GenSim: A General Social Simulation Platform with Large Language Model based Agents

Jiakai Tang[1], **Heyang Gao**[1], **Xuchen Pan**[3], **Lei Wang**[1], **Haoran Tan**[1], **Dawei Gao**[3], **Yushuo Chen**[3], **Xu Chen**[1*], **Yankai Lin**[1], **Yaliang Li**[3], **Bolin Ding**[3], **Jingren Zhou**[3], **Jun Wang**[2], **Ji-Rong Wen**[1]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2] University College London
[3]Alibaba Group

Abstract

With the rapid advancement of large language models (LLMs), recent years have witnessed many promising studies on leveraging LLM-based agents to simulate human social behavior. While prior work has demonstrated significant potential across various domains, much of it has focused on specific scenarios involving a limited number of agents and has lacked the ability to adapt when errors occur during simulation. To overcome these limitations, we propose a novel LLM-agent-based simulation platform called *GenSim*, which: (1) **Abstracts a set of general functions** to simplify the simulation of customized social scenarios; (2) **Supports one hundred thousand agents** to better simulate large-scale populations in real-world contexts; (3) **Incorporates error-correction mechanisms** to ensure more reliable and long-term simulations. To evaluate our platform, we assess both the efficiency of large-scale agent simulations and the effectiveness of the error-correction mechanisms. To our knowledge, GenSim represents an initial step toward a general, large-scale, and correctable social simulation platform based on LLM agents, promising to further advance the field of social science. The relevant code and project are open-sourced on https://github.com/TangJiakai/GenSim.

1 Introduction

Social science, which focuses on human behavior, communication, and organization, is playing an increasingly significant role as world civilization advances. One important research paradigm in social science is collecting real human data. For instance, to study the effectiveness of positive psychology interventions, (Bolier et al., 2013) recruited over 6,000 participants to observe their responses to controlled experiments. While the paradigm of collecting real human data is widespread in social science research, it suffers from significant drawbacks, such as high cost, poor controllability, and challenges in reproducibility, which have troubled researchers for a long time.

In the field of artificial intelligence (AI), researchers have discovered that language serves as a crucial carrier of intelligence (Zhao et al., 2023), and the objective of "next-token prediction" using a massive training corpus (*i.e.*, large language models, LLMs) has the potential to achieve human-like intelligence. With the advent of these high-intelligence models, a new "AI for Social Science" direction has emerged: leveraging LLMs as proxies for real humans to conduct social science experiments (Gao et al., 2024). This approach provides the opportunities to fundamentally address the above challenges faced by social science research, potentially paving the way for an entirely new research paradigm. For instance, Generative agents (Park et al., 2023) leverages 25 agents to simulate human daily life, and finds that these agents can autonomously host parties and conduct mayoral election. RecAgent (Wang et al., 2023) simulates user online behaviors, and studies the phenomenons of information cocoon and conformity behaviors. EconAgent (Li et al., 2024b) studies the macroeconomic behaviors using LLM-based agents in the context of dynamic markets.

While the above methods have shown promising results, they are primarily limited to specific scenarios and small-scale simulations. Moreover, when discrepancies arise between simulated behaviors and those observed in the real world, existing methods lack effective error-correction mechanisms. To address these limitations, we introduce GenSim, a general social simulation platform based on LLM agents. In specific, to avoid reinventing the wheel in simulating various social scenarios, we propose a general programming framework composed of three key modules on single-agent construction, multi-agent scheduling, and environment setup.

*Corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 143–150

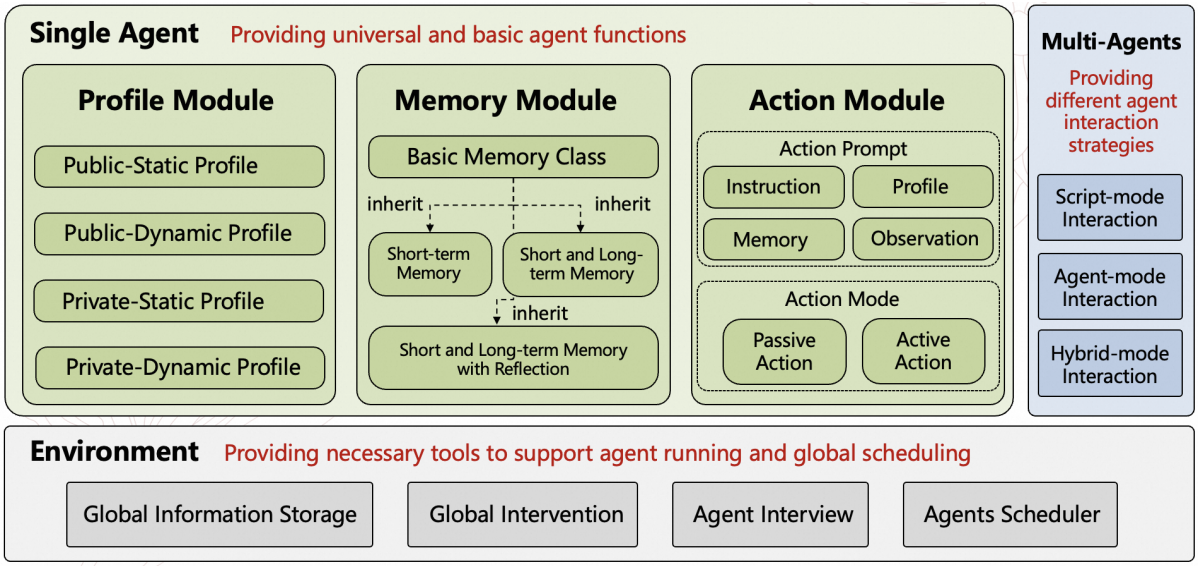

Figure 1: The general framework for social simulation

Additionally, we provide three default scenarios as references to help users quickly implement their customized simulations. To achieve large-scale simulations in real-world scenarios, we leverage distributed parallel technology to support one hundred thousands of agents in our platform. Finally, we design several error-correction mechanisms, allowing the platform to first perform self-evaluation or seek human feedback, and then fine-tune itself to ensure more reliable simulation. The comparison between our framework and the previous work can be seen in Figure 1.

In summary, the main contributions of this paper are as follows: (1) We propose a general, large-scale and correctable social simulation platform based on LLM agents. (2) We provide detailed usage examples to illustrate the capabilities of our platform. (3) We conduct a series of experiments to evaluate the platform's effectiveness and efficiency.

2 Features of GenSim

There are several unique features of our platform. To begin with, we abstract a set of general functions to facilitate any customized simulation scenario according to the users' requirements. Then, our platform supports one hundred thousand agents to better simulate large-scale populations in real-world contexts. At last, we provide a series of error-correction mechanisms to ensure more reliable simulation. The first two can be seen as static features from the generality and scalability perspectives, respectively, while the last one extends previous work from the dynamic perspective, making sure our platform can continually correct and improve itself.

2.1 General Simulation Framework

Our framework consists of three modules focusing on single agent, multi-agents, and environments (see Figure 1). In the **single agent** module, users can flexibly configure the agent's profile, memory, and action components. The profile includes both public information, such as gender, name, and birthplace, as well as private attributes such as income and health condition. To enable the agent to retain behaviors in various ways, users can assemble different memory components like short-term memory, long-term memory, and the reflection mechanism to build the agent's memory. The actions of the agents are driven by LLM prompts, where users can flexibly configure them to include agent profiles, memories and so on.

In the **multi-agents** module, inspired by the work (Zhou et al., 2024), we design two strategies for generating agent interactions: script mode and agent mode. In specific, in script mode, all interactions are treated as a whole and generated in a single call to the LLM. For example, one can directly prompt LLMs to generate a dialogue between a doctor and a teacher in one step. In this strategy, the LLM acts as a meta-agent, producing the dialogue from a third-person perspective. In agent mode, interactions are generated by different agents, each representing a distinct role, and each agent generates outputs from a first-person perspective. This interaction generation process

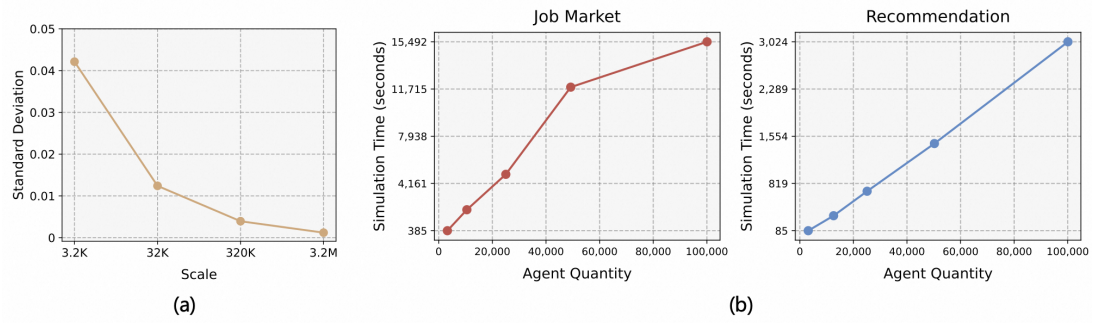

Figure 2: (a) The fluctuation of the simulation results with different agent scales. (b) Time costs with different numbers of agents

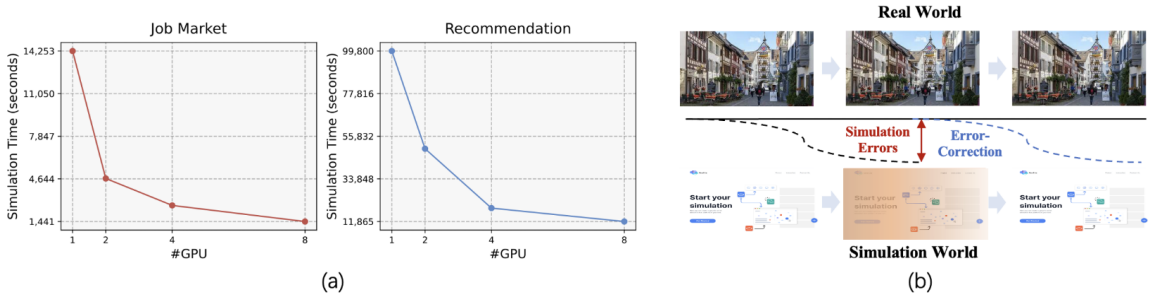

Figure 3: (a) Time costs with different numbers of GPUs. (b) The error-correction mechanism.

requires multiple calls to the LLM. In the example above, two agents are deployed, and each agent's output is determined by the complete history of their interactions.

In the **environment** module, we store all the information beyond the agents necessary for running the simulation, such as the recommendation algorithm used in a web user simulator (Wang et al., 2023). Additionally, we allow users to globally intervene in the platform, which is useful for counterfactual inferences. We also provide essential functions to facilitate interviewing, searching, and storing different agents.

Based on the above general framework, users can easily create customized simulations. To provide additional references, we offer three default scenarios: job market, recommender system, and group discussion. These scenarios can not only facilitate related research but also can provide code bases, enabling users to construct new scenarios with minimal effort.

2.2 Large-scale Simulation

While there are many previous studies on leveraging LLMs to simulate human social behaviors (Gao et al., 2024), the number of agents in their simulators are usually very small. In such cases, the users need to sample a small set of individuals from the real-world large-scale populations, and then leverage agents to simulate the sampled individuals, assuming that these samples can accurately approximate real-world populations. However, the sampled small number of individuals may lead to very large fluctuations of the simulation results. To verify this statement, we conduct a preliminary experiment by simulating the user-item rating behaviors on a movie website. In specific, we base our experiment on the well-known dataset MovieLens-32M (Harper and Konstan, 2015), which consists of 200,948 users' 32M ratings on 87,585 movies. For each user-movie pair, we use LLMs to simulate the user's rating on the movie in the range of $\mathbf{R} = [0.5, 1.0, ..., 5.0]$. To study the fluctuation of the simulation results with different agent scales, we first sample 3.2K, 32K, 320K and 3.2M user-item pairs from the complete dataset, and then, for each case, we repeat the simulation of predicting user-item ratings for 10 times. Formally, suppose $\boldsymbol{p}_i$ represents the rating distribution of the ith experiment, where $i = 1, 2, \ldots, 10$. For each rating $r \in \mathbf{R}$, we compute the standard deviation across all experiments as

$$v(r) = \sigma(\boldsymbol{p}_1(r), \boldsymbol{p}_2(r), \ldots, \boldsymbol{p}_{10}(r))$$

where $\sigma(\cdot)$ denotes the standard deviation operation. We use the sum of the standard deviations for all possible ratings to measure the fluctuation of the simulation results. The experiment results are

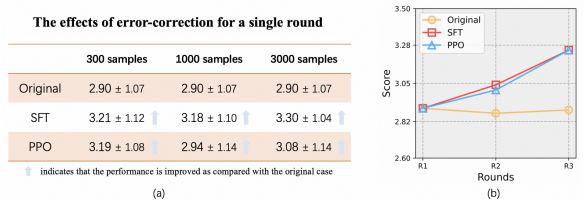

Figure 4: The effects of our error-correction mechanisms in the job market scenario for both single round (a) and multi rounds (b) settings.

presented in Figure 2(a), where we can see: as the number of samples becomes larger, the fluctuation of the simulation results is greatly lowered. This result suggests that if we only have a small number of agents, then the simulation results can be not reliable, since it can be hardly reproduced due to the large simulation fluctuation.

To solve the above problem, in our platform, we support up to one hundred thousand agents to better simulate real-world scenarios. To accelerate the simulation speed, we firstly employ an actor-based model (Hewitt et al., 1973) to facilitate automatic parallel optimization. Specifically, the "actor" function acts as an independent unit that processes computations once it receives all the necessary messages. This method ensures that each agent, representing a participant, only performs computations when the essential input messages are available, thus enabling efficient parallel optimization. Furthermore, a dynamic workflow is designed to accommodate the probabilistic outputs of LLMs, which are not pre-determinable. This dynamic approach enhances flexibility, allowing the execution path to adapt based on variations in model inference. Last but not least, our distributed framework supports multi-machine parallel simulation, overcoming the limitations of single-machine simulation scalability.

Using proposed platform, we evaluate the simulation speed in the job market and recommendation scenarios, where we run our simulator for one round for both settings[1]. All experiments in the paper use LLaMA3-8B as the foundational model for the agents. The results are presented in Figure 2(b), from which we can see: as the number of agents becomes larger, the time cost increases, and when we have 10w agents, they cost 15492 and 3024 seconds for running one round in the job market and recommendation scenarios, respectively. In addition, we also evaluate the acceleration effects of distributed parallel computing in our platform. In specific, we measure the time costs of running our platform for one round with different numbers of GPUs. The results are presented in Figure 3(a), where we can see, as the number of GPUs becomes larger, the time cost decreases, which suggests that, with the help of distributed parallel computing, our platform can effectively take the advantages of more GPUs.

2.3 Simulation Error Correction

Most previous LLM-agent-based simulation platforms lack error-correction mechanisms, which means that if unexpected results occur during the simulation process, they can be accumulated and amplified as the simulation progresses (see Figure 3(b)). To solve this problem, in our platform, we provide two strategies for correcting the simulation errors. The first one is based on LLMs, where we leverage GPT-4o to score on or revise the simulated result, utilizing the capabilities of LLMs as judges (Zheng et al., 2023). The second one is based on real humans, where we provide interfaces for the users to score or revise the simulated agent behaviors. Between these two approaches, the first is more efficient and requires no human intervention, though it may be less accurate due to the inherent biases of LLM. The second approach is more aligned with real humans but can be labor-intensive and less efficient. Suppose the simulation result is represented by a (q, a) pair, where q is the prompt for driving an agent action, and a is the action. For each of the above strategies, there are two forms of feedback provided by LLMs or real humans. Let the score for (q, a) be s, and a' be the revised results[2]. Then, we use (q, a, s) and (q, a') to fine-tune the backbone LLMs using PPO (Schulman et al., 2017) and SFT, respectively.

To evaluate the effectiveness of our designed error-correction mechanisms, we conduct experiments based on the job market scenario with LLMs as the feedback provider. To begin with, we evaluate whether PPO and SFT can improve the simulation results in a single round. In the experiments, we select different numbers of samples for labeling, and use GPT-4o to measure the reasonableness of the simulation results, assigning scores from 1 to 5 based on reasonableness level, from low to high. From Figure 4(a), we can see: for both PPO

[1]For all experiments in this paper, we used a server with a 192-core CPU, eight A100-40G GPUs, and 440 GB of memory.

[2]It should be noted that s and a' may not co-exist for the same (q, a) pair.

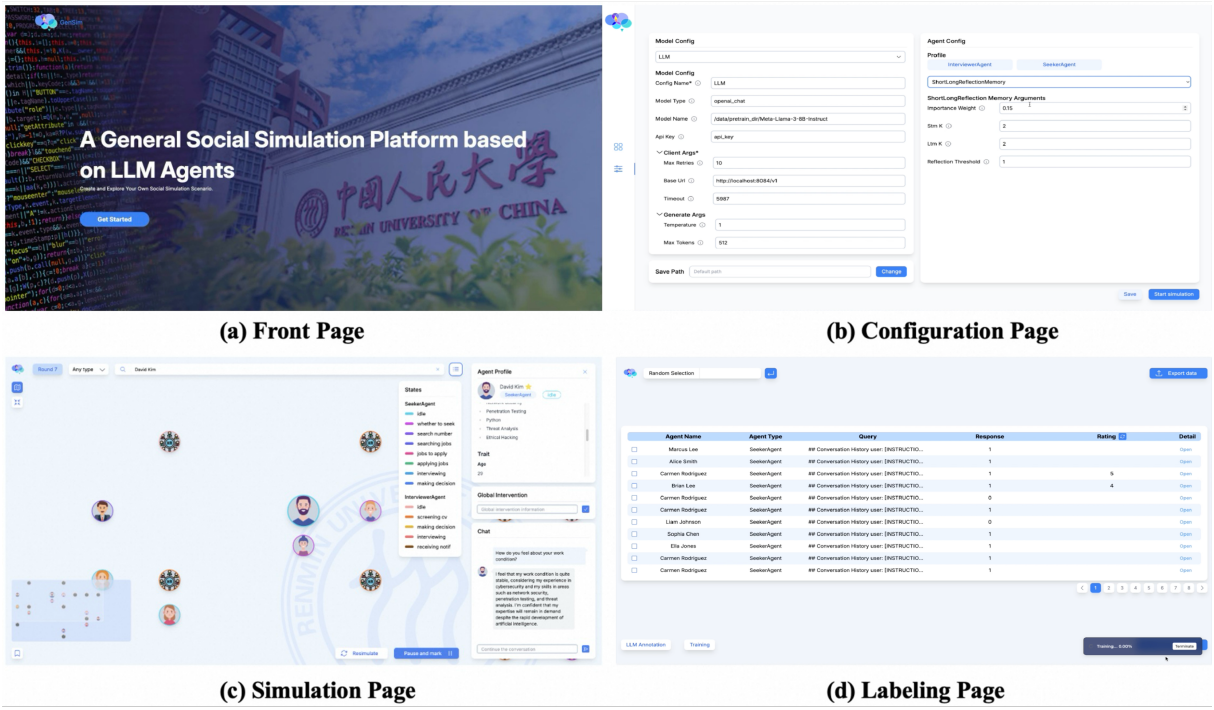

Figure 5: The interfaces of GenSim.

and SFT, they can improve the simulation performance across different numbers of labeled samples. Compared with PPO, the results of SFT are better, which is reasonable, since the revised action a' used in SFT may include more effective and comprehensive information.

Next, we evaluate the effectiveness of the error-correction mechanisms in a multi-round setting. Specifically, we fine-tune the backbone LLMs in the earlier round and use the updated models to simulate the results in the subsequent round. We present the results in Figure 4(b). We can see: if we do not conduct error-correction (the yellow line), the simulation performance is unsatisfactory. When using PPO (the blue line) or SFT (the red line), the simulation performance improves significantly, and these improvements continue to increase as the number of simulation rounds grows.

3 Usages of GenSim

In this section, we introduce the basic methods for running a default scenario. For more details, readers can refer to a short live demonstration of our platform on YouTube https://www.youtube.com/watch?v=SZf8mvhkLvI.

The complete interfaces of our platform are presented in Figure 5. In specific, the user needs to click the 'Get Started' button to initiate our platform. The user can then configure the simulation by specifying scenarios, agent profile, memory type, number of agents, LLM parameters, etc. Once configured, the platform can be launched. The simulation interface consists of three sections: At the top, there is a search box that allows the user to search for an agent to observe its status and track its behavior. In the middle, there is a display window where the behaviors of multiple agents are shown. On the right, there is a functionality window where users can view agent profiles, intervene in the system, and interact with agents. After running the platform for several rounds, the user can stop the simulation and label the results. Finally, the backbone LLMs can be fine-tuned based on the labeled results and used in the subsequent simulation.

4 Related Work

4.1 Social Science

Social science refers to the discipline that uses scientific methods to study a variety of social phenomena. It can be divided into different sub-disciplines such as political science, sociology, economics, psychology, and more. The social phenomena studied in social science are usually very sophisticated due to various uncertain variables, such as dynamically changing natural and social factors.

In the early stages, social science researchers conducted experiments in laboratory settings to

Table 1: An overview of comparing GenSim with existing works on multi-agent social simulations. The symbol '∞' indicates that our proposed platform can support general social simulation scenarios.

Name	Simulation Domain	Agent Number	Self-evolution
Generative Agent	Daily Life	25	✗
RecAgent	Recommendation Systems	20	✗
EconAgent	Economic Market	100	✗
Social Simulacra	Social Network	1,000	✗
Agent Hospital	Healthcare	<100	✗
WarAgent	Warfare	<100	✗
GenSim (ours)	∞	**>100,000**	✓

conveniently control variables. For instance, (Gosnell, 1926) studied the effect of non-partisan mail campaigns on increasing voter turnout by interviewing 6,000 individuals. To enhance the reality and accuracy, *Field Experiment* and *Randomized Controlled Trial (RCT)* have gradually become the primary methods in social science research. For example, (Staiano et al., 2014) conducted a field experiment involving 60 participants over a period of six weeks, collecting various personal identity information. This helped in studying the monetary valuation of mobile personally identifiable information. (Bolier et al., 2013) designed 39 randomized controlled studies involving 6,139 participants, aiming to evaluate the impact of positive psychology interventions on both the general public and individuals facing specific psychosocial issues.

Although existing social science experimental methods have been widely adopted, there are several significant drawbacks such as irreversibility, high cost, poor controllability, and even potential ethical violations. These issues severely restrict the feasibility and scalability of social experiments. Moreover, long experimental durations and complex environmental factors may adversely impact the validity and timeliness of the experimental conclusions.

4.2 Multi-Agent Simulation

With the rise of large language models (Achiam et al., 2023; Team et al., 2023), the "AI for Social Sciences" direction presents unprecedented opportunities for researchers in both social sciences and artificial intelligence. The core idea is to utilize LLMs as human-like brains to create behavior decision-makers that can approximately imitate real individuals, known as LLM-based agents.

Recently, there are increasing amount of works focused on how to leverage LLM-based multi-agents to conduct social simulation experiments in various fields. Specifically, Generative Agent (Park et al., 2023), as a pioneering work in this field, simulates the daily life of 25 agents in a small town. Similarly, RecAgent (Wang et al., 2023) and Social Simulacra (Park et al., 2022) investigate social phenomena that may emerge from group behavior in recommendation systems and social networks, such as information cocoons and the spread of antisocial behavior. In addition, EconAgent (Li et al., 2024b), Agent Hospital (Li et al., 2024a), and WarAgent (Hua et al., 2023) each use multi-agent simulation in the fields of economic markets, healthcare, and warfare respectively, to model specific human behaviors in their specific domains, and further analyze individual actions and group social phenomena.

While the aforementioned methods demonstrate the powerful advantages of LLM-based agents in social simulation, there are still three significant shortcomings hindering the further development: existing works cannot support different domains and large-scale social simulations, and lack a error correction mechanism. In contrast, our framework effectively overcomes these issues. We can support general-domain simulations and large-scale multi-agents, and introduce a novel self-correcting self-evolution mechanism, achieving more reliable and accurate simulation performance. Overall comparison results are summarized in Table 1.

5 Conclusions

In this paper, we introduce a general, large-scale, and correctable social simulation platform based on LLM agents. This is the initial version of our platform, we believe there is still much room left for improvement. In the future, we plan to incorporate

more advanced simulation accelerating strategies, and develop more adaptive self-correction mechanisms to improve the simulation performance.

Limitation

Despite our pioneering effort in developing a general social simulation platform using LLM-based agents, there are still several limitations. **First**, the platform's generalization capability across diverse sociocultural contexts has not been fully verified. While agent simulations are more cost-effective than real-world experiments, their accuracy in modeling complex cultural norms, regional socioeconomic dynamics, and varied institutional settings requires further validation through comparative studies with traditional experimental approaches. **Second**, our reliance on LLM-as-a-judge for synthetic data evaluation introduces potential assessment biases. The prohibitive cost of human annotation prevented systematic verification of alignment between LLM-generated scores and expert human ratings, particularly in culturally sensitive scenarios where language models may encode latent value misalignments. This necessitates future research on calibration techniques for automated social evaluation metrics. **Third**, the self-correction mechanisms of LLMs in social simulations present unresolved reliability concerns. The inherent biases in LLMs' error identification and correction processes across varied social scenarios (especially those involving ethical dilemmas or conflicting group interests) remain unexamined. We leave these limitations as our future work to explore.

Ethical Considerations

The data used in this paper comes from public datasets or simulated experiments using synthetic data generated by Large Language Models (LLMs). For public datasets, we adhere to their usage licenses to ensure that our work does not present any ethical issues.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (No. 62422215 and No. 62472427), Beijing Outstanding Young Scientist Program NO.BJJWZYJH012019100020098, Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "DoubleFirst Class" Initiative, Renmin University of China, Public Computing Cloud, Renmin University of China, fund for building world-class universities (disciplines) of Renmin University of China, Intelligent Social Governance Platform.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Linda Bolier, Merel Haverman, Gerben J Westerhof, Heleen Riper, Filip Smit, and Ernst Bohlmeijer. 2013. Positive psychology interventions: a meta-analysis of randomized controlled studies. *BMC public health*, 13:1–20.

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.

Harold F. Gosnell. 1926. An experiment in the stimulation of voting. *American Political Science Review*, 20(4):869–874.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Carl Hewitt, Peter Bishop, and Richard Steiger. 1973. A universal modular actor formalism for artificial intelligence. In *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, IJCAI'73, page 235–245, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024a. 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. 2024b. 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.

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*, pages 1–18.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Jacopo Staiano, Nuria Oliver, Bruno Lepri, Rodrigo De Oliveira, Michele Caraviello, and Nicu Sebe. 2014. Money walks: a human-centric study on the economics of personal mobile data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 583–594.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, et al. 2023. User behavior simulation with large language model based agents. *arXiv preprint arXiv:2306.02552*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

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