SoAgent: A Real-world Data Empowered Agent Pool to Facilitate LLM-Driven Generative Social Simulation

Na Ta
Renmin University of China
Beijing, China
tanayun@ruc.edu.cn

Kaiyu Li
Wilfrid Laurier University
Waterloo, Canada
kli@wlu.ca

Yushu Zhou
Renmin University of China
Beijing, China

Yuhan Liu
Renmin University of China
Beijing, China
yuhan.liu@ruc.edu.cn

ABSTRACT

The advancement of Large Language Models (LLMs) has significantly transformed research methodologies for social sciences. Beyond the much enhanced textual level analysis, the ability of LLMs to simulate social behavior offers new opportunities for studying social mechanisms at both individual and collective level, which are traditionally resource-intensive or even ethically challenging. Therefore, generative agent-based models (GABMs) are introduced to address these limitations. However, existing agents are typically initialized with synthetic or online data, limiting their alignment with the real-world dynamics. To overcome this gap, we propose SoAgent (Social Simulation Agents), a novel framework for generating agents based on real census-like data. Based on data-driven analysis from real census samples (nation-level, half-decade, approximately 4,500 samples/year) and carefully crafted prompt engineering, SoAgent is designed specifically for broader social science applications beyond social media analysis, with more real-person-like features. Our experiments demonstrate that SoAgent significantly outperforms synthetic-profile agents in tasks such as questionnaire response simulation and the study of conspiracy theory diffusion, offering a more realistic and robust foundation for social simulation research. We also propose our vision about future work on this promising research direction.

VLDB Workshop Reference Format:
Na Ta, Kaiyu Li, Yushu Zhou, and Yuhan Liu. SoAgent: A Real-world Data Empowered Agent Pool to Facilitate LLM-Driven Generative Social Simulation. VLDB 2025 Workshop: The 2nd International Workshop on Data-driven AI (DATAI).

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://osf.io/nwcmu/?view_only=14b04a52f28141dca07de16bd354e9e4

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
* Na Ta is the corresponding author.

1 INTRODUCTION

The past three years have witnessed a great leap of the research methodology of social science studies enabled by Large Language Models (LLMs) [25]. As an example, in the field of content analysis, LLMs can effectively assist on traditional natural language processing tasks such as thematic and sentiment analysis, with user-friendly prompts, instead of complicated programming codes [21]. More importantly, social science studies often care about the interactions between individuals [24], the collective dynamics and trends of certain groups of people [22], and the driving mechanisms behind [4], which can potentially benefit from the ever-growing capability of LLMs to a large extent.

Existing Methods and Their Limitations. Traditionally, social science studies are accomplished by qualitative methods (e.g, in-depth in person interview lead by researchers themselves), or quantitative methods (e.g, surveys or questionnaires combined with statistical analysis), requiring a considerable number of human participants and related resources such as temporal and physical costs, which can be quite expensive [6], and sometimes infeasible due to practical limits or ethical concerns. To overcome these challenges, with the enhanced role-playing capabilities of LLMs, a growing number of scholars have incorporated LLMs into social science research, achieving promising results in scenarios such as simulated social surveys and effect evaluation of communications [2].

Specifically, researchers have established generative agent based models (GABM) and frameworks [16, 18, 20] to study the dynamics of social events and processes using simulated “worlds" to accommodate a number of agents so that the agents could act like individuals in the real world, such as presenting certain personalities, planning, reflecting and behaving, as well as establishing social relationships and interacting with each other, and even evolve as a group. Existing studies have proven the validity of the "silicon" samples [1], or general-purpose computational agents [19] in representations of human attitudes and ideas. However, most current methods tend to employ synthetic user profiles with basic attributes, or user persona derived from online data, which may lead to low or limited human representation. A few studies have tried to derive persona from from surveys of real users for certain domains with finer granularity [9, 29], yet they have not connected to the downstream tasks to apply and test those persona in the context of social simulations, nor expand the survey data source to a broader sample space to fit social science studies.

Our Proposal: SoAgent. We propose a novel real-world data-based agent pool, **SoAgent** (**So**cial simulation **Agent**s), to address the aforementioned limitations. The differences (therefore, advancements) from existing methods are as follows.

First, the agents in SoAgent are injected with multi-faceted features of continuous yearly large-scale nation-wide real samples of general social survey (see CGSS [23] in Section 2). The distribution of samples effectively replicate the real world, and the quality of the sample data are better guaranteed than those collected from social media, which is not comparable by synthetic data.

Second, based on multiple features learned and injected from finer granularity of real-world individuals, agents in SoAgent can simulate a broader range of social scenarios, beyond social media study. The comprehensive questionnaires are designed to encompass hundreds of questions including personal beliefs and preferences, social life, government policy matters, and others. Therefore, each respondent provides a much enriched and precise portrait of a real person for the agent to embody.

Third, with carefully designed prompts to utilize the capability of LLMs, SoAgent provides the ability to accomplish social simulation tasks at both individual level and collective level, to better address the needs of social science research. On the one hand, application domains such as marketing or consumer study focus on motivations and behaviors of individuals; on the other hand, theoretically, social scientists often ask the question of why people, as a group, would collectively present certain preferences or trends. SoAgent has been tested on both levels with satisfying performance, demonstrating it a promising agent pool for diversified social simulations.

2 ARCHITECTURE OF SOAGENT

2.1 Workflow

The SoAgent framework (Figure 1) initiates its workflow with a User Pool, which forms the foundation for modeling complex social dynamics. This user pool is divided into two distinct categories: synthetic users and real-world users. Synthetic users are generated with basic features whose distributions are kept in line with real world population to simulate, such as age, gender, and simulated social relationships, enabling controlled experimentation with customizable parameters. In contrast, real-world users are incorporated with authentic features, including opinion attitudes, polarization metrics, and actual social relationships, derived from empirical data source such as CGSS. The inclusion of synthetic users is to support quick validation of social simulation in case real data source are not applicable.

Given a specific user pool, the Agent Engine then construct an agent pool. The data preprocessing phase involves cleaning and structuring these features to ensure consistency and reliability, followed by feature selection to identify the most relevant attributes for simulation, thereby injecting those attributes to construct a number of agents. Further simulation can provide sampling rules to attract proper agents from the agent engine.

The Social Simulation Engine utilizes the selected features to model both individual and collective behaviors. The engine drives this process by simulating social phenomena, utilizing either synthetic or real-world features to capture the nuances of human interactions. Through incorporation of various simulation scenarios, the engine enables researchers to explore diverse areas and validate models against observed social patterns. In both Agent Engine and Social Simulation Engine, SoAgent actively communicate with LLMs to employ their role-play capabilities in simulating social interactions and dynamics.

Figure 2 provides a more detailed illustration of how SoAgent works, with the given simulation task of exploring how a conspiracy theory spread between people, which can be modeled either using synthetic agents with basic features such as Big Five psychological combinations, or authentic agents whose labels are derived from social media, nation-wide survey,etc.

2.2 Offline Agent Pool Preparation

2.2.1 Real-world Data Source. We can plug in outside data sources into SoAgent. To demonstrate, we use the Chinese General Social Survey (CGSS) [23] dataset to generate our agent pool. This dataset is derived from China's first large-scale, nationwide, comprehensive, and continuous social survey project [3]. Data collections are conducted online and offline through an online platform using both paid and unpaid methods. The unpaid portion was distributed and promoted by the survey organizers via social media platforms such as WeChat and Weibo, allowing voluntary participation by internet users. Invalid responses were filtered out in real-time during the survey process. Additional invalid entries were later removed following manual verification by two graduate-level reviewers. To protect participant privacy, personally identifiable information such as WeChat IDs, nicknames, and user comments was excluded from the final dataset. The yearly sample sizes are around 5,000, portraying people living in China, with their attitude to a couple of questions that are relevant to social events and governmental policy, as well as their sociodemographic features. According to the size of the comprehensive questionnaire, each respondent contributes at least 100 variables.

2.2.2 Data Preprocessing. We collected public CGSS results of the latest 5 years, and removed irrelevant columns (e.g., row sequence, devices and operating systems used to fill the survey), to form the initial dataset used for following steps (Table 1) . We then removed open-ended questions, since the variations of responses to such questions can deviate more drastically from the variations of responses to non-open-ended questions.

Table 1: Statistics of the CGSS Dataset

Year	# Questions	# Sub-questions†	# Samples	# Clusters
2018	47	185	5,415	10
2019	33	127	4,882	10
2020	34	91	5,000	10
2021	46	132	3,619	10
2022	60	199	3,788	10

† In the CGSS questionnaire, a number of questions are matrix questions, i.e., multiple sub-questions concerning different aspects of the main question.

2.2.3 Feature Selection. Generally, each question or sub-question in the survey can be treated as an attribute of a person, so that different combinations of responses to all questions can represent

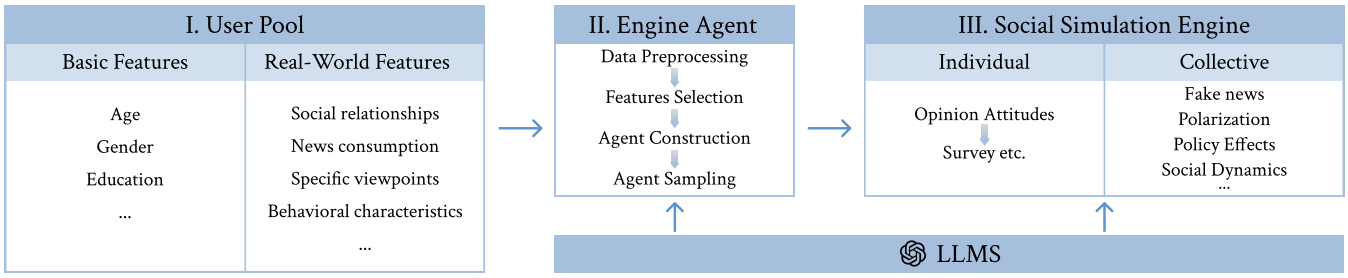

Figure 1: SoAgent Framework

different human samples. However, it is not optimal to attach all attributes to an agent when using survey samples to simulate a real person. On the one hand, it can be expensive for LLMs to instantiate an agent with more than 100 attributes derived from corresponding survey questions, because a considerable number of tokens must be used. On the other hand, verbosity might increase since a group of questions within one survey may be of the same topic, resulting in duplicated attributes attached to one agent. For example, "*How many years have you spent on education*" and "*What is your education level*", are semantically similar. Therefore, we performed a clustering of survey questions, so that each agent can be injected with one representative attribute from each cluster, such as social demographic features, or consumption behaviors. A comparison between traditional clustering methods such as DBScan [5] and LLM-based clustering strategies suggested that the latter works better in consistency among survey years and clear semantics assigned to each cluster. Therefore, we stick to the LLM-based clustering in following steps.

2.2.4 Agent Pool Construction. For a given survey (such as the CGSS survey at 2021), we first load the survey questions and clusters, and the corresponding sample data, where each sample record (p_{record}) represents a real respondent. Then we generate one agent **a** for each p_{record} as follows. For each survey question cluster, assign p_{record}'s responses to all questions within this cluster to **a**. After all p_{record} records are processed, we construct an agent pool of the same size to the survey samples, with each agent having representative features attached. By doing this, the distributions of samples are directly replicated into the constructed agents.

2.3 Online Simulation

2.3.1 Social Simulation Setup. The SoAgent's social simulation setup begins with defining the simulation task, which can focus on either individual-level predictions, such as personal opinion, or collective-level outcomes, such as group polarization or consensus formation. This step establishes the scope and objectives, ensuring that the simulation aligns with the research question. Next, agent screening rules are defined to select appropriate agents from the agent pool, which corresponds to both synthetic users with basic features (e.g., age, gender) and real-world users with authentic attributes (e.g., social relationships, opinion attitudes) from the user pool. These rules filter agents based on criteria relevant to the task, such as demographic characteristics or behavioral traits, to ensure the simulation reflects the desired population dynamics.

Once agents are selected, the simulation process advances to agent interaction assisted by LLMs. For collective tasks, agents engage in simulated interactions, mimicking real-world communication patterns to model social phenomena like information diffusion or conflict emergence. The Social Simulation Engine then computes the outcome variable, such as an individual's attitude shift or a group's polarization level, and verifies it against predefined benchmarks or empirical data to ensure accuracy. Finally, the framework supports exploratory tasks by allowing researchers to extend simulations, adjusting parameters or introducing new variables to investigate alternative scenarios or uncover emergent behaviors. This structured yet flexible setup enables comprehensive analysis of social systems within a controlled computational environment.

2.3.2 Sampling. Given a requirement to simulate a social science experiment with n agents using SoAgent, and assuming we already have a pool of N user profiles—each corresponding to an existing agent—the key question becomes: which subset of agents should be selected to ensure meaningful and unbiased simulation results. Intuitively, we aim to maximize coverage across diverse attribute values (e.g., demographics, education, age groups) to reduce bias and enhance the representativeness of the simulation. Prior studies [27] suggest that a balanced feature selection strategy leads to fair and more reliable outcomes in social simulations. For instance, if we focus on two attributes—educational level and age group—we would ideally select agents representing all education levels (graduate, undergraduate, secondary school, and below) as well as all major age brackets (teen, adult, senior, and elderly). This ensures that the simulation captures the heterogeneity of real-world populations and supports generalizable conclusions. Therefore, our goal is to select n individuals from the N available profiles such that the selection covers as many distinct attribute values as possible. This problem corresponds to the Maximum Set Coverage problem [7], which is known to be NP-hard. To address this, we adopt a greedy approach used in [13] that iteratively selects the row (individual) contributing the largest number of previously uncovered attribute values, continuing until n individuals have been selected from the user pool.

2.3.3 Prompt Engineering. We present crucial prompts used in this work. For the prompts of simulating agents' interactions, and single agent's reasoning on opinions, please refer to our pervious work [15].

(1) Prompts to cluster the CGSS questionnaire.

`System prompt``: You are a professional survey analyst who is good at analyzing and clustering survey questions.`

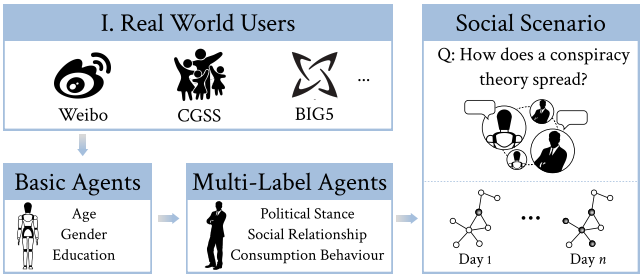

Figure 2: SoAgent Demonstration Example

```
Please output the results strictly in the json format required by the user and do not include any additional explanatory text.
Task: Please analyze the following questionnaire file and perform intelligent cluster analysis on the questions. Each cluster needs to be able to serve as a certain characteristic of a person. There must be at least ten categories, and each category corresponds to a category id.
Input: a CGSS yearly questionnaire.
Output: a json file of clusters of questions.
```

(2) Prompts for agents to respond CGSS questionnaire.

```
System prompt: Assume you a real person with following responses to given questions, traits: sample ID: 16 | Q2: your age? 25-29 years old | Q3: overall, are you interested in political news? quite interested | ...
Task: Stick to your personal characteristics, provide your answers to following questions:
Input: a group of CGSS questions to answer, and a sample response, e.g., Q37: did Covid-19 affected your life? Options are: 1: significant disruption, 2: considerable impact, 3: moderate impact, 4: minimal impact, 5: basically unaffected, 6: completely unaffected. ...
Output: Your responses will be formatted in json. For example: "Q22":"3":"middle", "Q15":"5":"not at all".
```

3 PRELIMINARY RESULTS

3.1 Experimental Setting

Datasets. We used the CGSS dataset (see section 2.2) to generate the testing agent pool, consisting 22,704 real-world respondent samples in total.

Social Science Tasks. We categorize the social science tasks into two types, the individual-level tasks, where the respondent are typically required to complete a task independently and their responses are evaluated or analyzed after the collection, and the community-level tasks, where individuals in a community interact with others thru specific media (e.g., social media or face-to-face), and researchers observe their attitudes and/or behaviors and changes during the experiment.

In this paper, we hope to simulate human behavior using LLM-powered agents. We consider a classical individual task, filling out a survey questionnaire. The LLM-powered agents are required to answer a set of questions and their answers are collected to compare with human responses. We consider one community-level tasks, the spreading of a conspiracy theory. We use agents to simulate individuals, and make the agents chat with each other everyday. We simulate the communication of 14 days in our experiments. And observe if the hypothesis and theories tested in the social science field still work for LLM-powered simulation.

Evaluation Metrics. For individual questionnaire task, we use the Root Mean Squared Error (RMSE) to estimate the accuracy of how LLM simulate human's attitude to a question. Our exploratory test focus on categorical values. If the LLM give the correct answer that is consistent with the ground truth, a value of 1 will be assigned. Otherwise, it will receive 0 credit. For the community level task to explore how agents communicate and spread a conspiracy theory, we evaluate how many individuals are affected by the given conspiracy theory, what features are significant in predicting the contagion, and compare these results with established results from existing studies.

Baselines. Our main claim is that simulating social behavior by randomly generate basic persona is not sufficient and cannot reflect the real world well. Thus we mainly compare two methods. The first method randomly generates human personality using the "Big Five Personality Traits (**Big5**) [26]" adopted by a number of social science studies to represent a single person's psychological traits, i.e., Openness to experience, Conscientiousness, Extroversion, Agreeableness, and Neuroticism [15]. The second methods, i.e., **SoAgent**, generates the selected most representative features and attitudes data from the collected real-world questionnaires data, and generate the prompt using these information, and feed the prompt of the agents from candidate pool.

Implementation Settings. Experiments run on a MacBook Pro with 3.3 GHz Intel Core i5 CPU, 16 GB RAM. Code is implemented in Python 3.8. For the LLM model, We selected the gpt-3.5-turbo-1106 model of OpenAI [17] to generate the agents in our experiments and observe the differences.

3.2 Experimental Results

3.2.1 Individual Level Task. We compare **SoAgent** with **Big5** on their performance on simulating real-world responses to CGSS yearly surveys. Table 2 presents the results on 22,704 agents for each method. It can be seen that SoAgent outperforms the baseline methods.

Table 2: Simulating Individual Tasks

Year	**SoAgent** Categorical RMSE	**Big5** Categorical RMSE
2018	0.99	1.47
2019	0.94	1.30
2020	0.83	1.37
2021	1.05	1.51
2022	0.92	1.49

3.2.2 Community Level Task. We compare **SoAgent** with **Big5** on their performance on simulating real-world dynamics of the spreading of a conspiracy theory. We used 5 conspiracy theories: (1) *911*, "the 911 event was conducted by the USA government";

Table 3: Simulating Collective Tasks

Conspiracy Theory	**SoAgent** Significant Predictors	**Big5** Significant Predictors
911	Age, Edu, P.S.*	Age, Edu
Bermuda	Edu	None
Climate Change	Edu, P.S.	Edu
Covid-19	Age, Edu, P.S.	Edu, P.S.
Moon	Edu, P.S.	None

* P.S. is short for political stance.

(2) *Bermuda*, "a region of the Atlantic Ocean—contains a hidden force or anomaly that causes ships and planes to disappear mysteriously"; (3) *Climate Change*, "scientific consensus on human-caused climate change is a deliberate hoax, and climate data is falsified to advance hidden agendas, such as political control, economic gain, or global governance"; (4) *Covid-19*, "covid originated from a Wuhan laboratory"; (5) *Moon*, "the Moon landing was a complete hoax orchestrated by NASA to assert American dominance during the Space Race". According to previous studies, the belief in a conspiracy theory is significantly related to one's age, education level and political stance [8, 10]. Since the agents in SoAgent are injected with such features, to make a fair comparison, we also attach these features to Big5 agents. Then we tested the significant predictors the two methods use to assign an infected status to an agent when this agent is exposed to a conspiracy theory. Table 3 presents the results on 90 agents for each test run on a 14-day communication time span, where each agent randomly exchange its opinion on a specific conspiracy theory with 3 other agents on a daily basis. Again, SoAgent outperforms the baseline method.

3.2.3 Exploratory Case Study. After the validations of SoAgent's capability to simulate social process at both individual and community levels, we further simulated a 14-day time window of 90 agents to study the temporal dynamics of the spread of a conspiracy theory, where each agent randomly exchange its opinion on a specific conspiracy theory with 3 other agents on a daily basis. Since this is a preliminary conceptual study, for the sake of API-calling expenses, we use simplified agents, which only have attributes of age, education, political stance, nationality set before the experiment. For the age attribute, possible values are *young*, *middle age* and *elderly*; for the education attribute, possible values are *bachelor degree and above*, *middle school* and *preliminary school and below*; for the political stance attribute, possible values are *extreme Democratic*, *leaning Democratic*, *neutral*, *leaning Republican* and *extreme Republican*. All agents are set to have the United States citizenship. Figures 3-7 presents the results, where 'susceptible' means that an agent does not believe in a conspiracy theory but could believe in the future, 'infected' means that an agent believes in a conspiracy theory, and 'recovered' means that an agent used to believe in a conspiracy theory but does not believe in it currently.

Our initial observations are: (1) The topic (or domain) of the conspiracy theory could affect the adoption of a conspiracy theory. Specifically, topics of high political overtones are more likely to be adopted by agents, such as the '911' conspiracy, with the 'Bermuda' conspiracy as a counter example, which is of low political tendency. (2) Agents of different ideology presents different levels to recover

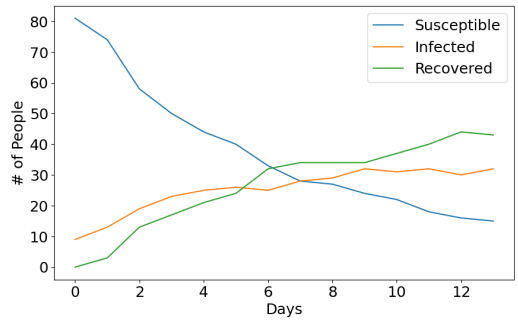

Figure 3: Spread of '911' Conspiracy (90 agents, 14days)

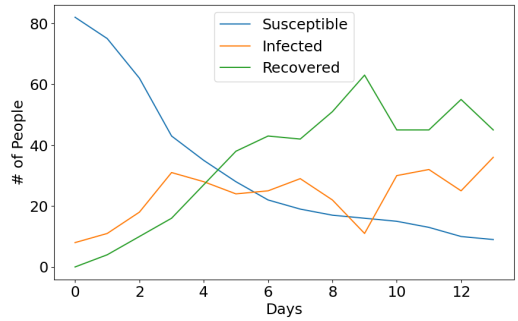

Figure 4: Spread of 'Bermuda' Conspiracy (90 agents, 14days)

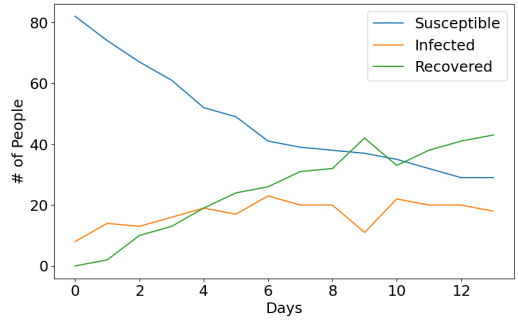

Figure 5: Spread of 'Climate' Conspiracy (90 agents, 14days)

from a conspiracy theory. Specifically, agents with the Democratic Party background are more open to be persuaded not to believe in a given conspiracy theory. (3) The impact of an agent's opinion climate (in our experiment, 3 other agents' opinions) on its adoption/rejection of a conspiracy theory is open to further validation. While these agent-based findings remain preliminary, they offer valuable pathways for subsequent verification.

4 CONCLUSION AND FUTURE WORK

In this work, we present SoAgent, a novel framework for generating realistic social simulation agents using real survey data and carefully engineered prompts. Our approach addresses a critical limitation of existing generative agent-based models (GABMs), which often rely on synthetic or online data, leading to misalignment with real-world social dynamics. By leveraging nation-level, longitudinal census data and advanced prompt engineering, SoAgent

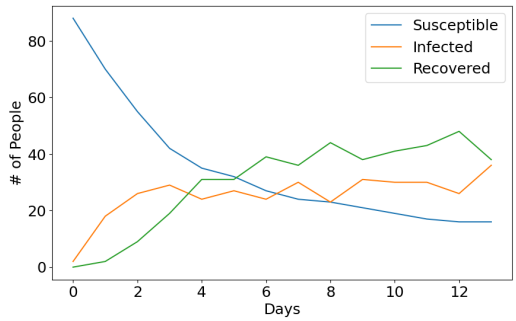

Figure 6: Spread of 'Covid-19' Conspiracy (90 agents, 14days)

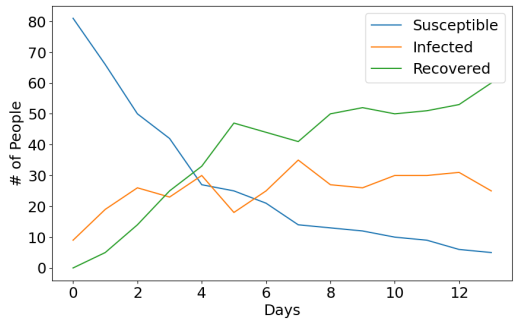

Figure 7: Spread of 'Moon' Conspiracy (90 agents, 14days)

enables more accurate simulations of human behavior in tasks such as questionnaire response generation and conspiracy theory diffusion analysis. Our long-term goal is to develop a benchmark comprising datasets that simulate a variety of social science tasks, along with standardized evaluation metrics to assess simulation performance. These simulations will be validated against real-world experimental results serving as ground truth. We believe this effort will be valuable for both social science applications and broader research communities. Besides, how to effectively input relational tables—especially questionnaire data which includes various types of user profiles and attitudes—into LLMs remains an open challenge. We will also explore hybrid interactions that combine human and agent communication within social activities [11, 12, 14, 28]. Finally, fine-tuning pre-trained LLMs to better align with the specific needs of social science simulations represents a promising direction for future work.

REFERENCES

[1] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[2] Christopher A Bail. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences* 121, 21 (2024), e2314021121.

[3] Chinese National Survey Data Archive. n.d.. Chinese National Survey Data Archive. http://www.cnsda.org/. Accessed: 2025-05-22.

[4] Steven N Durlauf and H Peyton Young. 2001. *Social dynamics*. Vol. 4. Mit Press.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, Vol. 96. 226–231.

[6] Robert M Groves. 2005. *Survey errors and survey costs*. John Wiley & Sons.

[7] Juris Hartmanis. 1982. Computers and intractability: a guide to the theory of np-completeness (michael r. garey and david s. johnson). *Siam Review* 24, 1 (1982), 90.

[8] Roland Imhoff, Felix Zimmer, Olivier Klein, João HC António, Maria Babinska, Adrian Bangerter, Michal Bilewicz, Nebojša Blanuša, Kosta Bovan, Rumena Bužarovska, et al. 2022. Conspiracy mentality and political orientation across 26 countries. *Nature human behaviour* 6, 3 (2022), 392–403.

[9] Soon-Gyo Jung, Joni Salminen, Kholoud Khalil Aldous, and Bernard J Jansen. 2025. PersonaCraft: Leveraging language models for data-driven persona development. *International Journal of Human-Computer Studies* 197 (2025), 103445.

[10] Jonas R Kunst, Aleksander B Gundersen, Izabela Krysińska, Jan Piasecki, Tomi Wójtowicz, Rafal Rygula, Sander van der Linden, and Mikolaj Morzy. 2024. Leveraging artificial intelligence to identify the psychological factors associated with conspiracy theory beliefs online. *Nature Communications* 15, 1 (2024), 7497.

[11] Kaiyu Li, Guoliang Li, Yong Wang, Yan Huang, Zitao Liu, and Zhongqin Wu. 2021. CrowdRL: An end-to-end reinforcement learning framework for data labelling. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 289–300.

[12] Kaiyu Li, Xiaohang Zhang, and Guoliang Li. 2018. A rating-ranking method for crowdsourced top-k computation. In *Proceedings of the 2018 International Conference on Management of Data*. 975–990.

[13] Kaiyu Li, Yong Zhang, Guoliang Li, Wenbo Tao, and Ying Yan. 2019. Bounded Approximate Query Processing. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (2019), 2262–2276. https://doi.org/10.1109/TKDE.2018.2877362

[14] Cong Lin, Yuxin Gao, Na Ta, Kaiyu Li, and Hongyao Fu. 2023. Trapped in the search box: An examination of algorithmic bias in search engine autocomplete predictions. *Telematics and Informatics* 85 (2023), 102068.

[15] Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 7886–7894. https://www.ijcai.org/proceedings/2024/873

[16] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. 2024. A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences* 121, 9 (2024), e2313925121.

[17] OpenAI. 2023. ChatGPT: GPT-4 Technical Preview. https://openai.com/chatgpt. Accessed: 2025-05-22.

[18] Joon Sung Park, Joseph C. 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, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (Eds.). ACM, 2:1–2:22. https://doi.org/10.1145/3586183.3606763

[19] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie J. Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. *CoRR* abs/2411.10109 (2024). https://doi.org/10.48550/ARXIV.2411.10109 arXiv:2411.10109

[20] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. *arXiv preprint arXiv:2502.08691* (2025).

[21] Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. 2025. Thematic-LM: A LLM-based Multi-agent System for Large-scale Thematic Analysis. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, 649–658. https://doi.org/10.1145/3696410.3714595

[22] Thomas C Schelling. 1971. Dynamic models of segregation. *Journal of mathematical sociology* 1, 2 (1971), 143–186.

[23] Chinese General Social Survey. n.d.. http://cgss.ruc.edu.cn/English/Home.htm. [Accessed 01-05-2025].

[24] Na Ta, Kaiyu Li, Yi Yang, Fang Jiao, Zheng Tang, and Guoliang Li. 2020. Evaluating public anxiety for topic-based communities in social networks. *IEEE Transactions on Knowledge and Data Engineering* 34, 3 (2020), 1191–1205.

[25] Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining* 15, 1 (2025), 1–30.

[26] Hui Wang, Yuxia Liu, Zhanying Wang, and Ting Wang. 2023. The influences of the Big Five personality traits on academic achievements: Chain mediating effect based on major identity and self-efficacy. *Frontiers in psychology* 14 (2023), 1065554.

[27] Xiaoying Xing, Hongfu Liu, Chen Chen, and Jundong Li. 2021. Fairness-aware unsupervised feature selection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3548–3552.

[28] Yi Yang, Na Ta, Kaiyu Li, Fang Jiao, Baijing Hu, and Zhanghao Li. 2021. Influential factors on collective anxiety of online topic-based communities. *Frontiers in Psychology* 12 (2021), 740065.

[29] Hye Sun Yun, Mehdi Arjmand, Phillip Sherlock, Michael K Paasche-Orlow, James W Griffith, and Timothy Bickmore. 2023. Keeping Users Engaged During Repeated Administration of the Same Questionnaire: Using Large Language Models to Reliably Diversify Questions. *arXiv preprint arXiv:2311.12707* (2023).