving) without cognitive simulation were also excluded.

Using this scope, we searched four databases using the query string provided in Appendix B, retrieving 1,676 papers published between January 2021 to December 2024. After removing duplicates, 1,573 unique papers remained. Two authors independently screened the paper titles and abstracts based on the inclusion criteria. If at least one author deemed a paper relevant, it proceeded to full-text screening, where two authors reviewed the paper in detail and resolved any disagreements through discussion (Fig. 3). The final set of selected studies comprised 122 publications.

3.2 Paper Annotation Method

Our team followed established open coding procedures (Brod et al., 2009) to conduct an inductive coding process to identify key themes. Three co-authors with extensive experience in LLM agents ("annotators," hereinafter) collaboratively annotated the papers on three dimensions: **agent attributes**, **task attributes**, and **evaluation metrics**.

To ensure consistency, two annotators independently annotated the same 20% of articles and then held a meeting to discuss and refine an initial set of categories for the three dimensions. After reaching a consensus, each annotator annotated half of the remaining papers and cross-validated the other half annotated by the other annotator. Once the annotations were completed, a third annotator reviewed

Table 1: Definition and examples of six agent attributes.

Agent attributes	Definition	Examples
Activity History	A record of past actions, behaviors, and engagements, including schedules, browsing history, and lifestyle choices.	Backstory, plot, weekly schedule, browsing history, social media posts, lifestyle
Belief and Value	The principles, attitudes, and ideological stances that shape an individual's perspectives and decisions.	Stances, beliefs, attitudes, values, political leaning, religion
Demographic Information	Personal identifying details such as name, age, education, career, and location.	Name, appearance, gender, age, date of birth, education, location, career, household income
Psychological Traits	Characteristics related to personality, emotions, interests, and cognitive tendencies.	Personality, hobby and interest, emotional
Skill and Expertise	The knowledge level, proficiency, and capability in specific domains or technologies.	Knowledge level, technology proficiency, skills
Social Relationships	The nature and dynamics of interactions with others, including roles, connections, and communication styles.	Parenting styles, interactions with players

the coded data and identified potential discrepancies. Any discrepancies were discussed among the annotators to ensure consistency until disagreements were resolved, ensuring reliability and validity through an iterative refinement process.

4 Survey Findings

Building on the annotated data, we systematically categorized agent attributes, task attributes, and evaluation metrics. We then present a structured RPA evaluation design guideline, outlining how to select appropriate evaluation metrics based on agent and task attributes.

4.1 Agent Attributes

We identified six categories of agent attributes, as shown in Tab. 1. *Activity history* refers to an agent's longitudinal behaviors, such as browsing history (Chen et al., 2024b) or social media activity (Navarro et al., 2024). *Belief and value* encompass the principles, attitudes, and ideological stances that shape an agent's perspectives, including political leanings (Mou et al., 2024c) or religious affiliations (Lv et al., 2024). *Demographic information* includes personal details such as name, age, education, location, career status, and household income. *Psychological traits* include an agent's personality (Jiang et al., 2023a), emotions, and cognitive tendencies (Castricato et al., 2024). *Skill and expertise* describe an agent's knowledge and proficiency in specific domains, such as technology proficiency or specialized professional skills. Lastly, *social relationships* define the social interactions, roles, and communication styles between agents, including aspects like parenting styles (Ye and Gao, 2024) or relationships between players (Ge et al., 2024).

4.2 Task Attributes

We identified seven key types of RPA downstream task attributes (Tab. 2). These tasks fall into two broad categories: those that use simulation as a research goal and those that use simulation as a tool to support specific research domains.

Among them, simulated individuals and simulated society primarily use simulation as the research goal. *Simulated individuals* involve modeling specific individuals or groups, such as end-users (Chen et al., 2024a), to study their behaviors and interactions in a controlled setting. *Simulated Society* focuses on social interactions, including cooperation (Bouzekri et al., 2024), competition (Wu et al., 2024b), and communication (Mishra et al., 2023), aiming to explore emergent social dynamics.

In contrast, the other task attributes employ simulation as a means to serve specific research domains. *Opinion dynamics* entails simulating political views (Neuberger et al., 2024), legal perspectives (Chen et al., 2024c), and social media discourse (Liu et al., 2024c) to analyze the formation and evolution of opinions. *Decision making* addresses the decision-making processes of stakeholders in investment (Sreedhar and Chilton, 2024) and public policy (Ji et al., 2024), providing insights into strategic behaviors. *Psychological experiments* explore human traits such as personality (Bose et al., 2024), ethics (Lei et al., 2024), emotions (Zhao et al., 2024), and mental health (De Duro et al., 2025), using simulated scenarios to study cognitive and behavioral responses. *Educational training* supports personalized learning by simulating teachers and learners, enhancing pedagogical approaches and adaptive education systems (Liu et al., 2024d). Finally, *writing* involves modeling readers or characters to facilitate

Table 2: Definition of seven task attributes.

Task attributes	Definition
Simulated Individuals	Simulating specific individuals or groups, such as users and participants.
Simulated Society	Simulating social interactions, such as cooperation, competition, and communication.
Opinion Dynamics	Simulating political views, legal perspectives, and social media content.
Decision Making	Simulating decision-making of stakeholders in investment, public policies, or games.
Psychological Experiments	Simulating human traits, including personality, ethics, emotions, and mental health.
Educational Training	Simulating teachers and learners to enable personalized teaching and accommodate learner needs.
Writing	Simulating readers or characters to support character development and audience understanding.

Table 3: Definitions and examples of seven evaluation metric categories.

Evaluation Metrics	Definitions	Examples
Performance	Assess RPAs' effectiveness in task execution and outcomes.	Prediction accuracy
Psychological	Measure human psychological responses to RPAs and the agents' self-awareness and emotional state.	Big Five Invertory
External Alignment	Evaluate how closely RPAs align with external ground truth or human behavior and judgments.	Alignment between model and human
Internal Consistency	Assess coherence between an RPA's predefined traits (e.g., personality), contextual expectations, and behavior.	Personality-behavior alignment
Social and Decision-Making	Analyze RPAs' social interactions and decision-making, including their effects on negotiation, societal welfare, markets, and social dynamics.	Social Conflict Count
Content and Textual	Evaluate the quality, coherence, and diversity of RPAs' text, including semantic understanding, linguistic style, and engagement.	Content similarity
Bias, Fairness, and Ethics	Assess biases, extreme or unbalanced content, or stereotyping behavior.	Factual error rate

Agent Attributes	Top 3 Agent-Oriented Metrics
Activity History	External alignment metrics, internal consistency metrics, content and textual metrics
Belief and Value	Psychological metrics, bias, fairness, and ethics metrics
Demographic Info.	Psychological metrics, internal consistency metrics, external alignment metrics
Psychological Traits	Psychological metrics, internal consistency metrics, content and textual metrics
Skill and Expertise	External alignment metrics, internal consistency metrics, content and textual metrics
Social Relationship	Psychological metrics, external alignment metrics, social and decision-making metrics

Table 4: Top 3 frequently used agent-oriented metrics for each agent attribute

Task Attributes	Top 3 Task-Oriented Metrics
Simulated Individuals	Psychological, performance, and internal consistency metrics
Simulated Society	Social and decision-making metrics, performance metrics, and psychological metrics
Opinion Dynamics	Performance metrics, external alignment metrics, and bias, fairness, and ethics metrics
Decision Making	Social and decision-making, performance, and psychological metrics
Psychological Experiment	Psychological, content and textual, and performance metrics
Educational Training	Psychological, performance, and content and textual metrics
Writing	Content and textual, psychological, and performance metrics

Table 5: Top 3 frequently used task-oriented metrics for each task attribute

character development (Benharrak et al., 2024) and audience engagement (Choi et al., 2024), contributing to storytelling and content generation research.

4.3 Agent- and Task-Oriented Metrics

We derived seven categories of evaluation metrics (Tab. 3) that are shared by agent- and task-oriented metrics despite differences in the specific metrics.

Agent-oriented metrics focus on intrinsic, task-agnostic properties that define an RPA's essential ability, such as underlying reasoning, consistency, and adaptability. These include *performance* metrics like memorization, *psychological* metrics such as emotional responses measured via entropy of valence and arousal, and *social and decision-making* metrics like social value orientation. Additionally, agent-oriented evaluations emphasize *internal consistency* metrics (e.g., consistency of information across interactions), *external alignment* metrics (e.g., hallucination detection), and *content and textual* metrics such as clarity. These evaluations ensure logical coherence, factual accuracy, and alignment with expected behavioral and cognitive frameworks, independent of any specific task.

Task-oriented metrics evaluate an RPA's effectiveness in performing specific downstream tasks, focusing on task-related aspects such as accuracy, consistency, social impact, and ethical considerations. *Performance* measures how well RPAs exe-

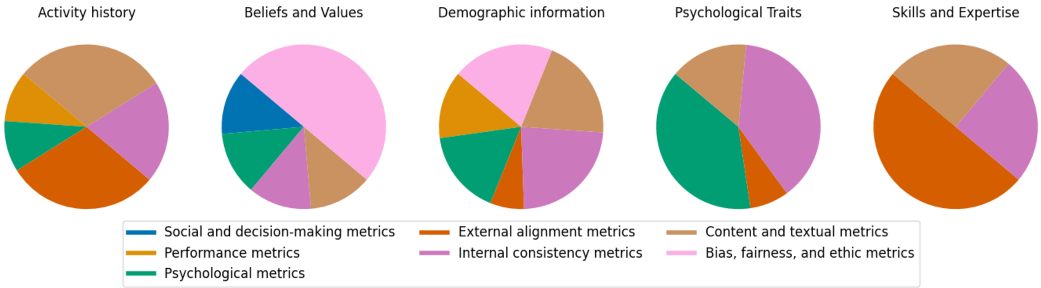

Figure 4: Proportional distribution of agent-oriented metrics across different agent attributes.

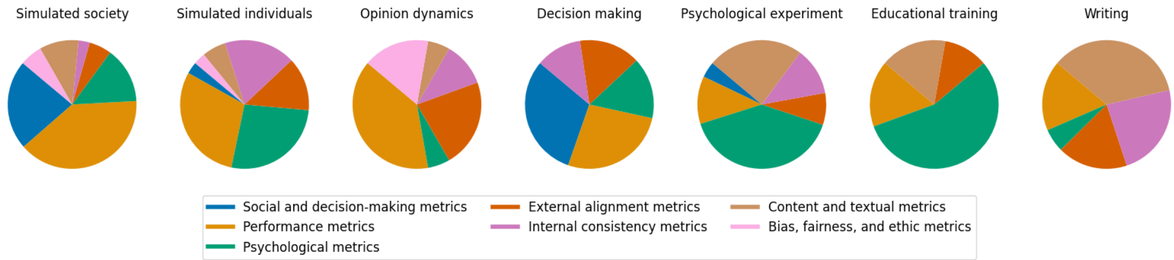

Figure 5: Proportional distribution of task-oriented metrics across different task attributes.

cute designated tasks, such as prediction accuracy. *Psychological* metrics assess human psychological responses to RPAs, including self-awareness and emotional states; for example, the Big Five Inventory. *External alignment* evaluates how closely RPAs align with external ground truth or human behavior; for instance, alignment between model and human. *Internal consistency* ensures coherence between an RPA's predefined traits, contextual expectations, and behavior; for example, personality-behavior alignment. *Social and decision-making* metrics analyze RPAs' influence on negotiation, societal welfare, and social dynamics; for instance, the social conflict count. *Content and textual quality* focuses on the coherence, linguistic style, and engagement of RPAs' generated text, such as content similarity. Lastly, *bias, fairness, and ethics* metrics examine biases, extreme content, or stereotypes; for instance, the factual error rate. Together, these seven metrics provide a comprehensive framework for evaluating RPAs' task performance and broader impact. To clarify how these metrics are adapted and implemented in practice, we compiled concrete examples across different use cases (see Appendix Table 11). For instance, the Big Five Inventory has been used in psychological experiments, educational training, and simulated societies, with variations in the number of items, rating targets (self vs. other), and timing. In contrast, metrics such as "response accuracy" appear more narrowly applied in simulated societies, and are implemented via expert judgment or through scenario-based behavioral probes. These examples highlight not only the flexibility of certain metrics but also the importance of aligning metric design with both agent capabilities and task structure.

4.4 RPA Evaluation Design Guideline

Building on our classification of agent attributes, task attributes, and evaluation metrics, we observed a recurring distinction between **agent-oriented** and **task-oriented** design and evaluation. This distinction revealed consistent associations between agent/task attributes and the evaluation metrics used. We interpret these associations through a layered theoretical lens. At the individual level, Goffman's dramaturgical theory (Goffman, 2023) frames agent attributes (e.g., personality, beliefs) as role-defining traits and task attributes as performance contexts, supporting the use of metrics that assess both internal coherence (e.g., internal consistency, psychological fidelity) and contextual fit (e.g., external alignment, task performance). Agent-Based Modeling (ABM) theory (Epstein, 1999) further explains how macro-level evaluation patterns can emerge from repeated agent-task pairings, providing theoretical support for our data-driven synthesis of design–evaluation couplings. These insights inform the development of our systematic guidelines for selecting evaluation metrics.

Step 1. Selecting Agent-Oriented Metrics Based on Agent Attributes We analyzed the distribution of agent attributes and agent-oriented metrics, as illustrated in Fig. 4. Our analysis reveals that, for each agent attribute, the top three categories of agent-oriented metrics account for the majority of all metric types. Based on this observation, our first guideline recommends selecting agent-oriented metrics according to agent attributes. Specifically, we suggest referring to Tab. 7 to identify the top three corresponding metrics. For instance, for Activity History, the recommended metrics are external alignment, internal consistency, and content and textual metrics. Likewise, for Beliefs and Values, the most relevant choices are psychological metrics and bias, fairness, and ethics metrics. In particular, there are no established agent-oriented evaluation metrics for social relationships. Based on Social Exchange Theory (Cropanzano and Mitchell, 2005), which explains relationship formation through reciprocal interactions and resource exchanges, we propose assessing social relationships with psychological metrics, external alignment metrics, and social and decision-making metrics.

Step 2: Selecting Task-Oriented Metrics Based on Task Attributes Additionally, we analyzed the distribution of task attributes and task-oriented metrics, as shown in Fig. 5. Consistent with our previous findings, we observed that for each category of task attributes, the top three task-oriented metrics account for the vast majority of all metrics. Based on this, our second guideline recommends selecting task-oriented metrics according to task attributes. Specifically, we suggest referring to Tab. 8 to identify the top three corresponding metrics. For instance, for the *Simulated Society* task, the recommended metrics are social and decision-making, performance, and psychological metrics. Similarly, for the *Opinion Dynamics* task, the most relevant choices are performance, external alignment, bias, fairness, and ethics metrics.

However, these two steps should not be treated as one-time decisions. As the agent design process evolves, evaluation results may prompt adjustments to the attributes of the agent and the task, thereby influencing the selection of evaluation metrics. Therefore, this two-step evaluation guideline should be used iteratively to ensure that the evaluation remains adaptive to changing agent capabilities and task requirements. This iterative approach enhances the reliability, relevance, and robustness of RPA evaluation experiments.

5 Case Study: How to Use RPA Design Guideline to Select Evaluation Metrics

We present **two case studies** to illustrate how our evaluation guidelines can be applied in practice. These examples are not intended to demonstrate superiority but to show the feasibility of aligning evaluation metrics with agent and task attributes, and how such alignment is reflected in existing studies. By adopting the perspective of the original authors, we compare the evaluation outcomes resulting from adhering to or deviating from the RPA evaluation guidelines.

5.1 A Well-Aligned Example: *Generative Agents: Interactive Simulacra of Human Behavior*

Park et al. (2023) designed agents with cognitive modules that included memory, planning, and reflection, along with demographic information, action history, and social relationships. Their evaluation approach demonstrates a strong alignment with both agent and task attributes, as outlined in our guideline.

For agent-oriented metrics, they selected five types that correspond to the top categories identified in our survey (see Fig. 4): Self-knowledge (Content/textual, Internal consistency), Memory and Plans (Internal consistency), Reactions (External alignment), and Reflections (Psychological). These metrics were tightly coupled with the agent's internal architecture. For example, they evaluated whether agents could recall and respond consistently: *"Generative agents equipped with a complete memory module are capable of recalling past experiences and answering questions in a manner that is consistent with their self-knowledge across a wide range of contexts."*

At the task level, their simulated society scenario guided the selection of four task-oriented metrics: Response accuracy (Performance), Relationship formation (Psychological), Information diffusion, and Coordination (Social and decision-making)—aligned with the dominant metric categories for simulated society tasks (see Fig. 5). These metrics enabled the evaluation of emergent behaviors such as event attendance and information propagation: *"The number of agents who knew about Sam's mayoral candidacy increased from one*

(4%) to eight (32%)... the agent community formed new relationships... network density increasing from 0.167 to 0.74."

By systematically linking metrics to both the agents' cognitive design and the societal dynamics of the task, this study exemplifies the practical application of our evaluation guideline.

5.2 A Misaligned Example: *A Generative Social World for Embodied AI*

As illustrated in Appendix E Fig. 9, this ICLR submission proposed agents with rich attributes—personas, social relationships, and behavioral histories—for tasks such as route planning and election campaigning. However, their evaluation choices diverged significantly from what our framework would suggest.

Although the agent design included psychological and social elements, the evaluation excluded agent-oriented metrics such as those assessing psychological realism or persona consistency. One reviewer commented: *"There is a lack of details on how social relationships are established from the characters' profiles... Reference to 'open-world knowledge' does not appear sufficient in light of the vast body of work dedicated to persona definition with LLMs."*

On the task side, the study focused on opinion dynamics and decision-making, which typically call for metrics like Psychological, Social and decision-making, External alignment, and Ethics-related measures. Yet the evaluation was limited to only Arrival rate, Time, and Campaign strategy alignment. This omission resulted in additional reviewer criticism: *"Only results for Route Planning are included; it would be nice to see results for the Election Campaign as well." "The election campaign environment is more about interactions with other people—not something that immediately requires a 3D environment."*

These critiques illustrate the very type of design–evaluation misalignment that our framework is intended to prevent. By failing to match their metrics with the agent and task characteristics they had modeled, the study limited the interpretability and credibility of its results—despite promising agent designs.

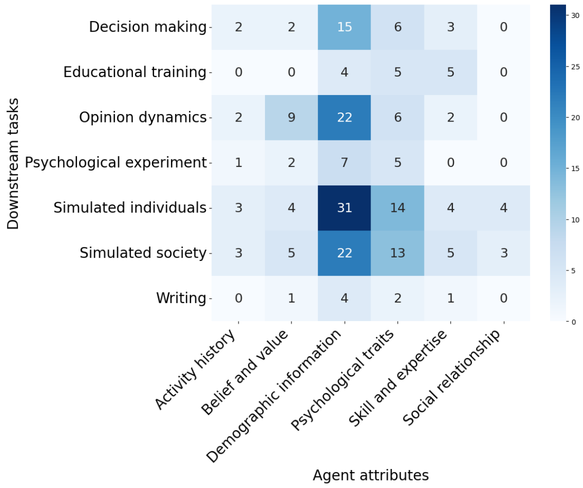

Figure 6: Relationships between agent attributes and downstream tasks. The numbers in the heatmap represent the paper counts.

6 Relationships Between Agent Attributes and Downstream Tasks

Both agent attributes and downstream task attributes play a crucial role in selecting appropriate RPA evaluation metrics. Researchers predefine these factors when designing and evaluating RPAs, yet their interrelation remains an open question. In this section, we analyze how agent attributes correspond to different downstream tasks, uncovering several recurring patterns (Fig. 6).

Demographic information and psychological traits are fundamental across all downstream tasks. Whether in decision-making, opinion dynamics, or simulated environments, these attributes consistently shape RPA design. As shown in Fig. 6, they are the most frequently incorporated factors, underscoring their central role in modeling agent behavior across diverse applications.

For tasks where simulation itself is the primary objective, such as Simulated Individuals and Simulated Society, the selection of agent attributes becomes broader. In addition to demographic and psychological factors, these tasks frequently incorporate skills, expertise, and social relationships, reflecting the need for richer agent representations to capture complex social and individual interactions. By contrast, tasks that use simulation as a means to study specific research fields tend to prioritize certain agent attributes. For instance, in Opinion Dynamics, beliefs and values play a distinctive role, as they directly influence how agents interact and form opinions. Similarly, tasks related to Educational Training and Writing exhibit a different pattern, emphasizing skills and expertise over broad demographic or psychological considerations.

In contrast, attributes such as activity history and social relationships receive significantly less emphasis across tasks. This raises a question: is their impact inherently limited, or are they simply underexplored in current RPA applications?

Overall, these findings highlight the nuanced interplay between agent attributes and downstream tasks. While demographic information and psychological traits are universally relevant, attributes like beliefs and values gain importance in specific contexts. At the same time, the relative absence of activity history and social relationships in current evaluations presents an open research question, particularly in scenarios requiring long-term modeling and complex social interactions.

7 Discussion

7.1 RPA as a Socio-Technical System

Our analysis in Section 6 explored how agent attributes are distributed across task types, a dimension often overlooked in RPA design discussions. Although these attributes are usually predefined, they reflect deeper modeling assumptions that shape how RPAs behave. By identifying patterns such as the frequent use of demographic and psychological traits, and the relative underuse of social relationships, we surface important design trends and open questions. Rather than offering definitive prescriptions, this analysis is intended to support future work in interrogating why certain attributes are emphasized over others, and how they relate to evaluation choices.

More broadly, RPAs should be viewed not just as algorithmic components but as socio-technical systems embedded in context. Their design has implications beyond performance, including ethical, cognitive, and societal dimensions. From psychological simulations to social modeling, RPAs hold promise as scalable, interactive tools—but only if their assumptions, behaviors, and roles are made explicit and reflect the systems they represent. This calls for iterative, human-centered design approaches that account for diversity in user expectations, cultural contexts, and domain constraints.

7.2 Designing the RPA Persona

RPAs' flexibility allows them to simulate a wide range of personas across tasks and domains. Yet designing these personas is nontrivial: agent traits must align with both their intended role and their surrounding context. Intrinsic characteristics such as personality, values, and domain expertise should be selected with the application's goals in mind, e.g., emphasizing empathy in therapeutic agents or strategic reasoning in policy simulations. Contextual grounding is equally important. Task-specific environments shape how agents should behave and what behaviors are deemed credible. A caregiving agent in a healthcare simulation, for example, must balance emotional expressiveness with adherence to clinical norms. Without sufficient contextual fidelity, agents risk being perceived as implausible or ineffective. Future research should explore how to scaffold personas through modular, context-aware components that support both behavioral consistency and scenario adaptability.

7.3 Challenges in Evaluating RPAs

RPAs' diversity and adaptability make unified evaluation inherently difficult. As our literature-based synthesis shows, agent- and task-oriented metrics vary significantly by application. No single set of metrics can capture all relevant qualities across domains, use cases, or user goals. For example, emotional plausibility is critical in psychological studies but secondary in economic modeling. Our proposed evaluation guideline offers a structured starting point, rooted in observed design–evaluation pairings. However, these should not be interpreted as prescriptive standards. Cross-task and cross-domain evaluation remains a core challenge due to inconsistent metric definitions, task framings, and agent behaviors. Addressing this will require adaptive, multi-dimensional evaluation strategies that incorporate not only technical performance but also user-centered concerns, normative judgments, and long-term behavioral consistency.

8 Conclusion

RPA evaluation lacks consistency due to varying tasks, domains, and agent attributes. Our systematic review of $1,676$ papers reveals that task-specific requirements shape agent attributes, while both task characteristics and agent design influence evaluation metrics. By identifying these interdependencies, we propose guidelines to enhance RPA assessment reliability, contributing to a more structured and systematic evaluation framework.

Limitations

RPAs are rapidly evolving and have widespread applications across various domains. While we aim to comprehensively review existing literature, we acknowledge certain limitations in our scope. First, our review may not encompass all variations of RPA evaluation approaches across different application domains. Second, new research published after December 2024 is not included in our analysis. As a result, our work does not claim to exhaustively cover all potential evaluation metrics. Instead, our goal is to provide a structured framework and actionable guidelines to help future researchers design more systematic and consistent RPA evaluations, even as the field continues to evolve.

Ethics Statement

Our work focuses on summarizing and analyzing the evaluation of RPAs, which we believe will be valuable to researchers in AI, HCI, and related fields such as psychological simulation, educational simulation, and economic simulation. We have taken care to ensure that this survey remains objective and balanced, neither overestimating nor underestimating trends. We do not anticipate any ethical concerns that arise from the research presented in this paper.

References

Ana Antunes, Joana Campos, Manuel Guimarães, João Dias, and Pedro A. Santos. 2023. Prompting for socially intelligent agents with chatgpt. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, IVA '23, New York, NY, USA. Association for Computing Machinery.

Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2024. Plurals: A system for guiding llms via simulated social ensembles. *Preprint*, arXiv:2409.17213.

Sarah Assaf and Timothy Lynar. 2024. Human testing using large-language models: Experimental research and the development of a security awareness controls framework.

Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined ai personas for on-demand feedback generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Ritwik Bose, Mattson Ogg, Michael Wolmetz, and Christopher Ratto. 2024. Assessing behavioral alignment of personality-driven generative agents in social dilemma games. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Elodie Bouzekri, Pascal E Fortin, and Jeremy R Cooperstock. 2024. Chatgpt, tell me more about pilots' opinion on automation. In *2024 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pages 99–106. IEEE.

Meryl Brod, Laura E Tesler, and Torsten L Christensen. 2009. Qualitative research and content validity: developing best practices based on science and experience. *Quality of life research*, 18:1263–1278.

Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, et al. 2024. Digital life project: Autonomous 3d characters with social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 582–592.

Gian Maria Campedelli, Nicolò Penzo, Massimo Stefan, Roberto Dessì, Marco Guerini, Bruno Lepri, and Jacopo Staiano. 2024. I want to break free! persuasion and anti-social behavior of llms in multi-agent settings with social hierarchy. *Preprint*, arXiv:2410.07109.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2024. Persona: A reproducible testbed for pluralistic alignment. *Preprint*, arXiv:2407.17387.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *ArXiv*, abs/2308.07201.

Chaoran Chen, Leyang Li, Luke Cao, Yanfang Ye, Tianshi Li, Yaxing Yao, and Toby Jia-jun Li. 2024a. Why am i seeing this: Democratizing end user auditing for online content recommendations. *arXiv preprint arXiv:2410.04917*.

Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-Jun Li. 2024b. An empathy-based sandbox approach to bridge the privacy gap among attitudes, goals, knowledge, and behaviors. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. 2024c. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *arXiv preprint arXiv:2408.08089*.

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. 2024d. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024e. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *Preprint*, arXiv:2308.10848.

Xuzheng Chen, Zhangshiyin, and Guojie Song. 2024f. Towards humanoid: Value-driven agent modeling based on large language models. In *NeurIPS 2024 Workshop on Open-World Agents*.

Haocong Cheng, Si Chen, Christopher Perdriau, and Yun Huang. 2024. Llm-powered ai tutors with personas for d/deaf and hard-of-hearing online learners. *ArXiv*, abs/2411.09873.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Yizhou Chi, Lingjun Mao, and Zineng Tang. 2024. Amongagents: Evaluating large language models in the interactive text-based social deduction game. *Preprint*, arXiv:2407.16521.

Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2024. Proxona: Leveraging llm-driven personas to enhance creators' understanding of their audience. *arXiv preprint arXiv:2408.10937*.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023a. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*.

Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2023b. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents.

Russell Cropanzano and Marie S Mitchell. 2005. Social exchange theory: An interdisciplinary review. *Journal of management*, 31(6):874–900.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. *arXiv preprint arXiv:2408.04203*.

Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. 2025. Introducing counsellme: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans. *Emerging Trends in Drugs, Addictions, and Health*, page 100170.

Joost C. F. de Winter, Tom Driessen, and Dimitra Dodou. 2024. The use of chatgpt for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences*.

Joshua M Epstein. 1999. Agent-based computational models and generative social science. *Complexity*, 4(5):41–60.

Jingchao Fang, Nikos Arechiga, Keiichi Namikoshi, Nayeli Bravo, Candice Hogan, and David A Shamma. 2024. On llm wizards: Identifying large language models' behaviors for wizard of oz experiments. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, pages 1–11.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *Preprint*, arXiv:2402.02896.

Chen Gao, Xiaochong Lan, Zhi jie Lu, Jinzhu Mao, Jing Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *ArXiv*, abs/2307.14984.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

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

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *Preprint*, arXiv:2410.07991.

Erving Goffman. 2023. The presentation of self in everyday life. In *Social theory re-wired*, pages 450–459. Routledge.

Zhouhong Gu, Xiaoxuan Zhu, Haoran Guo, Lin Zhang, Yin Cai, Hao Shen, Jiangjie Chen, Zheyu Ye, Yifei Dai, Yan Gao, Yao Hu, Hongwei Feng, and Yanghua Xiao. 2024. Agentgroupchat: An interactive group chat simulacra for better eliciting emergent behavior. *Preprint*, arXiv:2403.13433.

George Gui and Olivier Toubia. 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. *ArXiv*, abs/2312.15524.

models. In *Advances in Neural Information Processing Systems*, volume 36, pages 10622–10643. Curran Associates, Inc.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Juhye Ha, Hyeon Jeon, DaEun Han, Jinwook Seo, and Changhoon Oh. 2024. Clochat: Understanding how people customize, interact, and experience personas in large language models. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024a. Agentscourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Conference on Empirical Methods in Natural Language Processing*.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024b. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416, Miami, Florida, USA. Association for Computational Linguistics.

Zihong He and Changwang Zhang. 2024. Afspp: Agent framework for shaping preference and personality with large language models. *ArXiv*, abs/2401.02870.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.

Yin Jou Huang and Rafik Hadfi. 2024. How personality traits influence negotiation outcomes? a simulation based on large language models. *arXiv preprint arXiv:2407.11549*.

Jiarui Ji, Yang Li, Hongtao Liu, Zhicheng Du, Zhewei Wei, Weiran Shen, Qi Qi, and Yankai Lin. 2024. Srapagent: Simulating and optimizing scarce resource allocation policy with llm-based agent. *arXiv preprint arXiv:2410.14152*.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context. *ArXiv*, abs/2406.05972.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023a. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 10622–10643. Curran Associates, Inc.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023b. Personallm: Investigating the ability of large language models to express personality traits. In *NAACL-HLT*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Hyoungwook Jin, Seonghee Lee, Hyun Joon Shin, and Juho Kim. 2023. Teach ai how to code: Using large language models as teachable agents for programming education. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. In *Conference on Empirical Methods in Natural Language Processing*.

Tianjie Ju, Yiting Wang, Xinbei Ma, Pengzhou Cheng, Haodong Zhao, Yulong Wang, Lifeng Liu, Jian Xie, Zhuosheng Zhang, and Gongshen Liu. 2024. Flooding spread of manipulated knowledge in llm-based multi-agent communities. *ArXiv*, abs/2407.07791.

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe agents: Generative agents for low-cost real-time social interactions. *Preprint*, arXiv:2310.02172.

Mahammed Kamruzzaman and Gene Louis Kim. 2024. Exploring changes in nation perception with nationality-assigned personas in llms. *Preprint*, arXiv:2406.13993.

Ping Fan Ke and Ka Chung Ng. 2024. Human-ai synergy in survey development: Implications from large language models in business and research. *ACM Transactions on Management Information Systems*.

Kyusik Kim, Hyeonseok Jeon, Jeongwoo Ryu, and Bongwon Suh. 2024. Will llms sink or swim? exploring decision-making under pressure. In *Conference on Empirical Methods in Natural Language Processing*.

Kunyao Lan, Bingrui Jin, Zichen Zhu, Siyuan Chen, Shu Zhang, Kenny Q. Zhu, and Mengyue Wu. 2024. Depression diagnosis dialogue simulation: Self-improving psychiatrist with tertiary memory. *Preprint*, arXiv:2409.15084.

Unggi Lee, Sanghyeok Lee, Junbo Koh, Yeil Jeong, Haewon Jung, Gyuri Byun, Jewoong Moon, Jieun Lim, and † HyeoncheolKim. Generative agent for teacher training: Designing educational problem-solving simulations with large language model-based agents for pre-service teachers.

Yu Lei, Hao Liu, Chengxing Xie, Songjia Liu, Zhiyu Yin, Canyu Chen, Guohao Li, Philip Torr, and Zhen Wu. 2024. Fairmindsim: Alignment of behavior, emotion, and belief in humans and llm agents amid ethical dilemmas. *arXiv preprint arXiv:2410.10398*.

Yan Leng and Yuan Yuan. 2024. Do llm agents exhibit social behavior? *Preprint*, arXiv:2312.15198.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024a. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*.

Jiale Li, Jiayang Li, Jiahao Chen, Yifan Li, Shijie Wang, Hugo Zhou, Minjun Ye, and Yunsheng Su. 2024b. Evolving agents: Interactive simulation of dynamic and diverse human personalities. *ArXiv*, abs/2404.02718.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. 2024c. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024d. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536.

Sha Li, Revanth Gangi Reddy, Khanh Duy Nguyen, Qingyun Wang, May Fung, Chi Han, Jiawei Han, Kartik Natarajan, Clare R. Voss, and Heng Ji. 2024e. Schema-guided culture-aware complex event simulation with multi-agent role-play. *ArXiv*, abs/2410.18935.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023a. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *ArXiv*, abs/2310.06500.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023b. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *Preprint*, arXiv:2310.06500.

Xiaoyu Lin, Xinkai Yu, Ankit Aich, Salvatore Giorgi, and Lyle Ungar. 2024. Diversedialogue: A methodology for designing chatbots with human-like diversity. *Preprint*, arXiv:2409.00262.

Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Noah Wang, Jian Yang, JiakaiWang, Hongcheng Guo, Z.Y. Peng, Ge Zhang, Jiayi Tian, Xingyuan Bu, Ke Xu, Wenge Rong, Junran Peng, and Zhaoxiang Zhang. 2024a. Roleagent: Building, interacting, and benchmarking high-quality role-playing agents from scripts. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ryan Liu, Howard Yen, Raja Marjieh, Thomas L. Griffiths, and Ranjay Krishna. 2023. Improving interpersonal communication by simulating audiences with language models. *Preprint*, arXiv:2311.00687.

Tianjian Liu, Hongzheng Zhao, Yuheng Liu, Xingbo Wang, and Zhenhui Peng. 2024b. Compeer: A generative conversational agent for proactive peer support. In *ACM Symposium on User Interface Software and Technology*.

Xuan Liu, Jie Zhang, Song Guo, Haoyang Shang, Chengxu Yang, and Quanyan Zhu. 2025. Exploring prosocial irrationality for llm agents: A social cognition view. *Preprint*, arXiv:2405.14744.

Yuhan Liu, Zirui Song, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2024c. From a tiny slip to a giant leap: An llm-based simulation for fake news evolution. *arXiv preprint arXiv:2410.19064*.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024d. Personality-aware student simulation for conversational intelligent tutoring systems. In *Conference on Empirical Methods in Natural Language Processing*.

Yaojia Lv, Haojie Pan, Zekun Wang, Jiafeng Liang, Yuanxing Liu, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. 2024. Coggpt: Unleashing the power of cognitive dynamics on large language models. *arXiv preprint arXiv:2401.08438*.

Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *Plos one*, 19(3):e0298522.

Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023. e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967, Singapore. Association for Computational Linguistics.

Konstantinos Mitsopoulos, Ritwik Bose, Brodie Mather, Archna Bhatia, Kevin Gluck, Bonnie Dorr, Christian Lebiere, and Peter Pirolli. 2024. Psychologically-valid generative agents: A novel approach to agent-based modeling in social sciences. *Proceedings of the AAAI Symposium Series*.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. *Preprint*, arXiv:2407.06576.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. 2024a. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. 2024b. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *Preprint*, arXiv:2410.19346.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024c. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Annual Meeting of the Association for Computational Linguistics*.

Sonia K. Murthy, Tomer Ullman, and Jennifer Hu. 2024. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. *Preprint*, arXiv:2411.04427.

Keiichi Namikoshi, Alexandre L. S. Filipowicz, David A. Shamma, Rumen Iliev, Candice Hogan, and Nikos Aréchiga. 2024. Using llms to model the beliefs and preferences of targeted populations. *ArXiv*, abs/2403.20252.

Alejandro Leonardo Garc'ia Navarro, Nataliia Koneva, Alfonso S'anchez-Maci'an, Jos'e Alberto Hern'andez, and Manuel Goyanes. 2024. Designing reliable experiments with generative agent-based modeling: A comprehensive guide using concordia by google deepmind. *ArXiv*, abs/2411.07038.

Shlomo Neuberger, Niv Eckhaus, Uri Berger, Amir Taubenfeld, Gabriel Stanovsky, and Ariel Goldstein. 2024. Sauce: Synchronous and asynchronous user-customizable environment for multi-agent llm interaction. *arXiv preprint arXiv:2411.03397*.

Alison Nightingale. 2009. A guide to systematic literature reviews. *Surgery (Oxford)*, 27(9):381–384.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Pat Pataranutaporn, Kavin Winson, Peggy Yin, Auttasak Lapapirojn, Pichayoot Ouppaphan, Monchai Lertsutthiwong, Pattie Maes, and Hal E. Hershfield. 2024. Future you: A conversation with an ai-generated future self reduces anxiety, negative emotions, and increases future self-continuity. *ArXiv*, abs/2405.12514.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Preprint*, arXiv:2404.16698.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *Preprint*, arXiv:2408.15787.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Ruiyang Ren, Peng Qiu, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Huaqin Wu, Ji-Rong Wen, and Haifeng Wang. 2024a. Bases: Large-scale web search user simulation with large language model based agents. *ArXiv*, abs/2402.17505.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. 2024b. Emergence of social norms in generative agent societies: Principles and architecture. *Preprint*, arXiv:2403.08251.

Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. 2024. Y social: an llm-powered social media digital twin. *arXiv preprint arXiv:2408.00818*.

Joni O. Salminen, João M. Santos, Soon gyo Jung, and Bernard J. Jansen. 2024. Picturing the fictitious person: An exploratory study on the effect of images on user perceptions of ai-generated personas. *Computers in Human Behavior: Artificial Humans*.

Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. 2024. Generating personas using llms and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Lipton, and J Zico Kolter. 2025. Rethinking llm memorization through the lens of adversarial compression. *Advances in Neural Information Processing Systems*, 37:56244–56267.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Jinxin Shi, Jiabao Zhao, Yilei Wang, Xingjiao Wu, Jiawen Li, and Liangbo He. 2023. Cgmi: Configurable general multi-agent interaction framework. *ArXiv*, abs/2308.12503.

Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding human-ai workflows for generating personas. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 757–781, New York, NY, USA. Association for Computing Machinery.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009.

Sinan Sonlu, Bennie Bendiksen, Funda Durupinar, and Uğur Güdükbay. 2024. The effects of embodiment and personality expression on learning in llm-based educational agents. *ArXiv*, abs/2407.10993.

Karthik Sreedhar and Lydia Chilton. 2024. Simulating human strategic behavior: Comparing single and multi-agent llms. *arXiv preprint arXiv:2402.08189*.

Libo Sun, Siyuan Wang, Xuanjing Huang, and Zhongyu Wei. 2024. Identity-driven hierarchical role-playing agents. *Preprint*, arXiv:2407.19412.

Eduardo Ryô Tamaki and Levente Littvay. 2024. Chrono-sampling: Generative ai enabled time machine for public opinion data collection.

Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Erabal: Enhancing role-playing agents through boundary-aware learning. *Preprint*, arXiv:2409.14710.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, Miami, Florida, USA. Association for Computational Linguistics.

Jesus-Pablo Toledo-Zucco, Denis Matignon, and Charles Poussot-Vassal. 2024. Scattering-passive structure-preserving finite element method for the boundary controlled transport equation with a moving mesh. *Preprint*, arXiv:2402.01232.

Haley Triem and Ying Ding. 2024. “tipping the balance”: Human intervention in large language model multi-agent debate. *Proceedings of the Association for Information Science and Technology*, 61(1):361–373.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Deepank Verma, Olaf Mumm, and Vanessa Miriam Carlow. 2023. Generative agents in the streets: Exploring the use of large language models (llms) in collecting urban perceptions. *ArXiv*, abs/2312.13126.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *arXiv preprint arXiv:2305.13160*.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji rong Wen. 2023b. User behavior simulation with large language model based agents.

Qian Wang, Tianyu Wang, Qinbin Li, Jingsheng Liang, and Bingsheng He. 2024a. Megaagent: A practical framework for autonomous cooperation in large-scale llm agent systems. *Preprint*, arXiv:2408.09955.

Qian Wang, Jiaying Wu, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025. What limits llm-based human simulation: Llms or our design? *arXiv preprint arXiv:2501.08579*.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024b. Deem: Dynamic experienced expert modeling for stance detection. *ArXiv*, abs/2402.15264.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024c. 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.

Yi Wang, Qian Zhou, and David Ledo. 2024d. Storyverse: Towards co-authoring dynamic plot with llm-based character simulation via narrative planning. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, FDG '24, New York, NY, USA. Association for Computing Machinery.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. 2024e. Connecting the dots: Collaborative fine-tuning for black-box vision-language models. *arXiv preprint arXiv:2402.04050*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024f. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Zhenyu Wang, Yi Xu, Dequan Wang, Lingfeng Zhou, and Yiqi Zhou. 2024g. Intelligent computing social modeling and methodological innovations in political science in the era of large language models. *ArXiv*, abs/2410.16301.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Jiale Hong, Hai Zhao, and Min Zhang. 2024a. From role-play to drama-interaction: An llm solution. *arXiv preprint arXiv:2405.14231*.

Zengqing Wu, Shuyuan Zheng, Qianying Liu, Xu Han, Brian Inhyuk Kwon, Makoto Onizuka, Shaojie Tang, Run Peng, and Chuan Xiao. 2024b. Shall we talk: Exploring spontaneous collaborations of competing llm agents. *arXiv preprint arXiv:2402.12327*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and G. Li. 2024a. Can large language model agents simulate human trust behaviors? *ArXiv*, abs/2402.04559.

Qiuejie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang. 2024b. Human simulacra: Benchmarking the personification of large language models. *Preprint*, arXiv:2402.18180.

Zihan Yan, Yaohong Xiang, and Yun Huang. 2024. Social life simulation for non-cognitive skills learning. *ArXiv*, abs/2405.00273.

Frank Tian-fang Ye and Xiaozi Gao. 2024. Simulating family conversations using llms: Demonstration of parenting styles. *arXiv preprint arXiv:2403.06144*.

Leo Yeykelis, Kaavya Pichai, James J. Cummings, and Byron Reeves. 2024. Using large language models to create ai personas for replication and prediction of media effects: An empirical test of 133 published experimental research findings. *Preprint*, arXiv:2408.16073.

Chenxiao Yu, Zhaotian Weng, Yuangang Li, Zheng Li, Xiyang Hu, and Yue Zhao. 2024. Towards more accurate us presidential election via multi-step reasoning with large language models. *ArXiv*, abs/2411.03321.

Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2024. Persllm: A personified training approach for large language models. *arXiv preprint arXiv:2407.12393*.

Dong Zhang, Zhaowei Li, Pengyu Wang, Xin Zhang, Yaqian Zhou, and Xipeng Qiu. 2024a. Speechagents: Human-communication simulation with multi-modal multi-agent systems. *ArXiv*, abs/2401.03945.

Jintian Zhang, Xin Xu, Ruibo Liu, and Shumin Deng. 2023a. Exploring collaboration mechanisms for llm agents: A social psychology view. *ArXiv*, abs/2310.02124.

Long Zhang, Meng Zhang, Wei Lin Wang, and Yu Luo. 2025. Simulation as reality? the effectiveness of llm-generated data in open-ended question assessment. *arXiv preprint arXiv:2502.06371*.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2024b. Self-emotion blended dialogue generation in social simulation agents. *Preprint*, arXiv:2408.01633.

Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024c. See widely, think wisely: Toward designing a generative multi-agent system to burst filter bubbles. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Zhaowei Zhang, Ceyao Zhang, Nian Liu, Siyuan Qi, Ziqi Rong, Song-Chun Zhu, Shuguang Cui, and Yaodong Yang. 2023b. Heterogeneous value alignment evaluation for large language models. *arXiv preprint arXiv:2305.17147*.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Wang Jian, Dandan Liang, et al. 2024. Esc-eval: Evaluating emotion support conversations in large language models. *arXiv preprint arXiv:2406.14952*.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2023. Competeai: Understanding the competition dynamics of large language model-based agents. In *International Conference on Machine Learning*.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024a. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *ArXiv*, abs/2403.05020.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. Sotopia: Interactive evaluation for social intelligence in language agents. *Preprint*, arXiv:2310.11667.

Table 6: Inclusion and exclusion criteria.

Inclusion Criteria (IC)	
IC-1	The LLM agents in the paper simulate humanoid behavior with implicit personality (e.g., preference and behavior pattern) or explicit personality (e.g., emotion or characteristics).
IC-2	The LLM agents in the paper have cognitive activities such as decision-making, reasoning, and planning.
IC-3	The LLM agents in the paper are capable of completing complicated and general tasks.
IC-4	The LLM agents' action set in the paper is neither predefined nor finite.
Exclusion Criteria (EC)	
EC-1	The study does not employ LLM agents for simulation purposes but rather uses them as chatbots, task-specific agents, or evaluators.
EC-2	The paper's research objectives, methodologies, and evaluations are not focused on simulating human-like behavior with LLM agents, but rather on optimizing LLM algorithms.
EC-3	The study primarily investigates the perception or action capabilities of LLM agents without simulating the cognitive process.
EC-4	The LLM agents are restricted to handling specific, close-ended tasks.
EC-5	The LLM agents' actions are either predefined or limited.

A Inclusion and Exclusion Criteria

We summarize the inclusion and exclusion criteria in Table 6. Briefly, the **Inclusion Criteria (IC)** ensure that the reviewed studies focus on LLM agents exhibiting human-like behavior—either implicitly (e.g., preference or behavioral patterns) or explicitly (e.g., emotions or personality)—along with key cognitive processes such as reasoning and decision-making. Moreover, an open-ended action space and the capacity to tackle multifaceted tasks are essential attributes for inclusion.

By contrast, the **Exclusion Criteria (EC)** eliminate studies employing LLMs purely as chatbots, single-purpose systems, or evaluation tools, rather than as agents mimicking human cognition. Likewise, if the LLM agents are restricted to fixed, close-ended tasks or limited to algorithmic optimization without simulating cognitive processes, they fall outside the scope of this work.

B Query String

We employed the following query to guide our literature retrieval process:

```
("large language model" OR LLM)
AND (agent OR persona OR "human
digital twin" OR simulacra) AND
(simulat* OR generat* OR eval*)
AND "human behavior" AND cognit*
```

This query was designed to capture a broad spectrum of studies on large language models that simulate or replicate human-like behavior. It combines keywords related to LLM agents (*LLM*, *persona*, *simulacra*), their capabilities (*simulat**, *generat**, *eval**), and the focus on cognitively grounded human behavior (*cognit**). This ensures that the resulting literature is relevant to our exploration of how LLM-based systems can mimic or exhibit human-like cognition and behavior patterns.

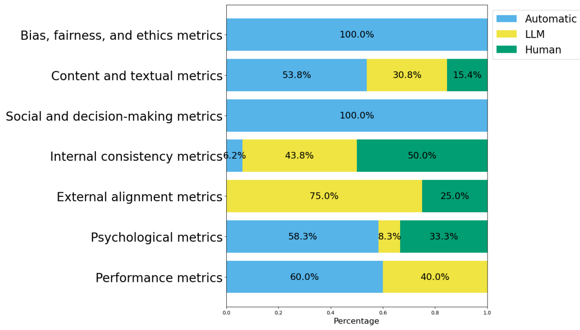

Figure 7: Usage ratio of evaluation approaches for each category of agent-oriented metrics.

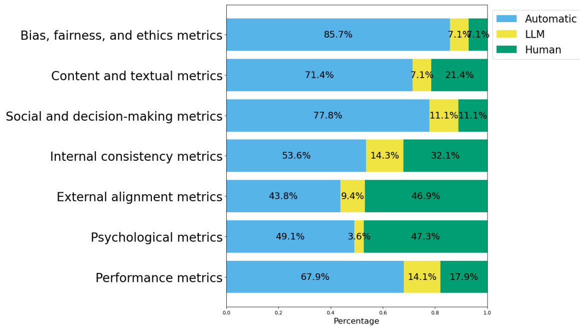

Figure 8: Usage ratio of evaluation approaches for each category of task-oriented metrics.

C Evaluation Approach Usage for Agent- and Task-Oriented Metrics

We present a breakdown of evaluation approach usage by agent-oriented metrics (Fig. 7) and task-oriented metrics (Fig. 8).

D Top Three Metrics for Agent and Task Attributes

We present two tables for referencing the top three frequently used metrics for agent attributes (Tab. 7) and task attributes (Tab. 8).

Agent Attributes	Top 3 Agent-Oriented Metrics
Activity History	External alignment metrics, internal consistency metrics, content and textual metrics
Belief and Value	Psychological metrics, bias, fairness, and ethics metrics
Demographic Info.	Psychological metrics, internal consistency metrics, external alignment metrics
Psychological Traits	Psychological metrics, internal consistency metrics, content and textual metrics
Skill and Expertise	External alignment metrics, internal consistency metrics, content and textual metrics
Social Relationship	Psychological metrics, external alignment metrics, social and decision-making metrics

Table 7: Top 3 frequently used agent-oriented metrics for each agent attribute

Task Attributes	Top 3 Task-Oriented Metrics
Simulated Individuals	Psychological, performance, and internal consistency metrics
Simulated Society	Social and decision-making metrics, performance metrics, and psychological metrics
Opinion Dynamics	Performance metrics, external alignment metrics, and bias, fairness, and ethics metrics
Decision Making	Social and decision-making, performance, and psychological metrics
Psychological Experiment	Psychological, content and textual, and performance metrics
Educational Training	Psychological, performance, and content and textual metrics
Writing	Content and textual, psychological, and performance metrics

Table 8: Top 3 frequently used task-oriented metrics for each task attribute

E Case Study: Flawed Example

Fig. 9 visualized how the authors in the flawed example selected their evaluation metrics how further evaluation metrics could be uncovered through our proposed guideline.

F Questionnaire

G Metrics Glossary

We present two glossary tables for referencing the source of agent-oriented metrics (Tab. 9) and task-oriented metrics (Tab. 10). To clarify how these metrics are adapted and implemented in practice, we also provide concrete examples across different use cases for task-oriented metrics (Tab. 11).

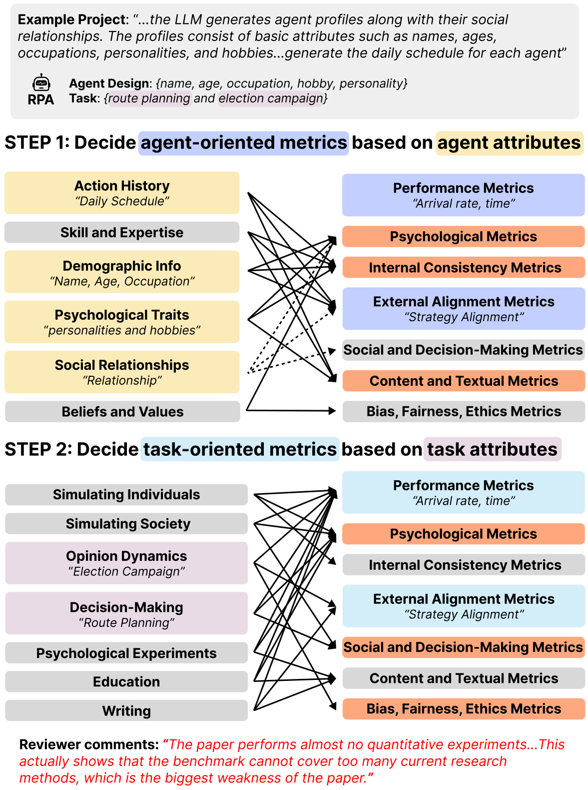

Figure 9: Case study of a flawed example in Section 5.2. Given agent attributes (yellow) and task attributes (pink). The original authors' selection of evaluation metrics (purple and blue). The missing metrics that are recommended by our proposed guideline (orange) align with the reviewer's criticism in red text.

Table 9: Agent-oriented evaluation metrics glossary.

Attribute	Category	Agent-oriented Metrics	Approach Source
Belief & Value	Bias, fairness, ethics metrics	Exaggeration (normalized average cosine similarity)	Automatic (Cheng et al., 2023)
Belief & Value	Bias, fairness, ethics metrics	Individuation (classification accuracy)	Automatic (Cheng et al., 2023)
Belief & Value	Bias, fairness, ethics metrics	Bias (performance disparity, prevalence, magnitude, variation, attitude shift)	Automatic (Gupta et al., 2024)
Belief & Value	Bias, fairness, ethics metrics	Bias (performance disparity, prevalence, magnitude, variation, attitude shift)	Automatic (Taubenfeld et al., 2024)
Demographic Information	Bias, fairness, ethics metrics	Exaggeration (normalized average cosine similarity)	Automatic (Cheng et al., 2023)
Demographic Information	Bias, fairness, ethics metrics	Individuation (classification accuracy)	Automatic (Cheng et al., 2023)
Demographic Information	Bias, fairness, ethics metrics	Bias (performance disparity, prevalence, magnitude, variation, attitude shift)	Automatic (Gupta et al., 2024)
Demographic Information	Bias, fairness, ethics metrics	Bias (performance disparity, prevalence, magnitude, variation, attitude shift)	Automatic (Neuberger et al., 2024)
Demographic Information	Bias, fairness, ethics metrics	Bias (performance disparity, prevalence, magnitude, variation, attitude shift)	Automatic (Taubenfeld et al., 2024)
Demographic Information	Bias, fairness, ethics metrics	Message toxicity	Automatic (Fang et al., 2024)
Activity History	Content and textual metrics	Coherence	LLM (Li et al., 2024e)
Activity History	Content and textual metrics	Clarity	Human (Chen et al., 2024b)
Activity History	Content and textual metrics	Diversity of dialog (Shannon entropy, intra-remote-clique, inter-remote-clique, semantic similarity, longest common subsequence similarity)	Automatic (Ha et al., 2024)
Belief & Value	Content and textual metrics	Diversity of dialog (Shannon entropy, intra-remote-clique, inter-remote-clique, semantic similarity, longest common subsequence similarity)	Automatic (Gu et al., 2024)
Demographic Information	Content and textual metrics	Coherence	LLM (Li et al., 2024e)
Demographic Information	Content and textual metrics	Attitudes (topic term frequency)	Automatic (Fang et al., 2024)
Demographic Information	Content and textual metrics	Diversity of dialog (Shannon entropy, intra-remote-clique, inter-remote-clique, semantic similarity, longest common subsequence similarity)	Automatic (Fang et al., 2024)
Demographic Information	Content and textual metrics	Clarity	Human (Chen et al., 2024b)
Demographic Information	Content and textual metrics	Diversity of dialog (Shannon entropy, intra-remote-clique, inter-remote-clique, semantic similarity, longest common subsequence similarity)	Automatic (Ha et al., 2024)
Demographic Information	Content and textual metrics	Linguistic complexity (utterance length, Kolmogorov complexity)	Automatic (Milička et al., 2024)
Psychological Traits	Content and textual metrics	Text similarity (BLEU, ROUGE)	Automatic (Zeng et al., 2024)
Psychological Traits	Content and textual metrics	Tone Alignment	LLM (Zeng et al., 2024)
Skills and Expertise	Content and textual metrics	Coherence	LLM (Li et al., 2024e)
Activity History	External alignment metrics	Hallucination	LLM (Shao et al., 2023)
Activity History	External alignment metrics	Entailment	LLM (Li et al., 2024e)
Activity History	External alignment metrics	Believability/Credibility(self-knowledge, memory, plans, reactions, reflections)	Human (Park et al., 2023)

Continued on next page

Attribute	Category	Agent-oriented Metrics	Approach	Source
Demographic Information	External alignment metrics	Entailment	LLM	(Li et al., 2024e)
Demographic Information	External alignment metrics	Believability/Credibility(self-knowledge, memory, plans, reactions, reflections)	Human	(Park et al., 2023)
Psychological Traits	External alignment metrics	Fact Accuracy	LLM	(Zeng et al., 2024)
Skills and Expertise	External alignment metrics	Hallucination	LLM	(Shao et al., 2023)
Skills and Expertise	External alignment metrics	Entailment	LLM	(Li et al., 2024e)
Activity History	Internal consistency metrics	Stability	LLM	(Shao et al., 2023)
Activity History	Internal consistency metrics	Consistency of information	Human	(Chen et al., 2024b)
Belief & Value	Internal consistency metrics	Attitude shift	LLM	(Wang et al., 2024e)
Demographic Information	Internal consistency metrics	Stability	LLM	(Shao et al., 2023)
Demographic Information	Internal consistency metrics	Attitude shift	LLM	(Neuberger et al., 2024)
Demographic Information	Internal consistency metrics	Attitude shift	LLM	(Taubenfeld et al., 2024)
Demographic Information	Internal consistency metrics	Behavior stability (mean, standard deviation)	Automatic	(Wang et al., 2024g)
Demographic Information	Internal consistency metrics	Consistency of information	Human	(Chen et al., 2024b)
Demographic Information	Internal consistency metrics	Consistency of psychological state / personalities	Human	(Chen et al., 2024b)
Demographic Information	Internal consistency metrics	Consistency of information	Human	(Zeng et al., 2024)
Psychological Traits	Internal consistency metrics	Stability	LLM	(Shao et al., 2023)
Psychological Traits	Internal consistency metrics	Consistency of information	Human	(Zeng et al., 2024)
Psychological Traits	Internal consistency metrics	Consistency of psychological state / personalities	Human	(Zeng et al., 2024)
Psychological Traits	Internal consistency metrics	Consistency of information	Human	(Cai et al., 2024)
Psychological Traits	Internal consistency metrics	Consistency of psychological state / personalities	Human	(Cai et al., 2024)
Skills and Expertise	Internal consistency metrics	Stability	LLM	(Shao et al., 2023)
Activity History	Performance metrics	Memorization	LLM	(Shao et al., 2023)
Demographic Information	Performance metrics	Memorization	LLM	(Chen et al., 2024b)
Demographic Information	Performance metrics	Communication ability (win rates)	Automatic	(Liu et al., 2024a)
Demographic Information	Performance metrics	Reaction (accuracy)	Automatic	(Liu et al., 2024a)
Demographic Information	Performance metrics	Self-knowledge (accuracy)	Automatic	(Liu et al., 2024a)
Activity History	Psychological metrics	Empathy	Human	(Chen et al., 2024b)
Belief & Value	Psychological metrics	Value	LLM	(Shao et al., 2023)
Demographic Information	Psychological metrics	Personality consistency	Automatic	(Wang et al., 2024c)
Demographic Information	Psychological metrics	Measured alignment for personality	Human	(Wang et al., 2024c)
Demographic Information	Psychological metrics	Sentiment	Automatic	(Fang et al., 2024)
Demographic Information	Psychological metrics	Empathy	Human	(Chen et al., 2024b)
Demographic Information	Psychological metrics	Belief (stability, evolution, correlation with behavior)	Automatic	(Lei et al., 2024)

Continued on next page

Attribute	Category	Agent-oriented Metrics	Approach	Source
Psychological Traits	Psychological metrics	Personality	Automatic	(Shao et al., 2023)
Psychological Traits	Psychological metrics	Belief (stability, evolution, correlation with behavior)	Automatic	(Shao et al., 2023)
Psychological Traits	Psychological metrics	Emotion responses (entropy of valence and arousal)	Automatic	(Shao et al., 2023)
Psychological Traits	Psychological metrics	Personality (Machine Personality Inventory, PsychoBench)	Automatic	(Jiang et al., 2023a)
Psychological Traits	Psychological metrics	Personality (vignette tests)	Human	(Jiang et al., 2023a)
Belief & Value	Social and decision-making metrics	Social value orientation (SVO-based Value Rationality Measurement)	Automatic	(Zhang et al., 2023b)

Table 10: Task-oriented evaluation metrics glossary.

Task	Category	Task-oriented Metrics	Approach Source
Decision Making	Social and economic metrics	Negotiation (Concession Rate, Negotiation Success Rate, Average Negotiation Round)	Automatic (Huang and Hadfi, 2024)
Decision Making	Social and economic metrics	Societal Satisfaction (average per-capita living area size, average waiting time, social welfare)	Automatic (Ji et al., 2024)
Decision Making	Social and economic metrics	Societal Fairness (variance in per capita living area size, number of inverse order pairs in house allocation, Gini coefficient)	Automatic (Ji et al., 2024)
Decision Making	Social and economic metrics	Macroeconomic (Inflation rate, Unemployment rate, Nominal GDP, Nominal GDP growth, Wage inflation, Real GDP growth, Expected monthly income, Consumption)	Automatic (Li et al., 2024d)
Decision Making	Social and economic metrics	Market and Consumer (Purchase probability, Expected competing product price, Customer counts, Price consistency between competitors)	Automatic (Gui and Toubia, 2023)
Decision Making	Social and economic metrics	Probability weighting	Automatic (Jia et al., 2024)
Decision Making	Social and economic metrics	Utility (Intrinsic Utility, Joint Utility)	Automatic (Huang and Hadfi, 2024)
Decision Making	Psychological metrics	Level of trust (distribution of amounts sent, trust rate)	Automatic (Xie et al., 2024a)
Decision Making	Psychological metrics	Risk preference	Automatic (Jia et al., 2024)
Decision Making	Psychological metrics	Loss aversion	Automatic (Jia et al., 2024)
Decision Making	Psychological metrics	Selfishness (Selfishness Index, Difference Index)	Automatic (Kim et al., 2024)
Decision Making	Performance metrics	Frequency (distribution of expert type)	Automatic (Wang et al., 2024b)
Decision Making	Performance metrics	Valid response rate	Automatic (Xie et al., 2024a)
Decision Making	Performance metrics	Web search quality (Mean reciprocal rank, Mean reciprocal rank)	Automatic (Ren et al., 2024a)
Decision Making	Performance metrics	Performance deviations/alignment from the baseline (accuracy, Jaccard Index, Cohen's Kappa Coefficient, Percentage Agreement, overlapping ratio between prediction and targets)	Automatic (Kim et al., 2024)
Decision Making	Performance metrics	Performance deviations/alignment from the baseline (accuracy, Jaccard Index, Cohen's Kappa Coefficient, Percentage Agreement, overlapping ratio between prediction and targets)	Automatic (Jin et al., 2024)
Decision Making	Performance metrics	Performance deviations/alignment from the baseline (accuracy, Jaccard Index, Cohen's Kappa Coefficient, Percentage Agreement, overlapping ratio between prediction and targets)	Automatic (Wang et al., 2024b)
Decision Making	Performance metrics	Performance deviations/alignment from the baseline (accuracy, Jaccard Index, Cohen's Kappa Coefficient, Percentage Agreement, overlapping ratio between prediction and targets)	Automatic (Wang et al., 2024f)
Decision Making	Internal consistency metrics	Behavioral alignment (lottery rate, behavior dynamic, Imitation and differentiation behavior, Proportion of similar and different dishes)	Automatic (Xie et al., 2024a)
Decision Making	Internal consistency metrics	Behavioral alignment (lottery rate, behavior dynamic, Imitation and differentiation behavior, Proportion of similar and different dishes)	Automatic (Zhao et al., 2023)

Continued on next page

Task	Category	Task-oriented Metrics	Approach	Source
Decision Making	Internal consistency metrics	Cultural appropriateness (Alignment between persona information and its assigned nationality)	LLM	(Li et al., 2024e)
Decision Making	External alignment metrics	Factual hallucinations (String matching overlap ratio)	Automatic	(Wang et al., 2024f)
Decision Making	External alignment metrics	Simulation capability (Turing test)	Human	(Ji et al., 2024)
Decision Making	External alignment metrics	Entailment	LLM	(Li et al., 2024e)
Decision Making	External alignment metrics	Realism	LLM	(Li et al., 2024e)
Educational Training	Psychological metrics	Perceived reflection on the development of essential non-cognitive skills	Human	(Yan et al., 2024)
Educational Training	Psychological metrics	Sense of immersion / Perceived immersion	Human	(Lee et al.)
Educational Training	Psychological metrics	Perceived intelligence	Human	(Cheng et al., 2024)
Educational Training	Psychological metrics	Perceived enjoyment	Human	(Cheng et al., 2024)
Educational Training	Psychological metrics	Perceived trust	Human	(Cheng et al., 2024)
Educational Training	Psychological metrics	Perceived sense of connection	Human	(Cheng et al., 2024)
Educational Training	Psychological metrics	Personality (Big Five Invertory, MBTI score, SD3 score, Linguistic Inquiry and Word Count framework, HEXACO)	Automatic	(Sonlu et al., 2024)
Educational Training	Psychological metrics	Personality (Big Five Invertory, MBTI score, SD3 score, Linguistic Inquiry and Word Count framework, HEXACO)	Automatic	(Liu et al., 2024d)
Educational Training	Psychological metrics	Perceived usefulness	Human	(Cheng et al., 2024)
Educational Training	Performance metrics	Density of knowledge-building	Automatic	(Jin et al., 2023)
Educational Training	Performance metrics	Effectiveness of questioning	Human	(Shi et al., 2023)
Educational Training	Performance metrics	Success criterion function outputs before operation and after operation	Human	(Li et al., 2023a)
Educational Training	External alignment metrics	Knowledge level (reconfigurability, persistence, and adaptability)	Automatic	(Jin et al., 2023)
Educational Training	External alignment metrics	Perceived human-likeness	Human	(Cheng et al., 2024)
Educational Training	Content and textual metrics	Story Content Generation (narratives staging score)	Automatic	(Yan et al., 2024)
Educational Training	Content and textual metrics	Willingness to speak	Human	(Shi et al., 2023)
Educational Training	Content and textual metrics	Authenticity	Human	(Lee et al.)
Opinion Dynamics	Psychological metrics	Opinion change	Human	(Triem and Ding, 2024)
Opinion Dynamics	Psychological metrics	Emotional density	Automatic	(Gao et al., 2023)
Opinion Dynamics	Performance metrics	Prediction accuracy (F1 score, AUC, MSE, MAE, depression risk prediction accuracy, suicide risk prediction accuracy)	Automatic	(Gao et al., 2023)
Opinion Dynamics	Performance metrics	Prediction accuracy (F1 score, AUC, MSE, MAE, depression risk prediction accuracy, suicide risk prediction accuracy)	Automatic	(Mou et al., 2024c)
Opinion Dynamics	Performance metrics	Prediction accuracy (F1 score, AUC, MSE, MAE, depression risk prediction accuracy, suicide risk prediction accuracy)	Automatic	(Yu et al., 2024)

Continued on next page

Task	Category	Task-oriented Metrics	Approach	Source
Opinion Dynamics	Performance metrics	Classification accuracy	Human	(Chan et al., 2023)
Opinion Dynamics	Performance metrics	Rephrase accuracy	Automatic	(Ju et al., 2024)
Opinion Dynamics	Performance metrics	Legal articles evaluation (precision, recall, F1)	Automatic	(He et al., 2024a)
Opinion Dynamics	Performance metrics	Judgment evaluation for civil and administrative cases (precision, recall, F1)	Automatic	(He et al., 2024a)
Opinion Dynamics	Performance metrics	Judgment evaluation for criminal cases (accuracy)	Automatic	(He et al., 2024a)
Opinion Dynamics	Performance metrics	Locality accuracy	Automatic	(Ju et al., 2024)
Opinion Dynamics	Performance metrics	Decision probability	Human	(Triem and Ding, 2024)
Opinion Dynamics	Performance metrics	Decision volatility	Human	(Triem and Ding, 2024)
Opinion Dynamics	Performance metrics	Alignment (compare simulation results with actual social outcomes)	Automatic	(Wang et al., 2024g)
Opinion Dynamics	Internal consistency metrics	Alignment (stance, content, behavior, static attitude distribution, time series of the average attitude)	Automatic	(Mou et al., 2024c)
Opinion Dynamics	Internal consistency metrics	Personality-behavior alignment	Human	(Navarro et al., 2024)
Opinion Dynamics	Internal consistency metrics	Similarity between initial and post preference (KL-divergence, RMSE)	Automatic	(Namikoshi et al., 2024)
Opinion Dynamics	Internal consistency metrics	Role playing	Human	(Lv et al., 2024)
Opinion Dynamics	External alignment metrics	Correctness	Human	(He et al., 2024a)
Opinion Dynamics	External alignment metrics	Logicality	Human	(He et al., 2024a)
Opinion Dynamics	External alignment metrics	Concision	Human	(He et al., 2024a)
Opinion Dynamics	External alignment metrics	Human likeness index	Automatic	(Chuang et al., 2023b)
Opinion Dynamics	External alignment metrics	Alignment between model and human (Kappa correlation coefficient, MAE), Authenticity (alignment of ratings between the agent and human annotators)	Human	(Chan et al., 2023)
Opinion Dynamics	External alignment metrics	Alignment between model and human (Kappa correlation coefficient, MAE), Authenticity (alignment of ratings between the agent and human annotators)	Human	(Lv et al., 2024)
Opinion Dynamics	Content and textual metrics	Turn-level Kendall-Tau correlation (naturalness, coherence, engagingness and groundedness)	Automatic	(Chan et al., 2023)
Opinion Dynamics	Content and textual metrics	Turn-level Spearman correlation (naturalness, coherence, engagingness and groundedness)	Automatic	(Chan et al., 2023)
Opinion Dynamics	Bias, fairness, and ethic metrics	Partisan bias	Automatic	(Chuang et al., 2023b)
Opinion Dynamics	Bias, fairness, and ethic metrics	Bias (cultural, linguistic, economic, demographic, ideological)	Automatic	(Qu and Wang, 2024)
Opinion Dynamics	Bias, fairness, and ethic metrics	Bias (mean)	Automatic	(Chuang et al., 2023a)
Opinion Dynamics	Bias, fairness, and ethic metrics	Extreme values	Automatic	(Chuang et al., 2023b)
Opinion Dynamics	Bias, fairness, and ethic metrics	Wisdom of Partisan Crowds effect	Automatic	(Chuang et al., 2023b)
Opinion Dynamics	Bias, fairness, and ethic metrics	Opinion diversity	Automatic	(Chuang et al., 2023a)
Psychological Experiment	Psychological metrics	Attitude change	Automatic	(Wang et al., 2023b)
Psychological Experiment	Psychological metrics	Average happiness value per time step	Automatic	(He and Zhang, 2024)

Continued on next page

Task	Category	Task-oriented Metrics	Approach	Source
Psychological Experiment	Psychological metrics	Belief value	Automatic	(Lei et al., 2024)
Psychological Experiment	Psychological metrics	Personality (Big Five Invertory, MBTI score, SD3 score, Linguistic Inquiry and Word Count framework, HEX-ACO)	Automatic	(He and Zhang, 2024)
Psychological Experiment	Psychological metrics	Personality (Big Five Invertory, MBTI score, SD3 score, Linguistic Inquiry and Word Count framework, HEX-ACO)	Automatic	(de Winter et al., 2024)
Psychological Experiment	Psychological metrics	Personality (Big Five Invertory, MBTI score, SD3 score, Linguistic Inquiry and Word Count framework, HEX-ACO)	Automatic	(Bose et al., 2024)
Psychological Experiment	Psychological metrics	Personality (Big Five Invertory, MBTI score, SD3 score, Linguistic Inquiry and Word Count framework, HEX-ACO)	Automatic	(Jiang et al., 2023b)
Psychological Experiment	Psychological metrics	Longitudinal trajectories of emotions	Automatic	(De Duro et al., 2025)
Psychological Experiment	Psychological metrics	Emotion	Automatic	(Lei et al., 2024)
Psychological Experiment	Performance metrics	Behavior reward	Automatic	(Lei et al., 2024)
Psychological Experiment	Internal consistency metrics	Behavioral similarity	Automatic	(Li et al., 2024b)
Psychological Experiment	Internal consistency metrics	Perception consistency (agent perceived safety, agent perceived liveliness)	LLM	(Verma et al., 2023)
Psychological Experiment	External alignment metrics	Rationality of the agent memory	Automatic	(Wang et al., 2023b)
Psychological Experiment	External alignment metrics	Believability of behavior	Automatic	(Wang et al., 2023b)
Psychological Experiment	Content and textual metrics	Salience of individual words	Automatic	(De Duro et al., 2025)
Psychological Experiment	Content and textual metrics	Absolutist words	Automatic	(De Duro et al., 2025)
Psychological Experiment	Content and textual metrics	Personal pronouns or emotions	Automatic	(De Duro et al., 2025)
Psychological Experiment	Content and textual metrics	Information entropy	Automatic	(Wang et al., 2023b)
Psychological Experiment	Content and textual metrics	Story (readability, personalness, redundancy, cohesiveness, likeability, believability)	Human	(Jiang et al., 2023b)
Psychological Experiment	Content and textual metrics	Story (readability, personalness, redundancy, cohesiveness, likeability, believability)	LLM	(Jiang et al., 2023b)
Simulated Individual	Social and economic metrics	Numbers of generated peer support strategies	Automatic	(Liu et al., 2024b)
Simulated Individual	Social and economic metrics	Perceived social support	Human	(Liu et al., 2024b)
Simulated Individual	Psychological metrics	Emotions	Human	(Pataranutaporn et al., 2024)
Simulated Individual	Psychological metrics	Agency	Human	(Pataranutaporn et al., 2024)
Simulated Individual	Psychological metrics	Future consideration	Human	(Pataranutaporn et al., 2024)
Simulated Individual	Psychological metrics	Self-reflection	Human	(Pataranutaporn et al., 2024)
Simulated Individual	Psychological metrics	Insight	Human	(Pataranutaporn et al., 2024)
Simulated Individual	Psychological metrics	Persona Perception Scale	Human	(Salminen et al., 2024)
Simulated Individual	Psychological metrics	Persona Perception Scale	Human	(Shin et al., 2024)

Continued on next page

Task	Category	Task-oriented Metrics	Approach	Source
Simulated Individual	Psychological metrics	Persona Perception Scale	Human	(Ha et al., 2024)
Simulated Individual	Psychological metrics	Persona Perception Scale	Human	(Chen et al., 2024b)
Simulated Individual	Psychological metrics	Sensitivity to personalization	Automatic	(Giorgi et al., 2024)
Simulated Individual	Psychological metrics	Agent self-awareness	LLM	(Xie et al., 2024b)
Simulated Individual	Psychological metrics	Personality (Big Five Invertory rated by LLM)	LLM	(Jiang et al., 2023a)
Simulated Individual	Psychological metrics	Positively mention rate	Automatic	(Kamruzzaman and Kim, 2024)
Simulated Individual	Psychological metrics	Optimism	Human	(Pataranutaporn et al., 2024)
Simulated Individual	Psychological metrics	Self-esteem	Human	(Pataranutaporn et al., 2024)
Simulated Individual	Psychological metrics	Pressure perceived scale	Human	(Liu et al., 2024b)
Simulated Individual	Performance metrics	Error rates (error of average, error of dispersion)	Automatic	(Lin et al., 2024)
Simulated Individual	Performance metrics	Model fit indices (Chi-square to degrees of freedom ratio, Comparative Fit Index, Tucker-Lewis Index, Root Mean Square Error of Approximation)	Automatic	(Ke and Ng, 2024)
Simulated Individual	Performance metrics	Knowledge accuracy (WikiRoleEval with human evaluators)	Human	(Tang et al., 2024)
Simulated Individual	Performance metrics	Knowledge accuracy (WikiRoleEval)	LLM	(Tang et al., 2024)
Simulated Individual	Performance metrics	Win rates	Automatic	(Chi et al., 2024)
Simulated Individual	Performance metrics	Comprehension	Automatic	(Shin et al., 2024)
Simulated Individual	Performance metrics	Completeness	Automatic	(Shin et al., 2024)
Simulated Individual	Performance metrics	Validity (average variance extracted, inter-construct correlations)	Automatic	(Ke and Ng, 2024)
Simulated Individual	Performance metrics	Composite reliability	Automatic	(Ke and Ng, 2024)
Simulated Individual	Performance metrics	Rated statement quality	Human	(Liu et al., 2023)
Simulated Individual	Performance metrics	Rated statement quality	LLM	(Liu et al., 2023)
Simulated Individual	Performance metrics	Conversational ability (CharacterEval)	LLM	(Tang et al., 2024)
Simulated Individual	Performance metrics	Roleplay subset of MT-Bench	LLM	(Tang et al., 2024)
Simulated Individual	Performance metrics	Professional scale (accuracy in replicating profession-specific knowledge)	LLM	(Sun et al., 2024)
Simulated Individual	Performance metrics	Language quality	LLM	(Zhang et al., 2024a)
Simulated Individual	Performance metrics	Prediction accuracy between real data and generated data (Replication success rate, Kullback-Leibler divergence)	Automatic	(Assaf and Lynar, 2024)
Simulated Individual	Performance metrics	Prediction accuracy between real data and generated data (Replication success rate, Kullback-Leibler divergence)	Automatic	(Tamaki and Littvay, 2024)
Simulated Individual	Performance metrics	Prediction accuracy between real data and generated data (Replication success rate, Kullback-Leibler divergence)	Automatic	(Park et al., 2024)
Simulated Individual	Performance metrics	Prediction accuracy between real data and generated data (Replication success rate, Kullback-Leibler divergence)	Automatic	(Yeykelis et al., 2024)

Continued on next page

Task	Category	Task-oriented Metrics	Approach	Source
Simulated Individual	Performance metrics	Accuracy of distinguishing between AI-generated and human-built solutions	Automatic	(Schuller et al., 2024)
Simulated Individual	Internal consistency metrics	Accuracy of reaction based on social relationship	Automatic	(Liu et al., 2024a)
Simulated Individual	Internal consistency metrics	Perceived connection between personas and system outcomes	Human	(Chen et al., 2024b)
Simulated Individual	Internal consistency metrics	Representativeness (Wasserstein distance, respond with similar answers to individual survey questions), Consistency (Frobenius norm, the correlation across responses to a set of questions in each survey)	Automatic	(Moon et al., 2024)
Simulated Individual	Internal consistency metrics	Role consistency (WikiRoleEval with human evaluators)	Human	(Tang et al., 2024)
Simulated Individual	Internal consistency metrics	Role consistency/attractiveness (WikiRoleEval, CharacterEval)	LLM	(Tang et al., 2024)
Simulated Individual	Internal consistency metrics	Consistency	Human	(Zhang et al., 2024a)
Simulated Individual	Internal consistency metrics	Consistency	Human	(Mishra et al., 2023)
Simulated Individual	Internal consistency metrics	Future self-continuity	Human	(Pataranutaporn et al., 2024)
Simulated Individual	Internal consistency metrics	Agreement between a synthetic annotator both with and without a leave-one-out attribute (Cohen's Kappa)	Automatic	(Castricato et al., 2024)
Simulated Individual	Internal consistency metrics	Consistency with the scenario and characters	Automatic	(Zhang et al., 2024a)
Simulated Individual	Internal consistency metrics	Quality and logical coherence of the script content	Automatic	(Zhang et al., 2024a)
Simulated Individual	Internal consistency metrics	Nation-related response percentage	Automatic	(Kamruzzaman and Kim, 2024)
Simulated Individual	External alignment metrics	Unknown question rejection (WikiRoleEval with human evaluators)	Human	(Tang et al., 2024)
Simulated Individual	External alignment metrics	Unknown question rejection (WikiRoleEval)	LLM	(Tang et al., 2024)
Simulated Individual	External alignment metrics	Accuracy of self-knowledge	Automatic	(Liu et al., 2024a)
Simulated Individual	External alignment metrics	Correctness	Human	(Zhang et al., 2024a)
Simulated Individual	External alignment metrics	Correctness	Human	(Milička et al., 2024)
Simulated Individual	External alignment metrics	Agreement score between human raters and LLM,	Automatic	(Liu et al., 2023)
Simulated Individual	External alignment metrics	Agreement score between human raters and LLM,	Automatic	(Jiang et al., 2023a)
Simulated Individual	External alignment metrics	Agreement score between human raters and LLM,	Automatic	(Liu et al., 2024a)
Simulated Individual	External alignment metrics	Human-likeness	Human	(Zhang et al., 2024a)
Simulated Individual	Content and textual metrics	Content similarity (ROUGE-L, BERTScore, GPT-based-similarity, G-eval)	Automatic	(Shin et al., 2024)
Simulated Individual	Content and textual metrics	Entity density of summarization	Automatic	(Liu et al., 2024a)
Simulated Individual	Content and textual metrics	Entity recall of summarization	Automatic	(Liu et al., 2024a)
Simulated Individual	Content and textual metrics	Dialog diversity	Automatic	(Lin et al., 2024)
Simulated Individual	Bias, fairness, and ethic metrics	Hate speech detection accuracy	Automatic	(Giorgi et al., 2024)
Simulated Individual	Bias, fairness, and ethic metrics	Population heterogeneity	Automatic	(Murthy et al., 2024)
Simulated Society	Social and economic metrics	Social Conflict Count	Automatic	(Ren et al., 2024b)

Continued on next page

Task	Category	Task-oriented Metrics	Approach	Source
Simulated Society	Social and economic metrics	Social Rules	Human	(Zhou et al., 2024b)
Simulated Society	Social and economic metrics	Social Rules	LLM	(Zhou et al., 2024b)
Simulated Society	Social and economic metrics	Financial and Material Benefits	Human	(Zhou et al., 2024b)
Simulated Society	Social and economic metrics	Financial and Material Benefits	LLM	(Zhou et al., 2024b)
Simulated Society	Social and economic metrics	Converged price	Automatic	(Toledo-Zucco et al., 2024)
Simulated Society	Social and economic metrics	Information diffusion	Automatic	(Park et al., 2023)
Simulated Society	Social and economic metrics	Relationship formation	Automatic	(Park et al., 2023)
Simulated Society	Social and economic metrics	Relationship	LLM	(Zhou et al., 2024b)
Simulated Society	Social and economic metrics	Coordination within other agents	Automatic	(Park et al., 2023)
Simulated Society	Social and economic metrics	Probability of social connection formation	Automatic	(Leng and Yuan, 2024)
Simulated Society	Social and economic metrics	Percent of social welfare maximization choices	Automatic	(Leng and Yuan, 2024)
Simulated Society	Social and economic metrics	Persuasion (distribution of persuasion outcomes, odds ratios)	Automatic	(Campedelli et al., 2024)
Simulated Society	Social and economic metrics	Anti-social behavior (effect on toxic messages)	Automatic	(Campedelli et al., 2024)
Simulated Society	Social and economic metrics	Norm Internalization Rate	Automatic	(Ren et al., 2024b)
Simulated Society	Social and economic metrics	Norm Compliance Rate	Automatic	(Ren et al., 2024b)
Simulated Society	Psychological metrics	NASA-TLX Scores	Human	(Zhang et al., 2024c)
Simulated Society	Psychological metrics	Helpfulness rating	Human	(Zhang et al., 2024c)
Simulated Society	Psychological metrics	Personality (Big Five Invertory, MBTI score, SD3 score, Linguistic Inquiry and Word Count framework, HEXACO)	Automatic	(Frisch and Giulianelli, 2024)
Simulated Society	Psychological metrics	Personality (Big Five Invertory, MBTI score, SD3 score, Linguistic Inquiry and Word Count framework, HEXACO)	Automatic	(Li et al., 2024b)
Simulated Society	Psychological metrics	Degree of reciprocity	Automatic	(Leng and Yuan, 2024)
Simulated Society	Psychological metrics	Pleasure rating	Human	(Zhang et al., 2024c)
Simulated Society	Psychological metrics	Trend of Favorability Decline	Automatic	(Gu et al., 2024)
Simulated Society	Psychological metrics	Negative Favorability Achievement	Automatic	(Gu et al., 2024)
Simulated Society	Psychological metrics	Trend of Favorability Decline	Automatic	(Gu et al., 2024)
Simulated Society	Psychological metrics	Negative Favorability Achievement	Automatic	(Gu et al., 2024)
Simulated Society	Performance metrics	Abstention accuracy	Automatic	(Ashkinaze et al., 2024)
Simulated Society	Performance metrics	Accuracy of information gathering	Automatic	(Kaiya et al., 2023)
Simulated Society	Performance metrics	Implicit reasoning accuracy	Automatic	(Mou et al., 2024b)
Simulated Society	Performance metrics	Prediction accuracy (F1 score, AUC, MSE, MAE, depression risk prediction accuracy, suicide risk prediction accuracy)	Automatic	(Lan et al., 2024)
Simulated Society	Performance metrics	Guess accuracy	Automatic	(Leng and Yuan, 2024)

Continued on next page

Task	Category	Task-oriented Metrics	Approach	Source
Simulated Society	Performance metrics	Classification accuracy	Automatic	(Li et al., 2024a)
Simulated Society	Performance metrics	Success rate	Automatic	(Kaiya et al., 2023)
Simulated Society	Performance metrics	Success rate	Automatic	(Li et al., 2023b)
Simulated Society	Performance metrics	Success rate	Automatic	(Li et al., 2023b)
Simulated Society	Performance metrics	Success rate for coordination (identification accuracy, workflow correctness, alignment between job and agent's skill)	Automatic	(Li et al., 2023a)
Simulated Society	Performance metrics	Success rate for coordination (identification accuracy, workflow correctness, alignment between job and agent's skill)	Automatic	(Li et al., 2023a)
Simulated Society	Performance metrics	Task Accuracy	Automatic	(Zhang et al., 2023a)
Simulated Society	Performance metrics	Task Accuracy	Automatic	(Lan et al., 2024)
Simulated Society	Performance metrics	Errors in the prompting sequence	Human	(Antunes et al., 2023)
Simulated Society	Performance metrics	Error-free execution	Automatic	(Wang et al., 2024a)
Simulated Society	Performance metrics	Goal completion	Human	(Mou et al., 2024b)
Simulated Society	Performance metrics	Goal completion	LLM	(Zhou et al., 2024a)
Simulated Society	Performance metrics	Goal completion	LLM	(Mou et al., 2024b)
Simulated Society	Performance metrics	Goal completion	LLM	(Zhou et al., 2024b)
Simulated Society	Performance metrics	Efficacy	Human	(Ashkinaze et al., 2024)
Simulated Society	Performance metrics	Knowledge	Human	(Zhou et al., 2024b)
Simulated Society	Performance metrics	Knowledge	LLM	(Zhou et al., 2024b)
Simulated Society	Performance metrics	Reasoning abilities	Automatic	(Chen et al., 2023)
Simulated Society	Performance metrics	Reasoning abilities	Human	(Chen et al., 2023)
Simulated Society	Performance metrics	Efficiency	Automatic	(Piatti et al., 2024)
Simulated Society	Performance metrics	Text understanding and creative writing abilities (Dialogue response dataset, Commongen Challenge)	LLM	(Chen et al., 2023)
Simulated Society	Performance metrics	Probabilities of receiving, storing, and retrieving the key information across the population	Automatic	(Kaiya et al., 2023)
Simulated Society	Performance metrics	Correlation between predicted and real results	Automatic	(Mitsopoulos et al., 2024)
Simulated Society	Internal consistency metrics	Behavioral similarity	Automatic	(Li et al., 2024b)
Simulated Society	Internal consistency metrics	Semantic consistency (cosine similarity)	Automatic	(Qiu and Lan, 2024)
Simulated Society	External alignment metrics	Alignment (Environmental understanding and response accuracy, adherence to predefined settings)	Automatic	(Gu et al., 2024)
Simulated Society	External alignment metrics	Strategy accuracy (strategies provided by the models vs. by human experts and evaluate the accuracy)	Automatic	(Zhang et al., 2024b)
Simulated Society	External alignment metrics	Believability of behavior	Human	(Zhou et al., 2024b)
Simulated Society	External alignment metrics	Believability of behavior	Human	(Park et al., 2023)

Continued on next page

Task	Category	Task-oriented Metrics	Approach	Source
Simulated Society	Content and textual metrics	Content similarity (ROUGE-L, BERTScore, GPT-based-similarity, G-eval, BLEU-4)	Automatic	(Li et al., 2024a)
Simulated Society	Content and textual metrics	Content similarity (ROUGE-L, BERTScore, GPT-based-similarity, G-eval)	Automatic	(Chen et al., 2024f)
Simulated Society	Content and textual metrics	Content similarity (ROUGE-L, BERTScore, GPT-based-similarity, G-eval)	Automatic	(Mishra et al., 2023)
Simulated Society	Content and textual metrics	Semantic understanding	Automatic	(Gu et al., 2024)
Simulated Society	Content and textual metrics	Complexity of generated content	Automatic	(Antunes et al., 2023)
Simulated Society	Content and textual metrics	Dialogue generation quality	Automatic	(Antunes et al., 2023)
Simulated Society	Content and textual metrics	Number of conversation rounds	Automatic	(Zhang et al., 2024c)
Simulated Society	Bias, fairness, and ethic metrics	Bias rate (herd effect, authority effect, ban franklin effect, rumor chain effect, gambler's fallacy, confirmation bias, halo effect)	Human	(Liu et al., 2025)
Simulated Society	Bias, fairness, and ethic metrics	Bias rate (herd effect, authority effect, ban franklin effect, rumor chain effect, gambler's fallacy, confirmation bias, halo effect)	LLM	(Liu et al., 2025)
Simulated Society	Bias, fairness, and ethic metrics	Bias rate (herd effect, authority effect, ban franklin effect, rumor chain effect, gambler's fallacy, confirmation bias, halo effect)	Automatic	(Liu et al., 2025)
Simulated Society	Bias, fairness, and ethic metrics	Equality	Automatic	(Piatti et al., 2024)
Writing	Psychological metrics	Qualitative feedback (expertise, social relation, valence, level of involvement)	Human	(Benharrak et al., 2024)
Writing	Performance metrics	Prediction accuracy (F1 score, AUC, MSE, MAE, depression risk prediction accuracy, suicide risk prediction accuracy)	Automatic	(Wang et al., 2024f)
Writing	Performance metrics	Success rate	Automatic	(Wang et al., 2024d)
Writing	Performance metrics	Behavioral patterns	Human	(Zhang et al., 2024c)
Writing	Internal consistency metrics	Consistency (user profile, psychotherapeutic approach)	Automatic	(Mishra et al., 2023)
Writing	Internal consistency metrics	Motivational consistency	LLM	(Wang et al., 2024d)
Writing	Internal consistency metrics	Audience similarity	Human	(Choi et al., 2024)
Writing	Internal consistency metrics	Quality of generated dimension & values (relevance, mutual exclusiveness)	Human	(Choi et al., 2024)
Writing	External alignment metrics	Factual error rate	Automatic	(Wang et al., 2024f)
Writing	External alignment metrics	Correctness (politeness, interpersonal behaviour)	Automatic	(Mishra et al., 2023)
Writing	External alignment metrics	Hallucination (groundedness of the chat responses)	Human	(Choi et al., 2024)
Writing	Content and textual metrics	Linguistic similarity	Human	(Choi et al., 2024)
Writing	Content and textual metrics	Fluency	Human	(Mishra et al., 2023)
Writing	Content and textual metrics	Perplexity	Automatic	(Mishra et al., 2023)
Writing	Content and textual metrics	Non-Repetitiveness	Human	(Mishra et al., 2023)
Writing	Content and textual metrics	response generation quality	Automatic	(Li et al., 2024a)
Writing	Content and textual metrics	Coherency	LLM	(Wang et al., 2024d)

Table 11: Use Cases of Task-Oriented Metric Implementation.

Metrics	Task	Implementation	Source
Accuracy	Opinion Dynamics	Accuracy is measured by evaluating how well the simulation replicates individual-level behaviors, attitudes, and emotions and population-level dynamics	(Gao et al., 2023)
Accuracy	Opinion Dynamics	Accuracy is measured by comparing predicted voting outcomes against actual election results—using voting probabilities, state-level winner predictions, and vote share percentages—to assess both individual- and aggregate-level performance in reflecting real-world election trends.	(Yu et al., 2024)
Accuracy	Opinion Dynamics	Accuracy is measured as the proportion of correctly classified instances out of the total number of instances	(Chan et al., 2023)
Agency	Simulated Individual	Agency is measured through self-reported scores and analyzed using a Welch one-way ANOVA	(Pataranutaporn et al., 2024)
Agent self-awareness	Simulated Individual	Agent self-awareness is measured through manually crafted self-report questionnaires containing fill-in-the-blank and multiple-choice questions about the agent's identity, relationships, and life experiences, with scores based on exact match accuracy to assess memory and introspective consistency.	(Xie et al., 2024b)
Agreement between LLMs and humans	Opinion Dynamics	Agreement between LLMs and humans is measured using KL-divergence, which captures alignment with population-level response distributions, and root mean square error (RMSE), which reflects similarity to individual survey responses, both computed on test data matched by demographic distribution.	(Namikoshi et al., 2024)
Alignment	Opinion Dynamics	Alignment is measured by comparing simulation results with actual social outcomes, assessing how closely the agent-based behaviors and emergent patterns replicate real-world events, decisions, or trends.	(Wang et al., 2024g)
Alignment from the baseline	Decision Making	Alignment from the baseline is measured by comparing final decisions across settings using the Jaccard Index, Cohen's Kappa Coefficient, and Percentage Agreement, which assess overlap, inter-rater reliability, and direct agreement with the baseline, respectively.	(Jin et al., 2024)
Anti-social behavior	Simulated Society	Anti-social behavior is measured by the percentage of toxic messages in each conversation	(Campedelli et al., 2024)
Attitude change	Psychological Experiment	Attitude change is measured by the average frequency of score changes across rounds for agents with friends, using an indicator function to detect shifts in user scores between consecutive rounds, thereby capturing conformity-related dynamics in social interactions.	(Wang et al., 2023b)
Authenticity	Opinion Dynamics	Authenticity is measured by computing Cohen's K between the agent's ratings and human annotators' ratings for the same questionnaire items, quantifying the consistency and alignment of responses at each iteration.	(Lv et al., 2024)
Authenticity	Opinion Dynamics	Authenticity is measured using the Kappa correlation coefficient (Kap.), which quantifies the alignment between agent and human annotator ratings while adjusting for chance agreement, providing a robust assessment of response consistency.	(Chan et al., 2023)
Average happiness per time step	Psychological Experiment	Average happiness per time step is used to measure agent ability to maintain a positive emotional baseline throughout the process of preference shaping.	(He and Zhang, 2024)
Behavior Alignment	Opinion Dynamics	Behavior alignment is measured by evaluating whether agents replicate user actions—specifically posting and retweeting—with performance assessed using accuracy and macro F1 score based on observed behavior in Twitter datasets.	(Mou et al., 2024c)

Continued on next page

Metrics	Task	Implementation	Source
Behavioral alignment	Decision Making	Behavioral alignment is measured by comparing lottery rates (%)—the proportion of times LLM agents and humans choose to gamble or trust—in decision-making games, assessing how closely LLM behaviors align with human choices.	(Xie et al., 2024a)
Behavioral alignment	Decision Making	Behavioral alignment is measured by examining whether agent behaviors conform to classic sociological and economic theories—such as differentiation and imitation—and by evaluating decision outcomes (e.g., dish quality scores) based on empirically derived functions that integrate factors like cost, price, and chef salary.	(Zhao et al., 2023)
Behavioral reward	Psychological Experiment	Behavioral reward is measured by summing the final policy rewards of all individuals in a group after the last trial, with rejection and missing rates analyzed across conditions and demographics, where higher rejection rates correlate with lower (more ethical) reward scores.	(Lei et al., 2024)
Behavioral similarity	Psychological Experiment	Behavioral similarity is measured by calculating the Euclidean distance between daily goal distributions, with the overall activity level representing the average behavioral variation across all day pairs, capturing consistency or divergence in agent planning over time.	(Li et al., 2024b)
Belief value	Psychological Experiment	Belief value is measured as the strength or confidence of an agent in its decision, with changes over time reflecting the stability or adaptability of beliefs—where higher values indicate stronger conviction and lower values suggest reduced decisional steadfastness.	(Lei et al., 2024)
Believability of behavior	Psychological Experiment	Believability of simulated user behaviors is evaluated by assessing the realism of user actions in both recommender system interactions and chatting/broadcasting scenarios, typically through comparison with human behavior patterns or human judgment of authenticity.	(Wang et al., 2023b)
Bias	Opinion Dynamics	Bias is measured by comparing model responses across diverse cultural, linguistic, economic, demographic, and ideological contexts, using simulated scenarios and survey-based benchmarks to identify disparities and deviations from human data.	(Qu and Wang, 2024)
Bias	Opinion Dynamics	Bias is measured as the average of agents' opinions at the final time step, indicating the overall directional leaning of the group's final stance.	(Chuang et al., 2023a)
Coherency	Writing	Overall coherency evaluation is performed by prompting an LLM to assess the coherence of the generated plot and provide suggestions for improvement, offering a qualitative measure of narrative consistency.	(Wang et al., 2024d)
Content Alignment	Opinion Dynamics	Content alignment is measured by classifying agent-generated content into five predefined categories, with evaluation based on accuracy, macro F1 score, and cosine similarity between simulated and real-world content to assess both categorical and semantic alignment.	(Mou et al., 2024c)
Conversation	Psychological Experiment	Conversations are analyzed at two levels: the text level, examining features like absolutist words, personal pronouns, and emotions, and the network level, focusing on the salience and connectivity of individual words within the conversational structure.	(De Duro et al., 2025)
Conversational ability	Simulated Individual	Conversational ability is assessed within the CharacterEval framework using a set of metrics that evaluate an RPLA's capacity to sustain engaging, coherent, and immersive dialogue, as part of a broader focus on realistic role-based interactions.	(Tang et al., 2024)

Continued on next page

Metrics	Task	Implementation	Source
Cultural appropriateness	Decision Making	Cultural appropriateness is measured by assessing the alignment between persona information and its assigned nationality, using GPT-4O to evaluate whether generated simulations reflect culturally coherent behaviors and norms across diverse regional scenarios.	(Li et al., 2024e)
Decision probability	Opinion Dynamics	Decision probabilities are measured by encoding each step's label (e.g., 'AFFIRM' = 1, 'not AFFIRM' = 0, hallucinations/'NONE' = 3), tallying affirmative and non-affirmative outcomes per case, and analyzing these distributions to assess the likelihood of specific decisions across agents and case complexity.	(Triem and Ding, 2024)
Decision volatility	Opinion Dynamics	Decision volatility is measured by logging a binary value for each debate round transition—"1" if the agent changed its opinion between rounds, and "0" if it remained consistent—to track when and how often opinion shifts occur during a debate and to identify patterns in decision progression.	(Triem and Ding, 2024)
Density of knowledge-building	Educational Training	Density of knowledge-building is measured by analyzing the frequency of knowledge-building messages in learning dialogues	(Jin et al., 2023)
Dialogue response	Opinion Dynamics	Dialogue response is evaluated using turn-level Spearman and Kendall-Tau correlations between model outputs and human judgments on four key aspects: naturalness, coherence, engagingness, and groundedness, aligning with established benchmarking methods.	(Chan et al., 2023)
Effectiveness of questioning	Educational Training	The effectiveness of questioning is measured by analyzing how teachers pose and direct questions—both broadly and selectively—based on predefined teaching plans and real-time assessments of student status and classroom dynamics, reflecting strategic instructional engagement.	(Shi et al., 2023)
Emotion	Psychological Experiment	Emotion is measured by computing the entropy of normalized valence and arousal distributions for each individual, using histogram-based probability distributions to quantify the variability and complexity of emotional expression through entropy.	(Lei et al., 2024)
Emotional density	Opinion Dynamics	Emotional density is measured by analyzing the intensity and fluctuation of emotions expressed in agent interactions over time, capturing the dynamic process of emotion propagation and identifying key peaks in emotional response that align with real-world events.	(Gao et al., 2023)
Emotions	Simulated Individual	Emotions are measured by analyzing changes in self-reported emotional states—such as negative emotion, anxiety, feeling unmotivated, overwhelmed, and positive emotion—across intervention conditions using ANOVA tests, with significant differences indicating the emotional impact of specific interventions.	(Pataranutaporn et al., 2024)
Entailment	Decision Making	Entailment is measured by evaluating whether the content of the simulation logically aligns with the given assumptions, assessing the consistency and coherence between the generated output and its intended premise.	(Li et al., 2024e)
Error rate	Simulated Individual	Error rate is measured by calculating the normalized absolute difference between scalar dialogue features from LLM and CANDOR data, including metrics for average error (mean bias) and dispersion error (variability differences) to assess overall performance deviation.	(Lin et al., 2024)

Continued on next page

Metrics	Task	Implementation	Source
Extreme values	Opinion Dynamics	Extreme Values is measured as the proportion of LLM responses that exceed predefined realism thresholds, indicating the model's tendency to produce unrealistic outputs, and is excluded from other evaluation metrics to ensure fair comparison with human data.	(Chuang et al., 2023b)
Factual hallucinations	Decision Making	Factual hallucinations are measured by performing string matching between generated answers and ground-truth aliases from the TriviaQA dataset, with the metric computed as the proportion of correct answer mentions over the total number of trivia questions.	(Wang et al., 2024f)
Frequency	Decision Making	Frequency is measured by the occurrence count of each expert type across the dataset, where higher-frequency experts are deemed more reliable and generalizable for stance detection tasks, and low-frequency experts—often from unrelated domains—are filtered based on a threshold of total appearances.	(Wang et al., 2024b)
Future Consideration	Simulated Individual	Future Consideration is measured through self-reported scores reflecting individuals' attention to and planning for future outcomes, with differences across intervention conditions analyzed using a Welch one-way ANOVA due to unequal variances.	(Pataranutaporn et al., 2024)
Future self-continuity	Simulated Individual	Future Self-Continuity is measured through changes in similarity, vividness, and positivity toward one's future self, using self-report scales analyzed via Welch one-way ANOVA, with significant differences across intervention conditions indicating how strongly participants perceive connection and continuity with their future selves.	(Pataranutaporn et al., 2024)
Hallucination	Writing	Hallucination is measured by evaluating the groundedness of chat responses, identifying instances where personas inaccurately reference specific video or channel content, indicating a lack of factual alignment.	(Choi et al., 2024)
Human likeness index	Opinion Dynamics	The Human Likeness Index (HLI) measures the extent to which LLM agents resemble human behavior by combining two components—partisan bias and the wisdom of the crowd deviation.	(Chuang et al., 2023b)
Information entropy	Psychological Experiment	Information entropy is used to measure the severity of the information cocoon phenomenon by quantifying the diversity of item categories recommended to each user, where lower entropy values indicate more narrow and homogeneous exposure, reflecting stronger cocooning effects.	(Wang et al., 2023b)
Insight	Simulated Individual	Insight is measured through a composite self-report score, with changes across intervention conditions analyzed using a one-way ANOVA, confirming equal variances and assessing how interventions influence participants' depth of understanding or awareness.	(Pataranutaporn et al., 2024)
Judgment evaluation for civil and administrative cases	Opinion Dynamics	Judgment Evaluation for civil and administrative cases is conducted using GPT-4 to compare key points—including rulings, monetary amounts, and interest rates—between the agent's and reference judgments, with precision, recall, and F1 scores computed through micro-averaged counts of matching and non-matching key points.	(He et al., 2024a)
Judgment evaluation for criminal cases	Opinion Dynamics	Judgment Evaluation for criminal cases is measured by calculating the accuracy of the agent system in predicting three core elements—charge, prison term, and fine—each evaluated separately to ensure alignment with case facts and contextual factors like courtroom behavior and defense statements.	(He et al., 2024a)

Continued on next page

Metrics	Task	Implementation	Source
Knowledge accuracy	Simulated Individual	Knowledge accuracy is measured using the WikiRoleEval benchmark by evaluating whether the role-playing language agent (RPLA) provides factually correct responses aligned with its assigned role-specific knowledge.	(Tang et al., 2024)
Knowledge level	Educational Training	Knowledge level is measured by evaluating the agent's performance on multiple-choice questions (MCQs)	(Jin et al., 2023)
Language quality	Simulated Individual	Language quality is measured using a 1–5 quality score that evaluates fluency, emotional expression, logical consistency, and grammatical correctness in dialogue, with scores assigned by ChatGPT based on transcribed text from a pre-trained ASR model.	(Zhang et al., 2024a)
Legal articles evaluation	Opinion Dynamics	Legal Articles Evaluation is measured using strict matching between the agent-generated and reference legal article lists, with precision, recall, and F1 scores computed via micro-averaging to assess the accuracy and completeness of legal reference identification.	(He et al., 2024a)
Level of trust	Decision Making	Level of trust is measured by the distribution of amounts sent in the Trust Game and the trust rate	(Xie et al., 2024a)
Locality accuracy	Opinion Dynamics	Locality Accuracy measures the agent's ability to answer unrelated questions correctly after knowledge manipulation, ensuring that changes to one fact (e.g., editing Messi's profession) do not improperly affect unrelated facts (e.g., Ronaldo's profession); it is computed as the proportion of agent responses that match the ground truth for locality prompts.	(Ju et al., 2024)
Logical reasoning and ethical considerations	Opinion Dynamics	Logical reasoning and ethical considerations are evaluated by a panel of human annotators using binary True/False criteria for each analysis, assessing correctness (fair and inclusive reasoning), logicality (absence of illogical or false claims), and concision (completeness without unnecessary detail).	(He et al., 2024a)
Longitudinal trajectories of emotions	Psychological Experiment	Longitudinal trajectories of emotions are measured by performing emotional profiling at each conversational turn, distributing human conversation quips into 10 aligned steps to enable comparison with LLM-generated responses and analyze emotional dynamics over time.	(De Duro et al., 2025)
Loss aversion	Decision Making	Loss aversion is measured by comparing LLMs' responses to equivalent gain and loss scenarios, assessing whether losses are weighted more heavily than gains, consistent with behavioral economic theory.	(Jia et al., 2024)
Macroeconomic	Decision Making	Macroeconomic performance is measured through inflation rate, unemployment rate, nominal GDP, nominal GDP growth, wage inflation, real GDP growth, expected monthly income, and consumption, capturing both economic stability and agents' income-consumption dynamics.	(Li et al., 2024d)
Market and Consumer	Decision Making	Market and Consumer dynamics are measured by purchase probability, expected competing product price, customer counts, and price consistency between competitors, reflecting consumer behavior and competitive market stability.	(Gui and Toubia, 2023)
Model fit	Simulated Individual	Model fit indices are evaluated using CFI, TLI, and RMSEA, with acceptable thresholds confirming overall model fit	(Ke and Ng, 2024)
Negotiation	Decision Making	Negotiation is measured by Concession Rate to assess offer flexibility over time, Negotiation Success Rate to evaluate outcome effectiveness, and Average Negotiation Round to capture the efficiency of reaching agreements.	(Huang and Hadfi, 2024)

Continued on next page

Metrics	Task	Implementation	Source
Number of generated peer support strategies	Simulated Individual	Number of generated peer support strategies is measured by counting the supportive message types generated by the conversational agent (CA) and conducting a thematic analysis of user interactions, categorizing each round into six agreed-upon topics to assess strategy diversity and relevance.	(Liu et al., 2024b)
Opinion change	Opinion Dynamics	Opinion change is measured by human-labeling each discussion step based on stance (e.g., affirm, reverse, remand), allowing for a quantitative analysis of shifts in position throughout the debate process.	(Triem and Ding, 2024)
Opinion diversity	Opinion Dynamics	Diversity is measured as the standard deviation of the final opinion distribution, capturing the extent of opinion variation among agents at the end of the simulation.	(Chuang et al., 2023a)
Overlapping ratio	Decision Making	Overlapping ratio is measured by calculating the proportion of predicted words from the Guesser that match the target words, providing an objective, annotation-free metric for evaluating model performance in the Codenames Collaborative task.	(Wang et al., 2024f)
Partisan bias	Opinion Dynamics	Partisan Bias is measured by calculating the average difference in normalized group means between Democratic and Republican groups for each question, adjusted by the expected direction of human bias	(Chuang et al., 2023b)
Perceived reflection on the development of essential non-cognitive skills	Educational Training	Perceived reflection of non-cognitive skills is measured through a customized 7-point Likert scale questionnaire, where users rate how well the system supports specific skills such as self-perception, motivation, perseverance, self-control, metacognition, social competencies, resilience, and creativity, based on established psychological frameworks.	(Yan et al., 2024)
Perceived social support	Simulated Individual	Perceived social support is measured using a multi-item questionnaire where participants rate statements about the CA's supportiveness, persona, and relationship quality—covering aspects like care, helpfulness, encouragement, coherence, and emotional connection—to evaluate users' subjective experience with the agent.	(Liu et al., 2024b)
Perception consistency	Psychological Experiment	Perception consistency is measured by analyzing the agent's perceived safety and liveliness scores across different scenarios or environments	(Verma et al., 2023)
Performance deviations	Decision Making	Performance deviations are measured by score changes in Game Theory tasks under different pressure conditions (e.g., competitive and outcome pressure), comparing models with high and low self-consciousness personas against a baseline, with significance determined by statistical tests	(Kim et al., 2024)
Persona Perception	Simulated Individual	Persona Perception is measured using the Persona Perception Scale (PPS), with data structured at the participant-persona dyad level and validated through Confirmatory Factor Analysis to ensure the instrument's reliability and construct validity for repeated measures.	(Salminen et al., 2024; Shin et al., 2024; Ha et al., 2024; Chen et al., 2024b)
Personality	Educational Training	Personality is measured using 15 items from the BFI-2-XS on a 5-point Likert scale, assessing participants' perceptions of the agent's Big Five personality traits.	(Sonlu et al., 2024)
Personality	Educational Training	Personality is measured using the 44-item Big Five Inventory (BFI), where the model responds to descriptive statements on a 5-point Likert scale, and the resulting scores are mapped to the five personality traits to evaluate personality expression in tutoring contexts.	(Liu et al., 2024d)

Continued on next page

Metrics	Task	Implementation	Source
Personality	Psychological Experiment	Personality is measured using the MBTI framework for general traits and SD3 for negative traits, with shaping modeled as a function of identity and the attitudes of close agents within social networks, capturing how social context and subjective consciousness influence the development of agent personalities.	(He and Zhang, 2024)
Personality	Psychological Experiment	Personality is measured by having each persona complete the BFI-10, a short version of the Big Five Inventory, with responses programmatically collected to assess traits across the five major personality dimensions.	(de Winter et al., 2024)
Personality	Psychological Experiment	Personality is evaluated by varying individual Big Five traits and measuring their correlation with behavioral outcomes in game scenarios to assess how specific traits influence social decision-making.	(Bose et al., 2024)
Personality	Psychological Experiment	Personality is measured by creating LLM personas with distinct traits, administering a personality assessment, and analyzing their story outputs using LIWC, followed by both human and LLM evaluations that rate the stories across six dimensions and infer the intended personality traits from the narratives.	(Jiang et al., 2023b)
Personality	Simulated Individual	Personality is induced using naive or word-based prompts targeting Big Five traits, and evaluated with brief personality inventories to measure the effectiveness of each method.	(Jiang et al., 2023a)
Personality	Simulated Society	Personality is measured by having agents repeatedly complete the BFI-44 questionnaire throughout the simulation to track objective changes in Big Five traits over time, with a parallel multi-day assessment process designed to capture and compare personality drift or variability.	(Li et al., 2024b)
Personality	Simulated Society	Personality is assessed explicitly by prompting LLM agents with Big Five Inventory (BFI) statements and collecting their responses on a 5-point Likert scale, following standard psychological methods to measure traits across the five personality dimensions.	(Frisch and Giulianelli, 2024)
Persuasion	Simulated Society	Persuasion is measured by analyzing the distribution of successful persuasion outcomes over time and computing odds ratios from logistic regression, which quantify the likelihood of goal achievement based on conversation dynamics and goal types.	(Campedelli et al., 2024)
Positively mention rate	Simulated Individual	Positively Mention Rate is measured as the percentage of positive adjective prompts that receive favorable responses, conditioned on the target country or region.	(Kamruzzaman and Kim, 2024)
Prediction accuracy	Decision Making	Prediction accuracy is measured by calculating the proportion of correct predictions made by each expert, comparing predicted labels to ground-truth labels across all instances where the expert appears, to assess reliability and reduce the impact of hallucinations.	(Wang et al., 2024b)
Pressure	Simulated Individual	Pressure is measured using the Perceived Stress Scale (PSS-10) at multiple time points and through daily self-reports on relief from CA interactions, assessing changes in perceived stress over time.	(Liu et al., 2024b)
Probability weighting	Decision Making	Probability weighting is measured by comparing the LLMs' responses to normative probability-based decisions, identifying whether they overweight small probabilities or underweight large ones	(Jia et al., 2024)

Continued on next page

Metrics	Task	Implementation	Source
Professional scale	Simulated Individual	Professional scale is measured by evaluating agents' fidelity of role representation across occupational dimensions, with higher scores indicating accurate alignment with assigned professions and lower scores reflecting effective differentiation from unrelated roles.	(Sun et al., 2024)
Rated statement quality	Simulated Individual	Rated statement quality is measured by collecting human ratings on a 0–10 Likert scale for generated responses across scenarios, evaluating their impact on the communicator's goal, and comparing the average scores of the framework's selections against baseline models to assess effectiveness and alignment with human judgment.	(Liu et al., 2023)
Rationality	Opinion Dynamics	Rationality is assessed through manual annotation on a five-point scale, evaluating the agent's reasoning based on clarity, relevance, emotional coherence, and consistency with its profile, with higher scores indicating human-like, contextually appropriate responses.	(Lv et al., 2024)
Rationality of the agent memory	Psychological Experiment	The rationality of the agent memory is evaluated by comparing the believability of memory functions—summarizing short-term memory and generating long-term reflections—against non-expert human outputs, with human annotators judging which result appears more human-like or indistinguishable.	(Wang et al., 2023b)
Realism	Decision Making	Realism is measured by assessing the plausibility and believability of the simulation within the given scenario, determining how naturally the simulated behaviors and events align with real-world expectations.	(Li et al., 2024e)
Rephrase accuracy	Opinion Dynamics	Rephrase Accuracy measures the agent's ability to provide correct responses to prompts that are semantically equivalent but syntactically varied, evaluating the robustness of knowledge across different phrasings, and is defined as the proportion of matching responses between rephrased and original prompts.	(Ju et al., 2024)
response accuracy	Simulated Society	Human evaluators rank the believability of agent response from the least to the most	(Park et al., 2023)
response accuracy	Simulated Society	The system triggers controlled social scenarios and measures changes in relationship scores to assess the appropriateness of agent responses	(Gu et al., 2024)
Response quality	Writing	Response quality is computed using a composite reward function that combines (1) contextual alignment and user relevance measured by BERTScore-F1 and (2) fluency and non-repetitiveness measured by perplexity and BERTScore similarity to the previous response	(Mishra et al., 2023)
Response quality	Writing	Response quality is evaluated using both automatic metrics—BLEU-N, ROUGE-L, METEOR, and classification accuracy—and human ratings of coherence, fluency, and engagingness to comprehensively assess generation effectiveness and interaction quality.	(Li et al., 2024a)
Risk preference	Decision Making	Risk preference is measured by analyzing the LLMs' choices in scenarios involving uncertainty, identifying patterns of risk aversion or risk-seeking behavior based on deviations from expected utility maximization.	(Jia et al., 2024)
Roleplay subset of MT-Bench	Simulated Individual	The roleplay subset of MT-Bench evaluates RPLA performance using 2-turn dialogues across predefined role-playing scenarios, with GPT-4 serving as the evaluator to assess response quality in alignment with the benchmark's multi-category design.	(Tang et al., 2024)

Continued on next page

Metrics	Task	Implementation	Source
Self-esteem	Simulated Individual	Self-esteem is measured through a composite self-report score, with changes across intervention conditions analyzed using a Welch one-way ANOVA due to unequal variances, revealing no statistically significant effects.	(Pataranutaporn et al., 2024)
Self-Reflection	Simulated Individual	Self-reflection is measured through a composite self-report score, with changes across intervention conditions analyzed using a one-way ANOVA to assess the impact of different interventions on participants' reflective thinking.	(Pataranutaporn et al., 2024)
Selfishness	Decision Making	Selfishness is measured using the Selfishness Index (SI), which quantifies how much a player prioritizes personal gain across rounds, and the Difference Index (DI), which captures the deviation of a player's selfishness from the group average, highlighting relative selfish behavior in multi-agent game settings.	(Kim et al., 2024)
Sense of immersion	Educational Training	Sense of immersion is measured through usability testing feedback, where participants report a heightened feeling of immersion and authenticity during generative AI-driven educational simulations.	(Lee et al.)
Sense of immersion / Perceived immersion	Educational Training	Sense of immersion is measured through user-reported experiences during usability testing, where participants assess the authenticity and engagement of the simulation, often attributed to the generative AI's ability to produce dynamic and unpredictable interactions.	(Lee et al.)
Sensitivity to personalization	Simulated Individual	Sensitivity to personalization is measured by comparing LLM outputs before and after adding socio-demographic attributes, using Cohen's K to assess agreement on labels for ambiguous posts—where lower K values indicate greater sensitivity and stronger effects of personalization.	(Giorgi et al., 2024)
Simulation capability	Decision Making	Simulation capability is measured using a Turing test, where human annotators compare LLM-generated responses to human responses in policy execution scenarios and assign rationality labels to assess how realistically the LLM simulates human behavior.	(Ji et al., 2024)
Skill	Simulated Individual	Multiple skill-based evaluations, including comprehension and completeness metrics, are used to assess the model's effectiveness in performing complex tasks requiring accurate understanding and thorough responses.	(Shin et al., 2024)
Societal Fairness	Decision Making	Societal Fairness is measured using variance in per-capita living area, inverse order pairs in allocation, Gini coefficient of house distribution, and social welfare gap between vulnerable and non-vulnerable groups.	(Ji et al., 2024)
Societal Satisfaction	Decision Making	Societal Satisfaction is measured using average per-capita living area size, average individual waiting time, and social welfare, which reflects the cumulative satisfaction of all participants.	(Ji et al., 2024)
Stance Alignment	Opinion Dynamics	Stance alignment is measured by classifying generated content into support, neutral, or oppose, and further quantified using the mean absolute error (MAE) of attitude scores to capture the degree of alignment with expected positions.	(Mou et al., 2024c)
Story	Psychological Experiment	Story evaluation involves both human and LLM evaluators rating stories on six dimensions—readability, personalness, redundancy, cohesiveness, likeability, and believability—and inferring the personality traits of the LLM personas based on narrative content.	(Jiang et al., 2023b)

Continued on next page

Metrics	Task	Implementation	Source
Story Content Generation	Educational Training	Story content generation is measured using a narratives staging score based on the five-act structure, where each script is segmented by word count and analyzed for language trends and shifts across acts to evaluate narrative coherence and development.	(Yan et al., 2024)
Success rate	Educational Training	Success is measured by comparing criterion function outputs before and after operation across scenarios, focusing on agents' ability to identify capable candidates, propose accurate workflows, and correctly assign roles	(Li et al., 2023a)
User experience	Educational Training	User experience was measured through a 9-item questionnaire on a 7-point Likert scale, assessing perceived intelligence, enjoyment, usefulness, trust, sense of connection, and human-likeness for each AI tutor.	(Cheng et al., 2024)
Utility	Decision Making	Utility is measured through intrinsic utility functions representing each agent's normalized satisfaction based on offer price, and a joint utility function—inspired by the Nash bargaining solution—that quantifies the fairness of outcomes as the product of buyer and seller utilities.	(Huang and Hadfi, 2024)
Valid Response Rate	Decision Making	Valid Response Rate is used to assess whether the LLMs' sent amounts fall within the allowed monetary limits	(Xie et al., 2024a)
Validity	Simulated Individual	Validity is assessed through Confirmatory Factor Analysis (CFA) comparing BRASS- and human-generated items, evaluating convergent validity via average variance extracted (AVE >0.5) and item reliability, with most items performing well except one outlier with low factor loading, underscoring the need for human review.	(Ke and Ng, 2024)
Web search quality	Decision Making	Web search quality is measured using Mean Reciprocal Rank (MRR) to assess the accuracy of the top relevant result and Normalized Discounted Cumulative Gain (NDCG@1 and NDCG@3) to evaluate the overall ranking quality against an ideal ordering of relevance.	(Ren et al., 2024a)
Willingness to speak	Educational Training	Willingness to speak is measured by assigning willingness intensity scores to students when a question is posed, reflecting their likelihood to respond based on individual character traits, and compared against random selection to highlight the alignment between personality and participation.	(Shi et al., 2023)
Win rates	Simulated Individual	Win rates are measured as the proportion of games won by agents, used to evaluate their overall performance, emergent behaviors, and strategy effectiveness across different game setups.	(Chi et al., 2024)
Wisdom of Partisan Crowds Effect	Opinion Dynamics	The Wisdom of Partisan Crowds Effect is measured by calculating the reduction in normalized group error over time, quantifying how much LLM agent group estimates move closer to the ground truth through social interaction, with more negative indicating stronger collective improvement.	(Chuang et al., 2023b)