Evaluating the LLM Agents for Simulating Humanoid Behavior

Chaoran Chen
cchen25@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Bingsheng Yao
b.yao@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Yanfang Ye
yye7@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Dakuo Wang
d.wang@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Toby Jia-Jun Li
toby.j.li@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

ABSTRACT

Large Language Model (LLM)-based agents have showcased remarkable abilities in simulating human-like behavior, prompting widespread application across multiple fields. The growing use of these agents brings to light the critical need for robust evaluation metrics to assess their performance and for clear guidelines to direct their use across various downstream tasks. In this literature review, we identify three primary challenges in the evaluation of LLM agents: the overlook of evaluation metrics, the absence of a detailed taxonomy for aligning simulated humanoid behavior data with specific downstream tasks, and the disconnect between agent-oriented and task-oriented metrics. To tackle these issues, we summarized existing evaluation metrics and developed a comprehensive taxonomy for the research goals and the evaluation of LLM agents in simulating humanoid behavior. Through a systematic literature review, we aim to provide guidance for researchers in evaluating such LLM agents, ultimately mitigating the gap between the evaluation and the downstream tasks.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

large language models, evaluation methods, generative agent, simulation, humanoid behavior

1 INTRODUCTION

The remarkable capabilities of large language models (LLMs) in producing nuanced, human-like knowledge and behavior have sparked significant interest in using LLM agents to mimic humanoid behavior across various domains, including social system audits [40, 41], entertainment [5], education [34], privacy [7], psychology [3], economics [24], and judicial trials [21]. As shown in Fig. 1, the diverse applications of LLM agents have led to the emergence of distinct terminologies to describe their functionalities, including social simulacra, digital twin, and persona-based role-playing, each with its unique focus:

(1) **Persona-based role-playing** enables human to act as fictional characters and experience their emotion and behavior. These fictional characters *represent a specific individual or group*, detailed with attributes such as demographic background, preferences, goals, and habits [7, 9].
(2) **Digital twin** aims to create a *digital replica* of a real person and the environment in cyberspace, encompassing a wide range of human attributes from physical and physiological traits to cognitive performance and emotional states [35].
(3) **Social simulacra** is designed to generate multiple agents to *simulate social interactions* within a specific social space, defined by its goals and rules [41].

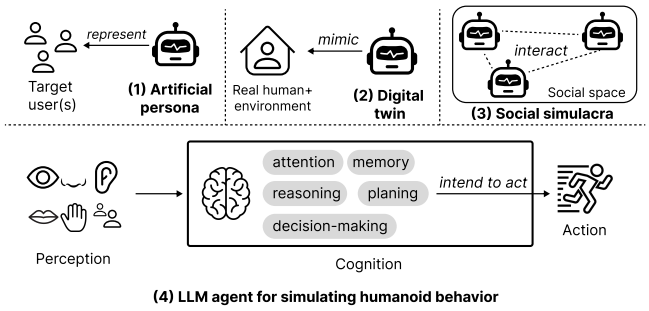

Figure 1: Comparison of terms about LLM agents for simulating humanoid behavior

The three types of LLM agents are subsets of the LLM agents studied in this literature review. Despite their differences, we distill a unified definition of LLM agents for simulating humanoid behavior through the perception-cognition-action model [15]: an LLM agent for simulating humanoid behavior is *a digital entity that can replicate human-like interactions and personalities, exhibit cognitive skills for reasoning, decision-making, and planning, tackle diverse or complex tasks, and execute non-predefined actions.*

As such LLM agents become increasingly sophisticated and their applications are getting more widespread, they underscores the pressing need for both robust evaluation metrics to measure their performance and clear guidelines for their application in diverse downstream tasks. A systematic evaluation method is essential to ensure that LLM agents can generate plausible human cognition and behavior, maintain accuracy, consistency, stability, and safety, as well as prove effective in real-world downstream tasks. Existing evaluation metrics can be categorized into two types:

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

(1) **Agent-oriented** metrics: These metrics evaluate the intrinsic and task-agnostic performance of LLM agents, including believability of behavior [40], memorization [43], consistency [43, 50, 52], hallucination [43], controllability [4], exaggeration [9], robustness [43, 52], diversity [9, 10], and empathy [7].
(2) **Task-oriented** metrics: These metrics are pertinent to the specific downstream tasks performed by LLM agents, such as the accuracy of responses to questions [49] and the success rate in collaboration [31].

However, there are three main challenges when evaluating LLM agents in simulating humanoid behavior for downstream tasks:

(1) The effectiveness of existing evaluation metrics has not been widely recognized.
   (a) One reason is the lack of uniformity in the definition of evaluation metrics. For instance, Xiao et al. [52] proposed consistency and robustness as two effective metrics for measuring simulation believability. However, Cheng et al. [9], based on its definition in the entertainment field, consider believability to be the agent's ability to provide the illusion of life, and thus permits the audience's suspension of disbelief. Therefore, they criticized that believability is susceptible to the biases and fallacies of human judgment.
   (b) Another reason is that even if the definitions of evaluation metrics are consistent, different evaluation purposes can lead people to have different attitudes towards a metric. For example, Shao et al. [43]. used memorization as one of the evaluation metrics to assess whether an agent can accurately portray a certain group of people. However, the evaluation purpose of Cheng et al. [9]. was to assess the agent simulation's susceptibility to caricature (the degree to which it overemphasizes unique traits of the agent that stand out more than a relevant response to the situation). This led them to believe that memorization only reproduces already-known behavior and does not facilitate new insight into human behavior.
(2) The existing evaluation metrics lack a clear taxonomy to map the generated data into whether it is related to downstream tasks, and whether it can be obtained by automatic calculation or through human evaluation.
(3) There is a gap between agent-oriented evaluation metrics and task-oriented evaluation metrics.
   (a) Task-oriented metrics directly measure agents performance in downstream tasks. However, it is not only necessary to evaluate the outcomes of the task itself but also to assess the data generated by the agent, as the quality of the generated data directly affects the agent's performance in downstream tasks and can guide the fine-tuning of agent simulation.
   (b) Agent-oriented metrics target the generated data itself, examining lower-level data validity without fully considering the downstream task during the evaluation phase. This means that further abstraction and examination of the generated data based on the downstream task are not conducted. Therefore, even if an agent passes some metrics, it cannot truly prove the effectiveness of the data in downstream tasks. For example, an agent might generate a sequence of latitude and longitude coordinates that seems reasonable, but if these are actually applied to a downstream task such as simulating a person's movement trajectory, then the trajectory formed by these coordinates might not match the trajectory of a real person.

To address these challenges, it is necessary to summarize the existing evaluation practices of LLM agents in simulating humanoid behavior. Yet, to the best of our knowledge, no survey has proposed a systematic literature review about the evaluation methods and deployment guidance for such agents.

Hence, we want to address the following three research questions in this literature review:

- **RQ1:** What are existing evaluation metrics for LLM agents in simulating human behavior?
- **RQ2:** What is a comprehensive taxonomy to cover and categorize the evaluation metrics?
- **RQ3:** How can we identify a suitable evaluation method by taking both the agents and the downstream tasks into account?

Since evaluating LLM agents is a relatively new direction, we surveyed both peer-reviewed papers from top-tier venues and preprint ones available on Arxiv. The final corpus comprises 45 papers, meticulously selected from a pool of 815 publications since 2020, in which the OpenAI published GPT-3.

In summary, by holistically examining the previous evaluation metrics of LLM agents in simulating humanoid behavior, we find that research in this field has significantly increased in the past six months and has diverse applications, ranging from HCI to AI and software engineering. Research on using LLM agents to simulate humanoid behavior identifies two main goals: simulating human-like behavior (e.g., social interactions and cognitive dynamics) and applying simulations for further research. We summarize the existing evaluation metrics and build a comprehensive taxonomy to cover and categorize them, which elucidate their applicable conditions and bridge the gap between agent-oriented metrics and task-oriented metrics. The detailedness of key demographic and psychological attributes essential for crafting these simulations leans towards a less intricate portrayal, indicating a strategic focus on capturing essential human traits within current technological constraints. We hope this literature review and the taxonomy can not only steer researchers towards conducting more appropriate evaluations, but also mitigate the gap between the evaluation and the application of LLM agents for simulating humanoid behavior.

2 LITERATURE REVIEW METHOD

We conduct a *systematic literature review (SLR)* to address our research questions. Following previous guideline [38], we aim to identify all relevant research papers that focus on applying LLM agents to simulate humanoid behavior and give a balanced and unbiased summary of the literature. We used four databases, including Google Scholar, ACM Library, IEEE Xplore, and ACL Anthology, as they are the primary online academic databases, and are widely used in prior literature review [44, 53]. We did not restrict the publication venues because many of the LLM related paper are too timely to be formally published in the peer-reviewed conferences

Table 1: Inclusion criteria (IC) and exclusion criteria (EC).

Criteria	
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 planing.
IC-3	The LLM agents in the paper are capable to complete complicate and general tasks.
IC-4	The LLM agents' action set in the paper is neither predefined nor finite.
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 agents' actions are either predefined or limited.

or journals. Below, we detail our literature selection process, including the scope of our literature review, the inclusion/exclusion criteria, and the search terms.

2.1 Scope of literature review

Our literature review on LLM agents for simulating humanoid behavior is centered on agents that emulate cognitive processes. As depicted in Figure 1, these agents are primarily designed to replicate human decision-making and reasoning, transforming cognition into deliberate actions. Consequently, we scrutinized whether the studies showcased the LLM agents' ability to simulate human-like behavior, particularly cognitive processes, within their *research objectives, methodologies, or evaluation techniques*. Studies merely concentrating on the perception or action capabilities of LLM agents were excluded:

(1) Studies on perception, such as those involving LLM agents in autonomous driving or robotics that utilize sensors or computer vision for data collection, do not align with our focus on simulation and were omitted from our review.
(2) We also excluded studies where actions are predefined or limited (e.g., reinforcement learning agents operating within a specific and close-ended task) or focus on concrete action planning (e.g., a robot plans the movement trajectories in the physical world).

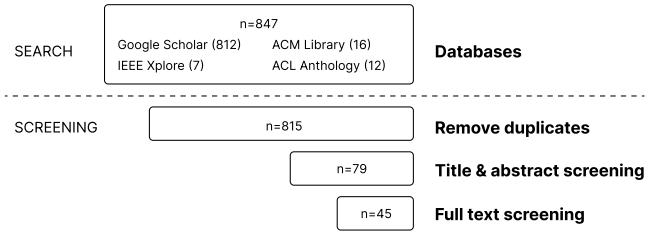

Figure 2: Screening process.

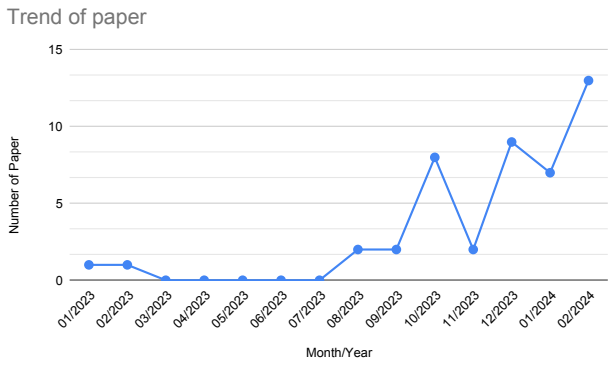

Figure 3: Trend of paper counts each month in the selected paper (2023-2024).

2.2 Inclusion/exclusion criteria, search query and screening process

In line with our research scope, we filtered papers using selection criteria, summarized in Table 1. To be selected, a primary study must satisfy all the inclusion criteria and no exclusion criteria.

We search the four databases using this string and retrieved a total of 847 papers: `("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*`.

Fig. 2 shows the screening process. After removing duplicates, 815 papers remained. These papers were independently screened by two authors based on reading the paper titles and abstracts to determine if they met the inclusion criteria. The independent screening results were compared, and Cohen's Kappa was calculated to be 0.844, indicating a strong inter-rater reliability. If at least one author considered a paper eligible, it proceeded to the full text screening stage, where two authors would read the full text and discuss whether to include it. Our final set of selected primary studies has a total of 45 publications. Table 2 shows the final selected papers grouped by the type of metrics.

3 RESULTS AND FINDINGS

3.1 Bibliometrics

Fig. 3 shows the trend of paper counts for each month from January, 2023 to Febuary, 2024. It shows that research in this field has significantly increased in the past six months. Within the 45 selected

Agent-oriented metrics
Park et al. [40], Lv et al. [33], Shao et al. [43], Cheng et al. [9], [20], Gerosa et al. [18], Jin et al. [25], Cai et al. [4], Wang et al. [48], Zhang et al. [57], Chen et al. [7]
Task-oriented metrics
Sreedhar and Chilton [45], Xie et al. [54], Wang et al. [49], Gao et al. [17], Li et al. [31], Park et al. [40], Wu et al. [51], Frisch and Giulianelli [16], Lu et al. [32], Mitsopoulos et al. [37], Chan et al. [6], Chuang et al. [10], He and Zhang [22], Pang et al. [39], Chuang et al. [12], Coletta et al. [13], Antunes et al. [1], Zhang et al. [56], Jinxin et al. [26], Leng and Yuan [29], Li et al. [30], Taubenfeld et al. [46], Verma et al. [47], Gui and Toubia [19], Zhao et al. [58], Kaiya et al. [27], Lee et al. [28], Chen et al. [8], Jin et al. [25], Wang et al. [48], Benharrak et al. [2], De Winter et al. [14], Zhou et al. [60], Mishra et al. [36], Salminen et al. [42], Chen et al. [7]

Table 2: Selected paper

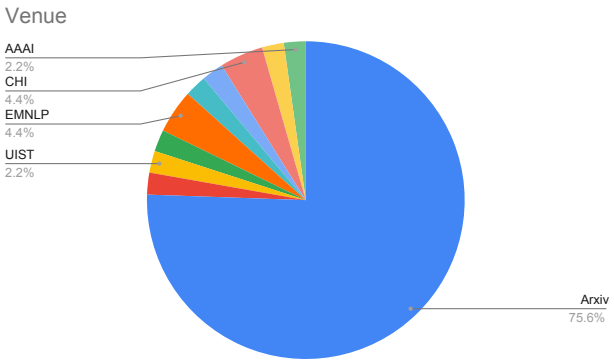

Figure 4: Venues of the selected paper.

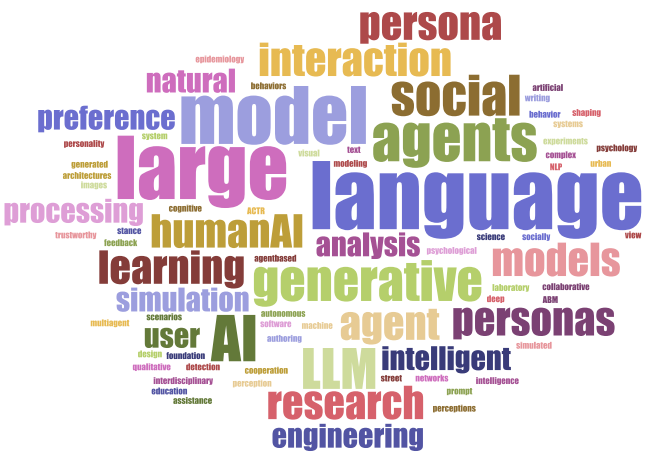

Figure 5: Keywords of the selected paper.

papers, 35 were preprint on Arxiv or ResearchGate, 2 at CHI, 2 at EMNLP, 1 at UIST, 1 at AAAI, 1 at NeurIPS, 1 at ASE, 1 at CHBAH, and 1 at IVA, as shown in Fig. 4. This suggests that research on LLM agents for simulating humanoid behavior is in a timely phase, with the majority of papers in preprint. The applications of LLM agents for simulating humanoid behavior extend across diverse venues, ranging from HCI to AI and software engineering. Fig. 5 shows the keyword cloud for the selected paper. Most of the keywords include "large language model(s)", "LLM", "generative agent(s)", "persona(s)", "human-AI interaction" and "simulation", aligning with our query.

3.2 Findings

3.2.1 Summary of existing agent-oriented evaluation metrics. We summarized previous agent-oriented metrics into 9 categories:

(1) **Believability of behavior** [40]: The degree to which agents' actions and responses in a simulated environment appear realistic and convincing to human.
(2) **Memorization/Replication** [9, 18, 43]: The agent's ability to recall relevant information about the character being portrayed.
(3) **Consistency** [4, 7, 18, 20, 43, 50, 52, 57]:
  (a) Consistency in agent's attitudes, values, preference, etc.;
  (b) Consistency in agent's speaking styles, tones, and plots;
  (c) Consistency in emotions (i.e., the change of emotions should be continuous but not volatile);
  (d) Consistency in reactions to context.
(4) **Hallucination/Incredibility** [7, 18, 43]: The agent's ability to discard knowledge and skills that it should not have.
(5) **Controllability** [4]: Measuring whether altering psychological traits can cause noticeable different behaviors for the agent.
(6) **Exaggeration** [9]: The degree to which the agent is tailored for specific traits compared with a neutral/base agent.
(7) **Robustness/Stability** [43, 52]: The agent's ability to maintain robust when faced with perturbations, especially after prolonged periods of acting.
(8) **Diversity/Individuation** [9, 10, 18, 20]:
  (a) Diversity of multiple agents in specific attributes (e.g., opinions);
  (b) Diversity for the global profile, i.e., differentiability from other agents.
(9) **Empathy** [7, 42]: Measuring the leval of human's emotional or cognitive empathy toward the agent.

3.2.2 The taxonomy of research goals. The research goals for using LLM agents to simulate humanoid behavior can be categorized into two primary types: 1) for the purpose of simulating humanoid behavior and 2) by means of simulating humanoid behavior (Table 3).

Research goals	Detailed categories	References
For the purpose of simulating humanoid behavior	Simulating social behavior	collaboration and social learning [6, 8, 29, 49, 51], social networking [17, 23, 27, 37, 40], task coordination [31],trust [54], competition [58]
	Simulating cognitive dynamics	human strategies [45], subrational behavior [13], opinion dynamics [10, 33], personality [4, 14, 22], social scene [1]
	Simulating the digital twins of famous figures	[43, 60]
	Simulating human conversation	[56]
By means of simulating humanoid behavior	Investigating problems in the simulation	the influence of treatment variation on LLM's generation [19], the bias of LLM generated personas [20], the challenges in human-AI collaboration [59]
	Categorizing the scope of simulation	[9]
	Evaluating the generated data	the causal relationship of behavior over time [32], the personality consistency and linguistic behavior of LLM [16], the alignment between LLM and human value [39, 57], the influence of portrait image on user perception for LLM-generated personas [42]
	Assisting the research in a specific domain	education [25, 26], economics [30], game [55], software engineering [18], writing [2], politics [11, 46], psychotherapy [36], privacy [7], urban planing [47]

Table 3: The taxonomy of research goals.

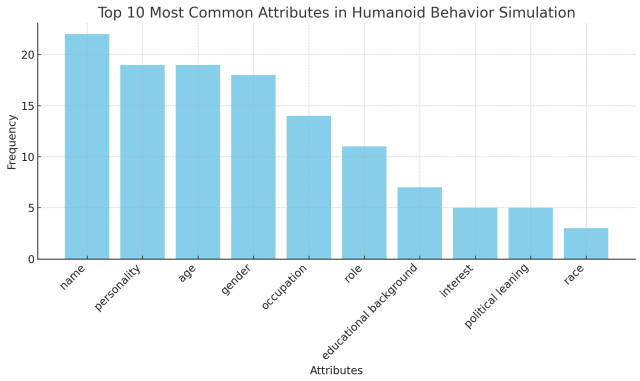

Figure 6: The 10 most common attributes in the LLM agents for simulating humanoid behavior.

In the first category, the focus is on simulating various aspects of human-like behavior through LLM agents, such as studying social interactions and cognitive processes. This encompasses efforts to replicate social behaviors such as trust and competition, and understanding cognitive dynamics, which involves human strategies and subrational behaviors. The second category simulates humanoid behavior for further evaluation or research in specific domain. It includes the analysis of the simulations themselves, such as identifying the issues that arise when simulating different treatment conditions and investigating the biases of LLM-generated personas. The research also focuses on the scope and validity of these simulations, assessing their impact and exploring the causal relationships of agent behaviors over time. Among the most impactful directions are the applications of these humanoid simulations in a variety of specific domains, such as education, economics, and writing, signifying a broad relevance and utility in practical contexts.

3.2.3 The taxonomy of evaluation metrics. Table 4 shows the taxonomy of evaluation metrics, categorized by two dimensions (i.e., agent-oriented vs. task-oriented and automatic vs. human). Most of the metrics are task-oriented and automatic. This highlights a predominant focus on automated evaluation of downstream tasks in prior research, driven partly by the higher costs and complexities of human assessment, and the straightforward insights derived from task-specific evaluations. However, enhancing an agent's downstream task performance requires more than just assessing task outcomes. The quality of data produced by the agent is crucial, as it directly influences task performance and informs the agent's simulation fine-tuning. Sole reliance on agent-centric metrics also falls short, since passing certain metrics doesn't guarantee the effectiveness in downstream applications.

3.2.4 The analysis of LLM agent attributes. In simulating humanoid behavior, research concentrates on key demographic attributes (Fig. 6) such as name, personality, age, gender, and occupation. Attributes like role, educational background, personal interest, political leaning, and race are also significant, highlighting the value placed on social and psychological facets of human behavior. We rated the detailedness of these attributes in a 5-point scale. The results shows a favor of less intricate portrayal, with the majority of attributes falling in between 1 and 3. This suggests a deliberate focus on capturing the essence of human traits while maintaining a practical balance in the granularity of the simulation to suit current technological capabilities and application needs.

Table 4: The taxonomy of evaluation metrics.

	Automatic	**Human**
Agent-oriented	personality, stability, hallucination, values, individuation, exaggeration, controllability, memorization consistency, generation diversity, bias, credibility, consistency, clarity, empathy	believability of behavior
Task-oriented	*Economics/game theory*: completeness of strategies, consistency of strategies with personality, adherence to the strategies, valid response rate, distribution of amount sent, behavioral alignment (trust rate, lottery rate, behavior dynamic), reward, probability distribution of action, inflation rate, unemployment rate, nominal GDP, nominal GDP growth, wage inflation, real GDP growth, expected monthly income, consumption, purchase probability, expected competing product price, distributions of number choices, collaboration type, price per round, accumulated escaped count.	
	Simulated society: prediction accuracy, AUC, F1 score, MSE, MAE, prediction error rate, temporal dissemination of events, emotional density, positive attitude rate, information diffusion, relationship formation, coordination within other agents, task completion time, LLM rated winning rate, complexity of generated content, dialogue generation quality, feasibility of action plans, guess accuracy, probability of social connection formation, probability of reporting own draw, probability of agents' motivations, distribution preferences, degree of reciprocity, percent of social welfare maximization choices, customer counts, imitation and differentiation behavior, proportion of similar and different dishes, average dish scores, price consistency between competitors, police success rate, club preference, accuracy of information gathering, probabilities of receiving, storing, and retrieving the key information across the population, rationality of the agent memory, information entropy, attitude change, success rate for coordination (identification accuracy, workflow correctness, alignment between job and agent's skill).	*Simulated society*: comprehensive efficiency, effectiveness, and usability of the system, human rated winning rate, relevance of the created artefacts, ramification of the scenarios, errors in the prompting sequence
	Writing: factual error rate, LIWC counts, consistency with the scenario and characters, quality and logical coherence of the script content, text understanding, creative writing abilities, reasoning abilities.	*Writing*: quality of feedback.
	Public health: correlation between predicted and real results.	
	Cognitive modeling: authenticity, rationality.	
	Question-answering system: classifcation accuracy, Kappa correlation coefficient, naturalness, coherence, engagingness, groundedness.	
	Opinion dynamics: bias, diversity, wisdom of crowd effect, human likeness index, extreme values, attitude score	
	Personality test: MBTI score, SD3 score, average happiness value per time step, Big Five Inventory score.	
	Teacher training: knowledge level of agents in understanding, implementation, and analysis; density of knowledge-building.	*Teacher training*: proportion of interaction behavior, willingness to speak, effectiveness of questioning.
	Urban studies: agent perceived safety, agent perceived liveliness	
	Conversation: consistency, human-likeness, engagement, quality, safety, correctness	
	dialogue for psychotherapy: gender-age consistency, persona consistency, psychotherapeutic approach correctness, politeness correctness, interpersonal behaviour correctness, perplexity, BERTScore-F1, Response-length	*dialogue for psychotherapy*: fluency, consistency, non-repetitiveness
		Algorithm audits: clarity, compassion, completeness, consistency, credibility, empathy, similarity, stereotypicality, transparency, human perceived connection between personas and system outcomes

REFERENCES

[1] 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*. 1–9.

[2] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2023. Writer-Defined AI Personas for On-Demand Feedback Generation. *arXiv preprint arXiv:2309.10433* (2023).

[3] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.

[4] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, et al. 2023. Digital Life Project: Autonomous 3D Characters with Social Intelligence. *arXiv preprint arXiv:2312.04547* (2023).

[5] Chris Callison-Burch, Gaurav Singh Tomar, Lara J Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. *arXiv preprint arXiv:2210.07109* (2022).

[6] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).

[7] Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jiajun Li. 2023. An Empathy-Based Sandbox Approach to Bridge Attitudes, Goals, Knowledge, and Behaviors in the Privacy Paradox. *arXiv preprint arXiv:2309.14510* (2023).

[8] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

[9] Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations. *arXiv preprint arXiv:2310.11501* (2023).

[10] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. Simulating Opinion Dynamics with Networks of LLM-based Agents. *arXiv preprint arXiv:2311.09618* (2023).

[11] Yun-Shiuan Chuang, Nikunj Harlalka, Siddharth Suresh, Agam Goyal, Robert D Hawkins, Sijia Yang, Dhavan V Shah, Junjie Hu, and Timothy T Rogers. 2024. The Wisdom of Partisan Crowds Comparing Collective Intelligence in Humans and LLM-based Agents. *ICLR 2024 Workshop on Large Language Model (LLM) Agents* (2024).

[12] Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. Evaluating LLM Agent Group Dynamics against Human Group Dynamics: A Case Study on Wisdom of Partisan Crowds. *arXiv preprint arXiv:2311.09665* (2023).

[13] Andrea Coletta, Kshama Dwarakanath, Penghang Liu, Svitlana Vyetrenko, and Tucker Balch. 2024. LLM-driven Imitation of Subrational Behavior: Illusion or Reality? *arXiv preprint arXiv:2402.08755* (2024).

[14] Joost CF De Winter, Tom Driessen, and Dimitra Dodou. 2023. The use of ChatGPT for personality research: Administering questionnaires using generated personas.

[15] Food, Drug Administration, et al. 2016. Applying human factors and usability engineering to medical devices: guidance for industry and Food and Drug Administration staff. *The Federal Register/FIND* 81 (2016).

[16] Ivar Frisch and Mario Giulianelli. 2024. LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models. *arXiv preprint arXiv:2402.02896* (2024).

[17] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984* (2023).

[18] Marco Gerosa, Bianca Trinkenreich, Igor Steinmacher, and Anita Sarma. 2024. Can AI serve as a substitute for human subjects in software engineering research? *Automated Software Engineering* 31, 1 (2024), 13.

[19] George Gui and Olivier Toubia. 2023. The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective. *arXiv preprint arXiv:2312.15524* (2023).

[20] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892* (2023).

[21] Sil Hamilton. 2023. Blind judgement: Agent-based supreme court modelling with gpt. *arXiv preprint arXiv:2301.05327* (2023).

[22] Zihong He and Changwang Zhang. 2024. AFSPP: Agent Framework for Shaping Preference and Personality with Large Language Models. *arXiv preprint arXiv:2401.02870* (2024).

[23] Pierre Hoes, Jan LM Hensen, Marcel GLC Loomans, Bert de Vries, and Denis Bourgeois. 2009. User behavior in whole building simulation. *Energy and buildings* 41, 3 (2009), 295–302.

[24] John J Horton. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report. National Bureau of Economic Research.

[25] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2023. " Teach AI How to Code": Using Large Language Models as Teachable Agents for Programming Education. *arXiv preprint arXiv:2309.14534* (2023).

[26] Shi Jinxin, Zhao Jiabao, Wang Yilei, Wu Xingjiao, Li Jiawen, and He Liang. 2023. Cgmi: Configurable general multi-agent interaction framework. *arXiv preprint arXiv:2308.12503* (2023).

[27] 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. *arXiv preprint arXiv:2310.02172* (2023).

[28] Unggi Lee, Sanghyeok Lee, Junbo Koh, Yeil Jeong, Haewon Jung, Gyuri Byun, Yunseo Lee, Jewoong Moon, Jieun Lim, and Hyeoncheol Kim. 2023. Generative Agent for Teacher Training: Designing Educational Problem-Solving Simulations with Large Language Model-based Agents for Pre-Service Teachers. (2023).

[29] Yan Leng and Yuan Yuan. 2023. Do LLM Agents Exhibit Social Behavior? *arXiv preprint arXiv:2312.15198* (2023).

[30] Nian Li, Chen Gao, Yong Li, and Qingmin Liao. 2023. Large language model-empowered agents for simulating macroeconomic activities. *arXiv preprint arXiv:2310.10436* (2023).

[31] Yuan Li, Yixuan Zhang, and Lichao Sun. 2023. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500* (2023).

[32] Jiaying Lu, Bo Pan, Jieyi Chen, Yingchaojie Feng, Jingyuan Hu, Yuchen Peng, and Wei Chen. 2024. AgentLens: Visual Analysis for Agent Behaviors in LLM-based Autonomous Systems. *arXiv preprint arXiv:2402.08995* (2024).

[33] Yaojia Lv, Haojie Pan, 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* (2024).

[34] Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. GPTeach: Interactive TA Training with GPT Based Students. (2023).

[35] Michael E Miller and Emily Spatz. 2022. A unified view of a human digital twin. *Human-Intelligent Systems Integration* 4, 1-2 (2022), 23–33.

[36] 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*. 13952–13967.

[37] Konstantinos Mitsopoulos, Ritwik Bose, Brodie Mather, Archna Bhatia, Kevin Gluck, Bonnie Dorr, Christian Lebiere, and Peter Pirolli. 2023. Psychologically-valid generative agents: A novel approach to agent-based modeling in social sciences. In *Proceedings of the AAAI Symposium Series*, Vol. 2. 340–348.

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

[39] Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. Self-Alignment of Large Language Models via Monopolylogue-based Social Scene Simulation. *arXiv preprint arXiv:2402.05699* (2024).

[40] 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*. 1–22.

[41] 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*. 1–18.

[42] Joni 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* (2024), 100052.

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

[44] Andy P Siddaway, Alex M Wood, and Larry V Hedges. 2019. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology* 70 (2019), 747–770.

[45] Karthik Sreedhar and Lydia Chilton. 2024. Simulating Human Strategic Behavior: Comparing Single and Multi-agent LLMs. *arXiv preprint arXiv:2402.08189* (2024).

[46] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. *arXiv preprint arXiv:2402.04049* (2024).

[47] 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 preprint arXiv:2312.13126* (2023).

[48] Lei Wang, Jingsen Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, and Ji-Rong Wen. 2023. RecAgent: A Novel Simulation Paradigm for Recommender Systems. *arXiv preprint arXiv:2306.02552* (2023).

[49] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing cognitive synergy in large language models: A task-solving

agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300* 1, 2 (2023), 3.

[50] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746* (2023).

[51] Zengqing Wu, Shuyuan Zheng, Qianying Liu, Xu Han, Brian Inhyuk Kwon, Makoto Onizuka, Shaojie Tang, Run Peng, and Chuan Xiao. 2024. Shall We Talk: Exploring Spontaneous Collaborations of Competing LLM Agents. *arXiv preprint arXiv:2402.12327* (2024).

[52] Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2023. How Far Are We from Believable AI Agents? A Framework for Evaluating the Believability of Human Behavior Simulation. *arXiv preprint arXiv:2312.17115* (2023).

[53] Yu Xiao and Maria Watson. 2019. Guidance on conducting a systematic literature review. *Journal of planning education and research* 39, 1 (2019), 93–112.

[54] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behaviors? *arXiv preprint arXiv:2402.04559* (2024).

[55] Ming Yan, Ruihao Li, Hao Zhang, Hao Wang, Zhilan Yang, and Ji Yan. 2023. LARP: Language-Agent Role Play for Open-World Games. *arXiv preprint arXiv:2312.17653* (2023).

[56] Dong Zhang, Zhaowei Li, Pengyu Wang, Xin Zhang, Yaqian Zhou, and Xipeng Qiu. 2024. SpeechAgents: Human-Communication Simulation with Multi-Modal Multi-Agent Systems. *arXiv preprint arXiv:2401.03945* (2024).

[57] Zhaowei Zhang, Nian Liu, Siyuan Qi, Ceyao Zhang, Ziqi Rong, Yaodong Yang, and Shuguang Cui. 2023. Heterogeneous Value Evaluation for Large Language Models. *arXiv preprint arXiv:2305.17147* (2023).

[58] Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2023. Competeai: Understanding the competition behaviors in large language model-based agents. *arXiv preprint arXiv:2310.17512* (2023).

[59] Qingxiao Zheng, Zhongwei Xu, Abhinav Choudhary, Yuting Chen, Yongming Li, and Yun Huang. 2023. Synergizing Human-AI Agency: A Guide of 23 Heuristics for Service Co-Creation with LLM-Based Agents. *arXiv preprint arXiv:2310.15065* (2023).

[60] Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832* (2023).