SSiLU: A Generalized Social Simulation Framework Empowered by LLM Agents and A Pool of 10 Million Real-World Users

Anonymous ACL submission

Abstract

Massive social simulation plays a vital role in predicting real-world trends. Previous studies use Large Language Models (LLMs) to replace traditional methods to enrich scenarios and improve simulation accuracy. However, they are faced with limitations such as rigid frameworks, small-scale simulations, and narrow evaluation criteria. To this end, we introduce ***SSiLU***, a generalized **S**ocial **Si**mulation framework powered by **LLM** agents and a pool of 10 million real-world **U**sers. Our framework features a large-scale user pool, a demographic distribution sampling strategy, and a unified simulation evaluation method. We evaluate its effectiveness by conducting large-scale simulations across political, journalistic, and economic scenarios. The results demonstrate that our framework enables social simulations that reflect large-scale population dynamics, ensuring diversity, trustworthiness, and representativeness with a standardized pipeline and minimal modifications.

1 Introduction

Massive social simulation aims to simulate social events at a large population scale, which has been of vital importance in forecasting potential real-world trends and capturing specific groups' preferences on particular topics or special events (Hoey et al., 2018; Murić et al., 2022; Mou et al., 2024a). Previous works also demonstrated that modeling massive social simulations by means of mathematical or statistical methods can significantly improve the efficiency and accuracy of traditional political and sociological analysis paradigms (Gao et al., 2022; Mou et al., 2024c).

The traditional and mainstream method for social simulation is agent-based modeling (ABM) (Schelling, 1969; Macal and North, 2009; Jusup et al., 2022; Chuang and Rogers, 2023), which employs heuristic-like rules or mathematical functions to simulate the actions of individuals (Tang, 2024), and then scales up these actions to forecast the collective result. With the rise of agent-based simulations powered by

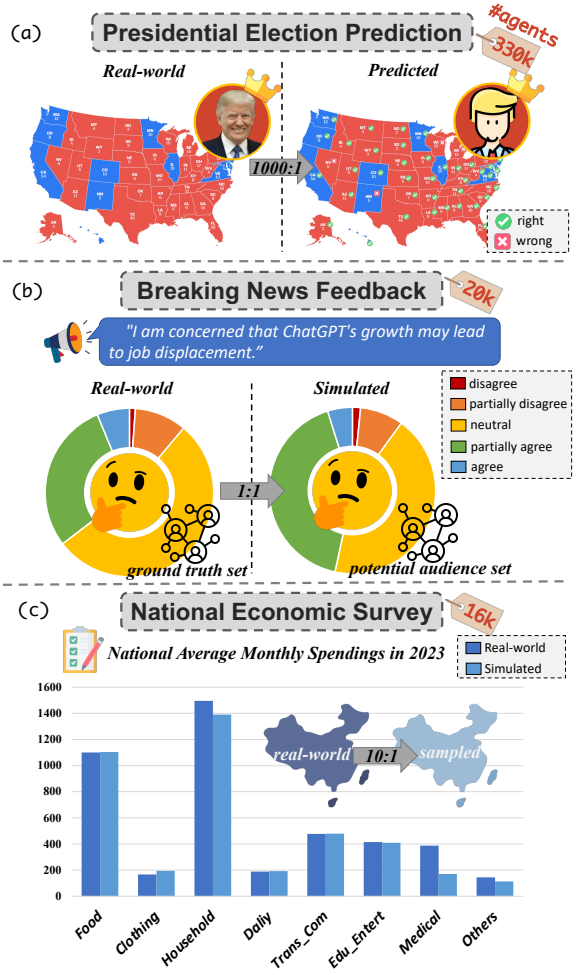

Figure 1: An illustration of the simulation results following the ***SSiLU*** framework in (a) presidential election prediction, (b) breaking news feedback, and (c) national economic survey scenarios. Different models are selected to demonstrate the broad applicability.

*Corresponding authors.

Large Language Models (LLMs), researchers have carried out social simulations in diverse scenarios and with different granularities (Shao et al., 2023; Mou et al., 2024b; Liu et al., 2024; Qi et al., 2024). However, despite LLMs' powerful role-playing abilities, existing studies struggle to address the following challenges.

Q1. How to construct a massive social simulation framework with high flexibility and customization? Current works mainly focus on constructing highly customized single scenarios like programming, legal, and medical tasks, which heavily depend on expert knowledge and contain a lot of handcraft design (Lee et al., 2023; Argyle et al., 2023). It is quite costly to build up wheels repeatedly and a paradigm that is able to guide any massive social simulation pipeline in a standard way can be of great help.

Q2. How to satisfy the large-scale population aligned with the real-world distribution? Accurate social simulation requires that the simulated individuals represent the diversity and aligned distribution of real-world populations, especially when the population is large. While random sampling can capture this diversity, it falls short when aligning to the demographic distribution of the real world and is prone to source-driven biases (Giorgi et al., 2022; Vraga, 2016; Cinelli et al., 2021; Yusuf et al., 2014; Ribeiro et al., 2018). As a result, a carefully designed sampling strategy that mirrors real-world demographic and behavioral distributions is essential for producing valid and reliable simulations.

Q3. How to evaluate the massive social simulation results in a systematic way? Evaluation metrics for social simulations vary depending on the specific context and task. Most existing works primarily focus on employing LLMs during the assessment to generate scores directly according to the output natural language content (Liu et al., 2024; Li et al., 2024), which offers a limited and unsystematic approach to assess the full scope of simulation outcomes. On the other hand, human assessment of the LLM-generated content can be quite costly. Consequently, it is crucial to design a unified and quantifiable evaluation method to benchmark simulation results and provide comprehensive analyses.

In this paper, we propose the ***SSiLU*** framework, a generalized massive social simulation paradigm driven by LLM agents based on a large-scale real-world user pool to cope with the above challenges. Typically, we construct a 10-million-size user pool by collecting real-world social media data to support diverse and massive social simulations. Given a customized massive social simulation task, the task-specific prior distribution containing multiple demographic features is obtained first. Then simulated agents are sampled from the user pool by diverse sampling strategies to align with the customized distribution. During the simulation, a questionnaire or scale is designed to uniformly evaluate the simulation results, and each individual is required to answer the question in consistency with their given profile and experience in the real world.

We carry out **three** types of massive social simulations: (a) presidential election prediction, (b) breaking news feedback, and (c) national economic survey following the SSiLU framework and compare the simulated results with real-world ground truths, as shown in Figure 1. The extensive and comprehensive experiments have demonstrated that the SSiLU framework is of great help in constructing a standard and accurate massive social simulation. To conclude, the contributions in this paper are as follows:

- **SSiLU**: a generalized social simulation framework driven by LLM agents based on a large-scale real-world user pool, allowing for diverse simulating scenarios with high confidence aligned with the real-world distribution.
- **10M User Pool**: a 10-million-size user pool containing real users' behaviors to support massive simulation by collecting and combining data from social media platforms.
- **Unified Evaluation Method**: a questionnaire-based approach designed to systematically quantify different simulation results, enabling direct comparison with real-world conditions.
- **Three Applicable Simulations**: presidential election prediction, breaking news feedback, and national economic survey can help relevant researchers carry out further studies based on the SSiLU framework.

2 Related Works

2.1 Social Simulation Research

Traditional social simulation methods mainly rely on opinion polls, expert judgment, and statistical models (Burnap et al., 2016; Bohannon, 2017). The

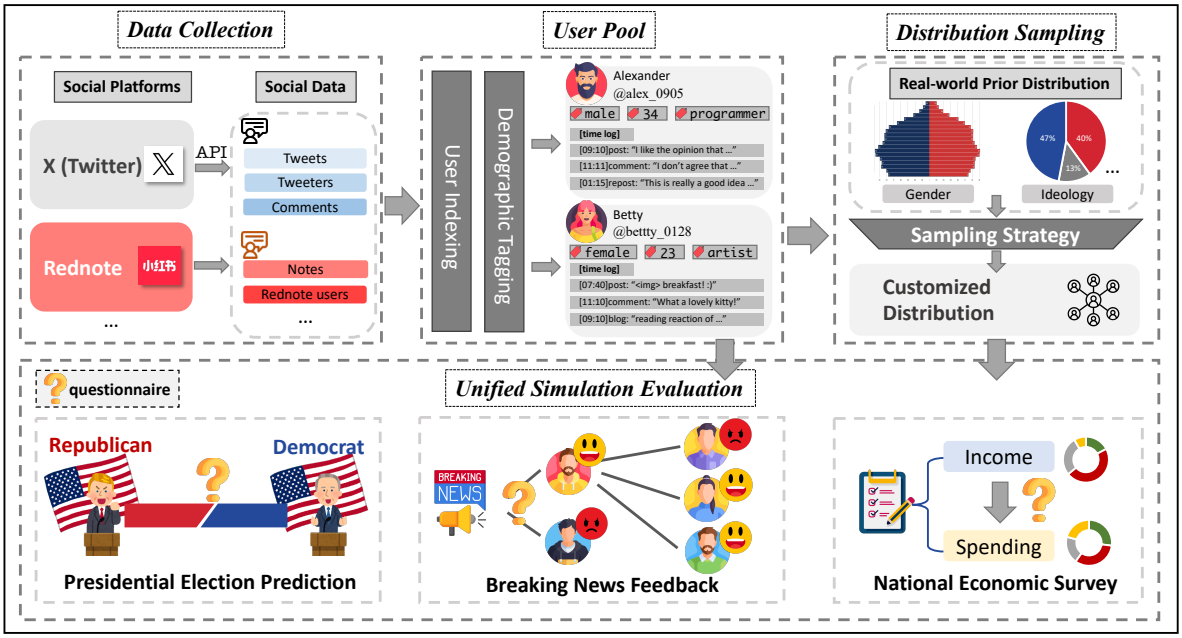

Figure 2: An illustration of ***SSiLU*** framework. We construct a 10M user pool through social media data. For a customized simulation, a prior distribution is calculated first to sample target agents, then a questionnaire-based simulation is conducted and compared with the real world.

ABM method provides a more objective and accurate prediction method by simulating individual behavior, combining micro-individual characteristics and macro-socioeconomic factors (Qiu and Phang, 2020; Sobkowicz, 2016). With the rapid development of LLM, researchers have discovered its potential to solve problems in social science (Linegar et al., 2023; Gujral et al., 2024). Preliminary research has shown positive outcomes in domains including electoral prediction, policy evaluation, etc. (Rozado, 2024; Moghimifar et al., 2024).

2.2 LLM Agent-based Simulation

Agent-based simulations powered by LLMs have gained wide attention recently for their promising application value and ability to solve general problems paradigm (Xi et al., 2023; Guo et al., 2024; Gao et al., 2024). Individual-level simulation focuses on highly reliable and reproducible human-like behavior (Shao et al., 2023; Xie et al., 2024; Sun et al., 2024), and task-level simulation pays more attention to the overall achievement of specific tasks and events (Du et al., 2023; Qian et al., 2024; Zhang et al., 2024; Yue et al., 2024). Task-level simulations also vary depending on different scenarios, wherein general-purpose scenarios highlight the intelligence within LLMs (Park et al., 2023; Mou et al., 2024b) while specific-domain scenarios emphasize the combination between workflows and domain specialization (Liu et al., 2024; Zhao et al., 2023; Lyu et al., 2024).

Source	# Users	# Posts
X	1,006,517	30,195,510
Rednote	9,158,404	40,963,735

Table 1: Statistical summary of the 10M user pool.

3 SSiLU

3.1 Overall Framework

The ***SSiLU*** framework follows a structured pipeline, as shown in Figure 2: (1) Social data are collected from multiple social media platforms, including both English- and Chinese-speaking communities. (2) Relevant users are extracted and annotated to construct a representative user pool. (3) Target groups are sampled from the user pool based on required demographic distributions. (4) Various large-scale social simulations are conducted and evaluated through a questionnaire-based unified evaluation, which closely resembles the real world.

3.2 Data Collection

Data Source The data source comprises X[1] and Rednote[2]. For X (formerly Twitter) data collection,

[1] https://x.com/
[2] https://www.xiaohongshu.com/

we use the official API to retrieve user posts. Specifically, before February 2023, posts were collected free of charge via Twitter, accounting for approximately 67% of the total posts. After February 2023, we began using the paid X API v2, which accounts for the remaining 33% of the data. For Rednote data collection, we obtained the data through an agreement with the Rednote platform, adhering to their privacy policy.

The diversity of data sources allows our user pool to encompass a broad distribution of user groups across different languages, cultures, and religions. We collect only posts (i.e., tweets and notes) along with engagement data, including the number of likes, comments, and reposts. These posts provide rich information from users.

Notice: all posts are collected from users who agree to share their public content, and no user profiles are collected.

Data Cleaning Anomalous data, such as advertising and robots, are filtered by calculating the post frequency and average text similarity. The detailed procedure can be found in Appendix A.

3.3 User Pool Construction

User Indexing We index users and construct a **user pool of 10 million users** based on the collected social media posts. Formally, we define **UserPool** as: $\mathbb{U} = \{u_i, p_i \mid i \in \mathbb{S}\}$, where the *i-th* user u_i derives from the collection of social media platforms $\mathbb{S}$ with his/her related posts $p_i = \{p_{i,1}, p_{i,2}, ...\}$. The statistical summary of the user pool is provided in Table 1.

Demographics Annotation Since user profiles are not accessible, we design a demographics annotation system to infer and tag demographic attributes. The process begins with multiple LLMs serving as initial annotators, classifying users across various demographic dimensions. Human annotators then evaluate and correct the LLM-generated labels, ensuring the reliability of the user tags dataset. The curated dataset is subsequently used to train demographic classifiers, cost-effectively enabling large-scale annotation. Specifically, we annotate users across 15 demographic dimensions: *age*, *gender*, *vocation*, *race*, *income*, *education*, *area*, *region*, *employment*, *marital*, *religious*, *party*, *ideology*, *BigFive personality*, and *hobbies*. Each attribute is inferred by a specialized classifier trained on the corresponding subset of the user tags dataset. See Appendix B for further details regarding annotation and classifier training.

3.4 Distribution Sampling Strategy

By constructing a 10M user pool, we enable the customization of group distribution for specific social simulations. The large scale and diversity of the user pool ensure flexible sampling strategies, which can be formulated as $U_S = Sampler(\mathbb{U}, D_P(i))$, where U_S and $D_P(i)$ denote the sampled users and prior distribution for the i-th task, respectively. For simulation scenarios with marginal demographic distributions (e.g., census data), we apply iterative proportional fitting (IPF) to estimate the joint distribution (Choupani and Mamdoohi, 2016). When the joint distribution is already known (e.g., online users), identical distribution sampling (IDS) is applied. Details of IPF and IDS are in Appendix C.

3.5 Unified Simulation Evaluation

The unified simulation evaluation involves a questionnaire scale to reflect the concern of the task quantitatively, which requires careful design involving domain experts. For simulations in a discrete label space, like representative election and attitudes simulation, the labels are directly transformed into options in the questionnaire. For simulations resulting in continuous results, like financial events, the options are formulated into numerical intervals. The strategy is formulated formally in Appendix D. Questionnaire answers from agents are converted into quantitative metrics, which are compared against real-world data or computed ground truth for evaluation.

4 Scenario Formulations

In this section, three large-scale social simulation scenarios are introduced following the ***SSiLU*** framework, i.e., presidential election prediction, breaking news feedback, and national economic survey. Each scenario is structured around four key components: *task formulation*, *prior distribution*, *questionnaire design*, and *comparison metrics*.

4.1 Presidential Election Prediction

Task Formulation The presidential election plays a pivotal role in shaping public engagement and party strategies (Bartels, 1996; Rosenstone, 1981). We use the U.S. presidential election campaign as a case to explore effective methods for achieving massive and diverse election simulations with LLMs, which follow an indirect voting system

through the Electoral College. Citizens vote for electors in their respective states, who then cast votes for the president. Each state has a set number of electors based on its congressional representation. Most states use a winner-takes-all system, where the candidate with the majority of votes receives all the state's electoral votes. We simulate the voting behavior of each agent in this task.

Prior Distribution Existing studies research the influence of demographics on elections (Major et al., 2018; Teixeira, 2009), which is considered a significant role in U.S. elections. We utilize data from the U.S. Census Bureau's Voting and Registration in the Election of November 2022, along with the 2020 Time Series Study from the American National Election Studies (ANES) (American National Election Studies, 2021) to capture the makeup of U.S. population, denoted as $\mathbb{A}$. Demographics including *age*, *gender*, *race*, *income*, *education*, *area*, *region*, *employment*, *marital*, *religious*, *party*, and *ideology* are considered to construct the overall prior distribution. Take all the users on the X in our user pool as $\mathbb{U}_X$, we employ iterative proportional fitting sampling (IPF) is to sample target agents from the user pool given marginal distributions, i.e., $U_S = IPF(\mathbb{U}_X, D_{\mathbb{A}})$.

Questionaire Design We design the presidential election questionnaire based on abundant polls carried out by different media and research institutes (Barnett and Sarfati, 2023; Keeter et al., 2021) to include both concerning issues and voting-behavior options, and optimize them into proper forms for LLM-based agents. The whole questionnaire can be found in Appendix G.1.

Comparision Metric Two metrics are used to comprehensively compare the simulated election results to the real world. (1) Accuracy rate (Acc) is measured by calculating the proportion of states for which the election simulation results align with the actual result. (2) Root Mean Square Error (RMSE) is measured by calculating the simulated vote share and the actual vote share for each state, which serves as a fine-grained evaluation metric.

4.2 Breaking News Feedback

Task Formulation Journalism shapes public perception and opinion by providing information and framing narratives through media coverage (van Dalen, 2024; Gómez-Calderón and Ceballos, 2024). Online social media platforms have gradually replaced the influence of traditional paper media. Every time when breaking news is released on social media platforms, its potential audience may hold different stances and react toward the news differently. We take *"the release of ChatGPT"* as our target news to evaluate the consistency and foreseeability of public attitudes.

Prior Distribution We take all the users on the rednote in our user pool as the universal set $\mathbb{U}_R$. We collect the users interested in the technology area as the **potential audience set** $\mathbb{P}$, and we take the users who have mentioned ChatGPT directly as the **ground truth set** $\mathbb{G}$ through keyword matching. It can be formulated that $\mathbb{G} \subset \mathbb{P} \subset \mathbb{U}_R$. The matched posts of users within $\mathbb{G}$ are used to calculate ground truth. The distribution of $\mathbb{P}$ is viewed as the prior distribution. We employ identical distribution sampling (IDS) on the $\mathbb{U}_R$, which can be formulated as $U_S = IDS(\mathbb{U}_R, D_{\mathbb{P}})$. During sampling, demographics like *gender*, *age*, *education*, and *consumption* are considered. Posts after the release of the news are masked so that $U_s \cap \mathbb{G} = \emptyset$.

Questionaire Design We design the public cognitive questionnaire based on the theory of the Affect, Behavior, and Cognition model (ABC model) (Liu et al., 2021). This model is particularly useful for analyzing acceptance pathways and the interactions between these components. Additionally, the 5-point Likert scale (Joshi et al., 2015) is combined to divide the questionnaire into six dimensions, i.e., public cognition (PC), perceived risks (PR), perceived benefits (PB), trust (TR), fairness (FA), and public acceptance (PA). The whole questionnaire can be found in Appendix G.2.

Comparison Metric Distribution evaluation involves two aspects: (1) RMSE is measured between the answers between simulated answers and ground truth answers in Likert dimensions. (2) KL-divergence (KL-Div) is measured by taking the 6-dimensional answer jointly as a distribution and calculating between the simulated results and the ground truth.

4.3 National Economic Survey

Task Formulation Economic simulation is a crucial part of massive social simulations as it models resource distribution, market dynamics, and financial behaviors, providing insights into economic stability and policy impacts (Dignum et al., 2020; Trimborn et al., 2020). Integrating economic fac-

Scenario	# Agents	# Demographics	Type	Sampling	Source	Language	# Questions	Ground truth
PresElectPredict	33,1836	12	label	IPF	X	EN	49	real world
BreakNewsFeed	20,000	7	label	IDS	rednote	ZH	18	calculated
NatEconSurvey	16,000	9	label+number	IDS	rednote	ZH	17	real world

Table 2: Detail settings of three simulation scenarios, where PresElectPredict, BreakNewsFeed, and NatEconSurvey denote three simulations, respectively. Details of IPF and IDS can be found in Appendix C.

tors with social interactions helps predict systemic outcomes, guiding decision-making in areas such as governance and crisis management. We conduct a simulation following a real-world national economic survey, which interviews Chinese citizens on their monthly spending, given the average salary of each province in China (NBS China, 2023b).

Prior Distribution The prior distribution is based on the methodology from the National Bureau of Statistics of China, which takes 160,000 families nationwide and calculates their incomes and spending as the national average statistics (NBS China, 2023b). Take the Chinese population as $\mathbb{C}$, we sample nationwide agents from the user pool in proportion to their *region* population and generate their *income* distribution according to the regional average income (NBS China, 2023a). Formally, $U_S = IDS(\mathbb{U}_R, D_{\mathbb{C}})$. The detailed method can be referred to in Appendix C.3.

Questionaire Design Spending details in China Statistical Yearbook 2024 (NBS China, 2024) are categorized into eight parts, i.e. *food*, *clothing*, *housing*, *daily necessities & services*, *communication & transportation*, *education & entertainment*, *healthcare*, and *others*. Consequently, the questionnaire design covers the above categories with multiple examples. Options are formulated into segmented interval options for each question. The whole questionnaire can be found in Appendix G.3.

Comparison Metric Distribution evaluation involves two aspects: (1) RMSE of the nine categories is measured between the simulated results and official statistics. (2) KL-Div is measured by taking the 8-dimensional spending as a distribution to evaluate the overall consistency.

5 Experiments

5.1 Experiment Settings

Models We select powerful large-scale LLMs from different model families. For open-source models, we select Llama-3-70b-Instruct (Dubey et al., 2024), Qwen2.5-72b-Instruct (Yang et al., 2024), DeepSeek-V3-671b and DeepSeek-R1-671b (Guo et al., 2025). For commercial models, we select GPT-4o[3] (OpenAI, 2024b) and GPT-4o-mini[4] (OpenAI, 2024a).

Implementation Details We compare the settings of all three scenarios for better understanding, as shown in Table 2. The PresiElectPredict covers a 1/1000 sample of the U.S. citizen population (over 33K agents). Thus, some results are not reported due to the cost restriction. The example prompts during the simulation can be found in Appendix F

In terms of LLM serving, Qwen2.5-72b-Instruct, and Llama3-70b-Instruct models are both deployed on 8*NVIDIA RTX4090 GPUs via vLLM (Kwon et al., 2023). We set max tokens to 2048, and set the temperature to 0.7 to encourage diversity during the generation. DeepSeek-V3-671b and DeepSeek-R1-671b are called through APIs.

5.2 Overall Results

The overall simulation results are shown in Table 3. We also report subset results for PresiElectPredict and NatEconSurvey.

Presidential Election Prediction We report the overall results and the battleground states' results separately. The prediction of battleground states is challenging even in the real world, and thus becomes the focus during the election process. According to the results, DeepSeek-V3-671b and Qwen2.5-72b show competitive performance both in Acc and RMSE. Typically, according to the winner-takes-all rule, **over 90% of state voting results are predicted correctly**, which means the simulation achieves a high-precision macroscopic reduction of the real-world election results. After the case study, we find that DeepSeek-R1-671b sometimes falls into overthinking, resulting in less accurate results. As only the voting-behavior question is reported here, we provide performance on the full-size questionnaire in Appendix E.1.

[3] `gpt-4o-2024-08-06`
[4] `gpt-4o-mini-2024-07-18`

Model	PresElectPredict				BreakNewsFeed		NatEconSurvey			
	Overall		*Battleground*				*Overall*		*Developed-Region*	
	Acc↑	RMSE↓	Acc↑	RMSE↓	KL-Div↓	RMSE↓	KL-Div↓	RMSE↓	KL-Div↓	RMSE↓
Llama3 $_{70B}$	0.843	0.064	0.733	0.045	0.668	0.199	**0.016**	**0.026**	**0.013**	**0.025**
Qwen2.5 $_{72B}$	**0.922**	**0.037**	0.800	**0.031**	**0.113**	0.059	0.066	0.048	0.043	0.039
DeepSeek-R1 $_{671B}$	\	\	0.670	0.065	0.383	0.082	0.059	0.045	0.045	0.036
DeepSeek-V3 $_{671B}$	**0.922**	0.046	**0.867**	0.041	0.263	0.072	0.035	0.036	0.023	0.030
GPT-4o-mini	\	\	0.800	0.039	0.195	0.114	0.046	0.045	0.030	0.036
GPT-4o	\	\	\	\	0.196	**0.055**	0.062	0.051	0.036	0.038

Table 3: Overall results of the three scenarios, where subset *Battleground* indicates battleground states in the U.S. in the presidential election and subset *Developed-Region* indicates top-10 developed regions in China in terms of GDP.

Breaking News Feedback The results measure the overall consistency of each model compared with the real-world users' reactions and attitudes. To this end, the performances of GPT-4o and Qwen2.5-72b are more aligned with real-world perspectives than other models in terms of KL-Div and RMSE, respectively. Generally, **the models consistently capture and accurately predict public trends and opinions**, which is also shown in §6.3. We also conduct simulations on more news and on ground truth from real-world humans in Appendix E.2 in a smaller size.

National Economic Survey We report the overall results and results for the top 10 regions by GDP (i.e., developed regions) separately. Generally, all the models closely align with real-world statistics. Llama3-70b shows a significant superiority over other models in the economic survey scenario, and all the models perform better in the *Developed-Region* subset than overall. The results demonstrate that **individuals' spending habits can be accurately reproduced under our framework, especially in developed regions**. We provide results on each question dimension with further discussion in §6.4.

Through the overall results of three simulations, ***SSiLU*** supports diverse and accurate massive social simulations with a standard pipeline and minimal changes with human experts in the loop. However, LLMs can impact the performance under different scenarios, which deserves further research.

6 Further Analysis

6.1 Ex: Multi-round Interactive Simulation

Notably, three simulations carried out in Section 5 are all single-round simulations without interactions among agents. Nevertheless, the architecture is already compatible with multi-round interaction simulations, which are not fully discussed due to the lack of large-scale interactive datasets. Consequently, we implement an extensive simulation under the *multi-round interactive setting* in a smaller size. Following previous work (Chuang et al., 2023), we employ 35 LLM-based agents as a group for two parties in the U.S. and simulate the wisdom of partisan crowds (Becker et al., 2019) through 3 rounds of simulations. Altogether, 12 paired groups (840 agents) are simulated.

Model	*HLI*↑	$\overline{\Delta\varepsilon}$ ↓	$\overline{\beta_{PB}}$ ↑	*Ext.*↓
GPT-4o	-3.08 ± 2.39	**-1.27 ± 2.49**	-4.35 ± 1.25	0.00
GPT-4o-mini	-6.35 ± 1.65	0.31 ± 0.41	-6.04 ± 1.82	0.00
DeepSeek-V3	**0.18 ± 1.96**	0.02 ± 0.24	**0.20 ± 1.97**	0.00
Qwen2.5-72b	-1.46 ± 2.08	-0.19 ± 0.56	-1.65 ± 1.66	0.00
Human	66.5 ± 6.79	-33.16 ± 6.74	33.35 ± 0.83	8.37

Table 4: Multi-round interactive simulation results. *HLI*, $\overline{\Delta\varepsilon}$, $\overline{\beta_{PB}}$, and *Ext.* are specified in Appendix E.3.

We follow the procedure of Chuang et al. during the simulation and evaluation, except that the user profiles are sampled from our user pool in proportion to the U.S. population, which is generated by LLM in previous work. The details can be found in Appendix E.3. As shown in Table 4, ***SSiLU*** **realizes multi-round interactive simulation with comparable results to previous works**.

6.2 Ablation: Are Prior Distribution and Real-World Knowledge Truly Important?

We conduct an ablation study on the presidential election prediction simulation to assess the impact of prior demographics distribution and real-world user knowledge. As shown in Table 5, prior demographics distribution significantly improves the accuracy of the simulation in both Acc and RMSE compared to random demographics distribution.

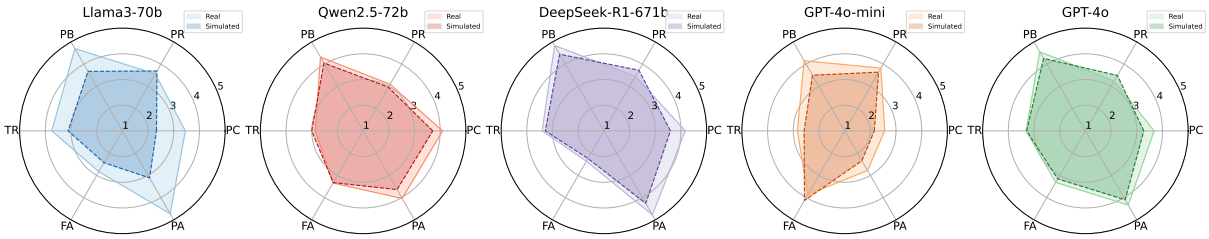

Figure 3: An illustration of the performances of the breaking news feedback simulation, where PC, PR, PB, TR, FA, and PA denote six dimensions from the Likert scale (see §4.2 questionnaire design), with 1-point standing for totally disagree and 5-point for totally agree. Table 16 in Appendix E.4 provides supplemental information.

Model	Acc↑	RMSE↓
Llama3-70b	**0.733**	**0.045**
- w/o Knowledge	0.533	0.051
- w/o Knowledge & PirorDistribution	0.600	0.386
Qwen2.5-72b	**0.800**	**0.031**
- w/o Knowledge	**0.800**	0.033
- w/o Knowledge & PriorDistribution	0.600	0.370
GPT-4o-mini	**0.800**	**0.039**
- w/o Knowledge	**0.800**	0.052
- w/o Knowledge & PriorDistribution	0.667	0.323

Table 5: Ablation experiment results on the presidential election prediction, where -w/o Knowledge denotes *without user posts* and -w/o Prior Distribution denotes *using random demographics distribution*.

Item	Llama3	Qwen2.5	4o-mini	4o	R1
Daily	**0.007**	0.009	**0.006**	0.010	**0.009**
Clothing	0.012	0.015	0.019	0.015	0.015
Transp_Comm	0.016	0.020	0.027	0.023	0.017
Educat_Entert	0.018	0.022	0.024	0.017	0.022
Medical	0.023	0.062	0.041	0.057	0.060
Food	0.037	0.031	0.031	0.040	0.032
Household	0.052	0.110	0.107	0.120	0.102
Others	0.008	**0.008**	0.010	**0.005**	**0.009**

Table 6: Detailed results on the national economic survey simulation reported in NRMSE, where the Item column indicates the components of spending. The best results are **bolded**; the second-best results are underlined.

Additionally, past posts from users improve the fine-grained performance, especially for Llama3-70b in Acc and all models in RMSE. The ablation study shows that **both prior distribution and real-world knowledge in the *SSiLU* pipeline are significant during the simulation**.

6.3 Can Group Preference and Perspectives Be Well Reflected?

During the Breaking News Feedback, the core concern is whether the preferences and perspectives of the target group are well captured and reflected in the results. As the ground truth of the simulation is calculated by prompting LLM agents from the ground truth set $\mathbb{G}$, the *Simulated* and *Real* results are paired for each model, as shown in Figure 3. **All the models tend to behave consistently with the ground truth.**. However, Llama3-70b performs poorly with a larger gap than other models. GPT-4o-mini exhibits divergent attitudes in the dimensions of fairness (FA) and public acceptance (PA), potentially due to the news content being associated with OpenAI. The cover area difference between *Real* and *Simulated* shows that **all models tend to perform more disagreeably in the *Simulated* results than the *Real***, which also underlines the potential risk of biases during the public opinion simulation, which is also shown in Table 16.

6.4 In Which Domain Do LLMs Predict Better/Worse?

The NatEconSurvey covers 8 spending dimensions, as mentioned in §4.3. Besides the average performance of these dimensions, model performances among these dimensions can also vary. We calculate the averaged RMSE nationwide on each spending level, as shown in Table 6. It is worth mentioning that all the models show high consistency. Eliminating the *others* item, **all the models perform best on *daily necessities* and worst on *housing***, which can reveal the LLM's preference on economic decision-making and highlight the challenge in *household* spending strategy.

7 Conclusion

We introduce the ***SSiLU*** for massive social simulations powered by LLM agents, featuring a 10-million-user pool enriched with real-world knowledge, a demographic distribution sampling strategy, and a unified simulation evaluation method. Through extensive simulations and diverse evaluations across political, journalistic, and economic scenarios, our results demonstrate the framework's effectiveness, scalability, and generalizability.

Limitations

SSiLU aims at generalized and standard massive social simulation, which depends on its large-scale user pool and adaptive simulation method. However, there may be some underlying limitations.

User Pool Bottleneck The generalization ability depends on the large-scale size of the user pool, which enables a large range of group distributions. Although we build a 10M user pool from multiple social media platforms, there may exist potential minority groups that cannot be fully represented, which will influence the performance of related simulations. Consequently, more groups are supposed to be included in the current user pool in future works.

Rigorous Expertise Requirement During the simulation pipeline, questionnaire design and prior distribution research involve expertise in relevant fields. Although the structure and pipeline require minimal changes during the simulation, rigorous expertise demands may pose certain challenges for researchers in conducting further studies, which is also a common challenge that needs to be considered and addressed in social simulations.

Ethics Statement

We clarify the potential ethical concerns as follows.

Ethical Concerns during the Collection of Data During the collection of the data from X, we employ the official API to request the posts. According to the X's privacy policy, only data from users who agree to share their data can be accessed via the API. Thus, there is no condition that the data are collected in contrast with the user's will. According to the X's developer terms, we do not, and will not, re-identify the user through the collected posts. We promise that all the demographic features are generated by LLMs or human annotators.

During the collection of the data from Rednote. We achieve an agreement with the platform and carry out the research in compliance with a similar private policy to the X.

Ethical Concerns during the Use of Data During the simulation, we sample users according to the target demographic distribution based on our annotated features. We do not collect, infer, or store any personally identifiable information (PII), and all data processing adheres to the principles of data minimization and anonymity. We neither intend to, nor are able to, simulate any real-world individual. Instead, our framework aims to simulate representative population groups based on demographic distributions. Personal experiences that appear in the data are only leveraged in aggregate to improve the accuracy of modeling group-level behaviors, rather than for reproducing individual behavior patterns.

In terms of the release of data, according to the platform's policy, we will release the X user pool with only user IDs and annotated features, which means all the posts will NOT be open-sourced. Researchers must request the corresponding posts via official APIs. This approach guarantees that whenever the user chooses to stop sharing data, the posts will not be collected for other simulations and reproductions.

References

American National Election Studies. 2021. Anes 2020 time series study full release [dataset and documentation]. February 10, 2022 version.

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):337–351.

Arnold Barnett and Arnaud Sarfati. 2023. The polls and the us presidential election in 2020.... and 2024. *Statistics and Public Policy*, 10(1):2199809.

Larry M Bartels. 1996. Uninformed votes: Information effects in presidential elections. *American journal of political science*, pages 194–230.

Joshua Becker, Ethan Porter, and Damon Centola. 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22):10717–10722.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

John Bohannon. 2017. The pulse of the people.

Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 2016. 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies*, 41:230–233.

Abdoul-Ahad Choupani and Amir Reza Mamdoohi. 2016. Population synthesis using iterative proportional fitting (ipf): A review and future research. *Transportation Research Procedia*, 17:223–233.

Yun-Shiuan Chuang and Timothy T Rogers. 2023. Computational agent-based models in opinion dynamics: A survey on social simulations and empirical studies. *arXiv preprint arXiv:2306.03446*.

Yun-Shiuan Chuang, Siddharth Suresh, Nikunj Harlalka, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. *arXiv preprint arXiv:2311.09665*.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Frank Dignum, Virginia Dignum, Paul Davidsson, Amineh Ghorbani, Mijke van der Hurk, Maarten Jensen, Christian Kammler, Fabian Lorig, Luis Gustavo Ludescher, Alexander Melchior, et al. 2020. Analysing the combined health, social and economic impacts of the corovanvirus pandemic using agent-based social simulation. *Minds and Machines*, 30:177–194.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Ming Gao, Zhongyuan Wang, Kai Wang, Chenhui Liu, and Shiping Tang. 2022. Forecasting elections with agent-based modeling: Two live experiments. *Plos one*, 17(6):e0270194.

Salvatore Giorgi, Veronica E Lynn, Keshav Gupta, Farhan Ahmed, Sandra Matz, Lyle H Ungar, and H Andrew Schwartz. 2022. Correcting sociodemographic selection biases for population prediction from social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 228–240.

Pratik Gujral, Kshitij Awaldhi, Navya Jain, Bhavuk Bhandula, and Abhijnan Chakraborty. 2024. Can llms help predict elections?(counter) evidence from the world's largest democracy. *arXiv preprint arXiv:2405.07828*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Bernardo Gómez-Calderón and Yaiza Ceballos. 2024. Journalism and artificial intelligence. the treatment of the chatbots in the spanish press. *index.comunicación*, 14(1):281–300.

Jesse Hoey, Tobias Schröder, Jonathan Morgan, Kimberly B Rogers, Deepak Rishi, and Meiyappan Nagappan. 2018. Artificial intelligence and social simulation: Studying group dynamics on a massive scale. *Small Group Research*, 49(6):647–683.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

Marko Jusup, Petter Holme, Kiyoshi Kanazawa, Misako Takayasu, Ivan Romić, Zhen Wang, Sunčana Geček, Tomislav Lipić, Boris Podobnik, Lin Wang, et al. 2022. Social physics. *Physics Reports*, 948:1–148.

Scott Keeter, Nick Hatley, Arnold Lau, and Courtney Kennedy. 2021. What 2020's election poll errors tell us about the accuracy of issue polling. *Pew Research Center Methods*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Sanguk Lee, Tai-Quan Peng, Matthew H Goldberg, Seth A Rosenthal, John E Kotcher, Edward W Maibach, and Anthony Leiserowitz. 2023. Can large language models capture public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *arXiv preprint arXiv:2311.00217*.

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

Mitchell Linegar, Rafal Kocielnik, and R Michael Alvarez. 2023. Large language models and political science. *Frontiers in Political Science*, 5:1257092.

Bingsheng Liu, Yinghua Xu, Yang Yang, and Shijian Lu. 2021. How public cognition influences public acceptance of ccus in china: Based on the abc (affect, behavior, and cognition) model of attitudes. *Energy Policy*, 156:112390.

Xiawei Liu, Shiyue Yang, Xinnong Zhang, Haoyu Kuang, Libo Sun, Yihang Yang, Siming Chen, Xuanjing Huang, and Zhongyu Wei. 2024. Ai-press: A multi-agent news generating and feedback simulation system powered by large language models. *arXiv preprint arXiv:2410.07561*.

Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2024. Gpt-4v(ision) as a social media analysis engine. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Charles M Macal and Michael J North. 2009. Agent-based modeling and simulation. In *Proceedings of the 2009 winter simulation conference (WSC)*, pages 86–98. IEEE.

Brenda Major, Alison Blodorn, and Gregory Major Blascovich. 2018. The threat of increasing diversity: Why many white americans support trump in the 2016 presidential election. *Group Processes & Intergroup Relations*, 21(6):931–940.

Farhad Moghimifar, Yuan-Fang Li, Robert Thomson, and Gholamreza Haffari. 2024. Modelling political coalition negotiations using llm-based agents. *arXiv preprint arXiv:2402.11712*.

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

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

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024c. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. *arXiv preprint arXiv:2402.16333*.

Goran Murić, Alexey Tregubov, Jim Blythe, Andrés Abeliuk, Divya Choudhary, Kristina Lerman, and Emilio Ferrara. 2022. Large-scale agent-based simulations of online social networks. *Autonomous Agents and Multi-Agent Systems*, 36(2):38.

NBS China. 2023a. Communiqué of the Seventh National Population Census of the People's Republic of China. Technical report. Accessed: 2025-02-14.

NBS China. 2023b. Explanatory Notes on Main Statistical Indicators – Population, Society, and Labor (China Statistical Yearbook 2023). Accessed: 2025-02-14.

NBS China. 2024. China Statistical Yearbook 2024. Accessed: 2025-02-14.

OpenAI. 2024a. GPT-4o Mini: Advancing Cost-Efficient Intelligence. Accessed: 2025-02-14.

OpenAI. 2024b. GPT-4o System Card. Technical report. Accessed: 2025-02-14.

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.

Pew Research Center. 2023. Views of artificial intelligence – topline survey results. `https://www.pewresearch.org/wp-content/uploads/2023/08/SR_23.08.28_views-of-ai_topline.pdf`. Topline survey data accompanying the report "Growing Public Concern About the Role of Artificial Intelligence in Daily Life".

Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2024. Representation bias in political sample simulations with large language models. *arXiv preprint arXiv:2407.11409*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Lin Qiu and Riyang Phang. 2020. Agent-based modeling in political decision making.

Filipe Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna Gummadi. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Steven J Rosenstone. 1981. Forecasting presidential elections.

David Rozado. 2024. The political preferences of llms. *arXiv preprint arXiv:2402.01789*.

Thomas C Schelling. 1969. Models of segregation. *The American economic review*, 59(2):488–493.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Pawel Sobkowicz. 2016. Quantitative agent based model of opinion dynamics: Polish elections of 2015. *PloS one*, 11(5):e0155098.

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

Shiping Tang. 2024. Idea, action, and outcome. *Innovation in the Social Sciences*, 2(2):123–170.

Ruy A Teixeira. 2009. *Red, blue, and purple America: the future of election demographics*. Rowman & Littlefield.

The New York Times. 2024a. As debate looms, trump is now the one facing questions about age and capacity. *The New York Times*. Politics news.

The New York Times. 2024b. Israel strikes hezbollah as nasrallah vows retribution. *The New York Times*. International affairs news.

The New York Times. 2024c. Offshore wind slowed by broken blades, rising costs and angry fishermen. *The New York Times*. Energy news.

Torsten Trimborn, Philipp Otte, Simon Cramer, Maximilian Beikirch, Emma Pabich, and Martin Frank. 2020. Sabcemm: A simulator for agent-based computational economic market models. *Computational economics*, 55(2):707–744.

Alec Tyson and Emma Kikuchi. 2023. Growing public concern about the role of artificial intelligence in daily life. *Pew Research Center*.

Arjen van Dalen. 2024. Revisiting the algorithms behind the headlines. how journalists respond to professional competition of generative ai. *Journalism Practice*, pages 1–18.

Emily Vraga. 2016. Party differences in political content on social media. *Online Information Review*, 40(5):595–609.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

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*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2024. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. *arXiv preprint arXiv:2407.09893*.

Nadia Yusuf, Nisreen Al-Banawi, and Hajjah Abdel Rahman Al-Imam. 2014. The social media as echo chamber: The digital impact. *Journal of Business & Economics Research (Online)*, 12(1):1.

Xinnong Zhang, Jiayu Lin, Libo Sun, Weihong Qi, Yihang Yang, Yue Chen, Hanjia Lyu, Xinyi Mou, Siming Chen, Jiebo Luo, et al. 2024. Electionsim: Massive population election simulation powered by large language model driven agents. *arXiv preprint arXiv:2410.20746*.

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*.

A Data Cleaning Details

A.1 Content Data Extraction

We extract only post-related content on all the social media platforms to avoid violating privacy policies. Specifically, the data list on each platform is shown in Table 7.

Platform	Data list
X	user ID, tweet, #likes, #coments, #retweets
Rednote	user ID, notes, #likes, #comments

Table 7: Data list for each social media platform during the data collection.

A.2 Abnormal Data Filtering

We filter the abnormal data to guarantee the quality through text similarity calculation. Typically, all the textual content from the same user is calculated by means of the word repetition ratio. The threshold is set to 0.3. If the ratio surpasses the threshold, the user is considered to be likely a robot or advertising and will be filtered.

B Demographics Annotation System

B.1 LLM Annotation

To save costs, we first sample a subset of the user pool and employ multiple power LLMs for annotation. Due to the long time span of this work, users from different data sources in the user pool have used the powerful LLMs available at the time. For users derived from the X, GPT-4o[5], Claude3.5-Sonnet[6], and Gemini-1.5[7] are employed. For users

[5] `gpt-4o-2024-08-06`
[6] `claude-3-5-sonnet-20240620`
[7] `gemini-1.5-pro`

derived from the Rednote, GPT-4o, Cluade3.5-Sonnet, and Qwen2.5-72b are employed.

B.2 Human Evaluation

We employ 7 professional human annotators to verify the results annotated by LLMs. Typically, each annotator is required to re-annotate the demographic factors without the LLM labels. All the data are verified by at least 2 human annotators. The overall consistency between humans and LLMs is shown in Table 8, which denotes the naive pair-wise matched accuracy.

Models	Human (X)	Human (Rednote)
GPT-4o	0.905	0.723
Claude3.5	0.901	0.659
Gemini-1.5	0.713	\
Qwen2.5	\	0.846
Majority votes	**0.956**	**0.849**

Table 8: Human annotators' verification results. We report the pairwise matched accuracy between humans and different LLMs.

We all calculate the pairwise matched accuracy and Kappa consistency among all annotators, which is shown in Table 9 and Table 10 for X and Rednote, respectively.

X	GPT-4o	Claude3.5	Gemini1.5	Majority	Human
GPT-4o	\	0.600	0.465	**0.919**	0.797
Claude3.5	0.796	\	0.426	0.914	0.785
Gemini1.5	0.645	0.620	\	0.807	0.550
Majority	0.896	0.865	0.749	\	0.891
Human	0.897	0.892	0.704	**0.946**	\

Table 9: Inter-annotator performance of X user pool. The *bottom left part* results are reported in **pairwise matched accuracy**. The *top right part* results are reported in **Kappa consistency**.

Rednote	GPT-4o	Claude3.5	Qwen2.5	Majority	Human
GPT-4o	\	0.403	0.515	**0.736**	0.586
Claude3.5	0.727	\	0.363	0.613	0.437
Qwen2.5	0.763	0.732	\	0.734	0.706
Majority	0.875	0.846	**0.883**	\	0.729
Human	0.799	0.767	0.867	**0.883**	\

Table 10: Inter-annotator performance of Rednote user pool. The *bottom left part* results are reported in **pairwise matched accuracy**. The *top right part* results are reported in **Kappa consistency**.

B.3 Classifier Training

We take the majority-voted labels from different LLMs to construct the training dataset. Considering the difference in mainstream language used on different platforms, we employ LongFormer (Beltagy et al., 2020) for X data and employ Bert-base-chinese (Devlin et al., 2019) for Rednote. The implementation details are shown in Table 11.

Params	LongFormer	Bert-base-chinese
train_size	10,000	10,000
# classifiers	5	4
max_tokens	4096	512
learning_rate	5e-5	5e-5
batch_size	16	32
optimizer	AdamW	AdamW
epochs	3	10
device	8*4090	2*4090

Table 11: Implementation details for demographic classifiers.

We report the performances of demographic classifiers on each demographic factor in Table 12.

Demos	LongFormer		Bert-base-chinese	
	Acc	F1	Acc	F1
Gender	0.875	0.904	0.926	0.958
Age	0.902	0.873	0.925	0.920
Party	0.849	0.846	\	\
Ideology	0.810	0.807	\	\
Race	0.779	0.768	\	\
Consumption	\	\	0.749	0.748
Education	\	\	0.954	0.975

Table 12: Performance of demographic classifiers on test set.

B.4 Overall Distribution of the User Pool

We employ the demographic classifiers to annotate all of the users in the user pool and the overall distributions are shown in Figure 4. For other demographics in specific simulations that are not considered in prior distribution, only users from the sampled user pool are annotated by majority votes of LLMs.

C Demographic Distribution Sampling Details

C.1 Iterative Proportional Fitting

In our study, we follow the classical IPF method to construct the joint distribution of all the attributes in our simulation. Specifically, we start with a two-way table with individual components denoted

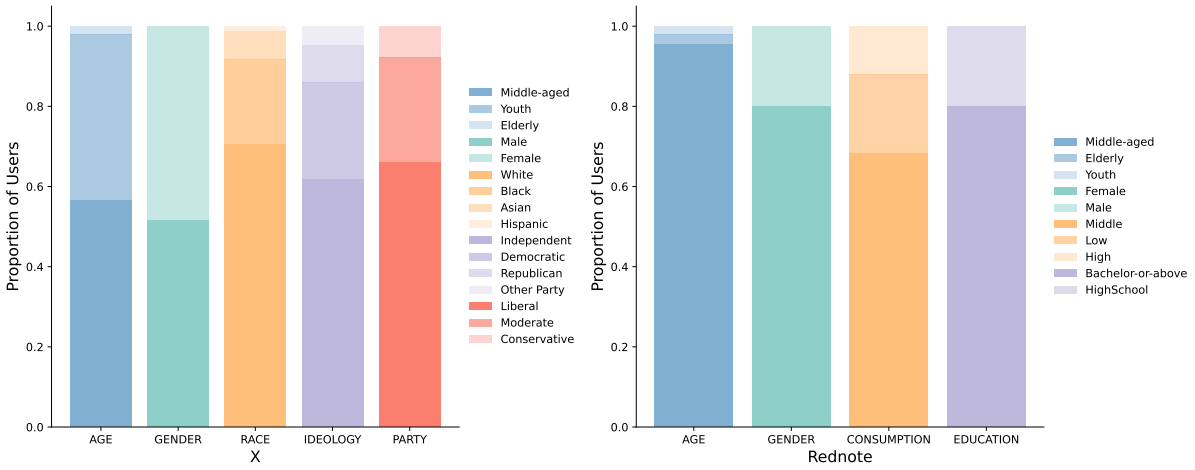

Figure 4: Demographic distribution on X and Rednote user pool.

as x_{ij} and targeted estimation $\hat{x}_{ij}$. The targeted estimation $\hat{x}_{ij}$ satisfies $\Sigma_j \hat{x}_{ij} = v_i$ and $\Sigma_i \hat{x}_{ij} = w_j$. The iterations are specified as follows:

Let $\hat{x}_{ij}^{(0)} = x_{ij}$. For $\alpha > 1$:

$$\hat{x}_{ij}^{(2\alpha-1)} = \frac{\hat{x}_{ij}^{(2\alpha-2)} v_i}{\Sigma_{k=1}^{J} \hat{x}_{ij}^{(2\alpha-2)}} \tag{1}$$

$$\hat{x}_{ij}^{(2\alpha)} = \frac{\hat{x}_{ij}^{(2\alpha-1)} w_j}{\Sigma_{k=1}^{I} \hat{x}_{ij}^{(2\alpha-1)}} \tag{2}$$

The iterations end when the estimated marginals are sufficiently close to the real marginals or when they stabilize without further convergence.

For the presidential election simulation, we implement the IPF algorithm for each state using five attributes: *gender*, *race*, *age group*, *ideology*, and *partisanship*. In most cases, the algorithm does not converge, but the gaps between the estimated and actual marginals are less than 5%, with 888 out of 918 marginals falling within this range. For the outliers, since IPF adjusts proportionally to the marginals, the overall ratio of marginals remains consistent. We then use the estimated joint distribution and marginals for our massive simulation.

C.2 Identical Distribution Sampling

Identical distribution sampling, also known as direct sampling, is applied when the joint distribution of multiple demographics is available. Given feature X and Y, the joint distribution can be formulated as $p(X, Y)$. Then identical distribution sampling can be formulated as follows:

$$(X_i, Y_i) \sim p(X, Y) \quad i = 1, 2, ..., n \tag{3}$$

For breaking news feedback simulations, as the ground truth set is directly from the Rednote, we can obtain all the users' demographics and calculate the joint distribution. Simultaneously, the scale of the user pool satisfies the direct sampling requirements.

C.3 Prior Distribution of National Economic Survey

For the national economic survey distribution, only average income is available from the official data. As a result, we generate the prior income distribution at the regional level. The income distribution across different regions exhibits significant heterogeneity, often characterized by a right-skewed pattern. To model this distribution, we adopt a mixture distribution approach, combining a log-normal distribution for the majority of the population with a Pareto distribution for the high-income segment. This hybrid model captures both the bulk of wage earners and the long-tail effect observed in high-income groups.

Formally, let X denote an individual's wage. We assume that for the lower and middle-income groups ($X < x_{min}$), incomes follow a log-normal distribution:

$$X \sim \log \text{Normal}\left(\mu, \sigma^2\right) \tag{4}$$

where

$$\mu = \ln\left(\frac{\mu_{\text{actual}}^2}{\sqrt{\sigma_{\text{actual}}^2 + \mu_{\text{actual}}^2}}\right), \quad \sigma = \sqrt{\ln\left(1 + \frac{\sigma_{\text{actual}}^2}{\mu_{\text{actual}}^2}\right)} \tag{5}$$

For the high-income group ($X \geq x_{min}$), wages follow a Pareto distribution:

$$P(X \geq x) = Cx^{-\alpha}, \quad x \geq x_{\min} \tag{6}$$

where α is the Pareto shape parameter determining the income concentration at the top. The proportion of individuals assigned to each distribution is governed by an empirical threshold ratio, typically set such that 90% of the population follows the log-normal distribution while 10% follows the Pareto distribution. This mixture approach provides a flexible yet robust framework for simulating realistic income distributions across diverse economic conditions. We set all the parameters empirically according to previous research and generate the income distribution for 31 regions in China (Hong Kong, Macao, and Taiwan are excluded).

D Algorithm for Unified Simulation Evaluation

A strategy of unified simulation evaluation is specified as follows:

Algorithm 1 Unified Evaluation Strategy

Input: Simulation result type
if result is continuous **then**
    Segment the result space into options at numeric intervals.
else if result is discrete **then**
    Map the label space to options with values (e.g., 1–5).
end if
Input: Whether simulation outputs a specific event (e.g., election winner)
if Yes **then**
    Value evaluation: Use Accuracy as the metric.
    Distribution evaluation: Use KL divergence and RMSE as the metrics.
else
    Only perform the Distribution evaluation using KL divergence and RMSE.
end if

E Supplementary Experiment Materials

E.1 Full Questionnaire Simulation Results in Presidential Election Prediction

For presidential election prediction simulation, as only **Q01-Voting Behavior** is reported in the main results, we provide the extended experiment using the full questionnaire to evaluate the LLM's behavior on election-related questions, as shown in Table 13. We simulate 5,454 agents and select superior LLMs based on this experiment to carry out the large-scale main experiment in main experiment.

	Overall	Voting-Subset
GPT-4o	**0.762**	**0.812**
GPT-4o-mini	0.754	0.803
Claude-3.5-Sonnet	0.737	0.775
Qwen2-7b-Instruct	0.675	0.764
Qwen2-72b-Instruct	0.748	0.784
Qwen2.5-72b-Instruct	0.750	0.804
Llama3-70b-Instruct	0.749	0.802

Table 13: Full questionnaire simulation results in presidential election prediction. Voting-Subset indicates questions containing specific party name, namely Q01, Q11, Q12, Q13, Q14, Q15, Q38.

E.2 Further Experiments in Breaking News Feedback

E.2.1 Breaking News Feedback Simulations on Other News

We simulate public opinion on the news titled "The Release of ChatGPT" in this paper, which belongs to the technology field. In addition to technology, we conducted three further news feedback simulations in the domains of politics (The New York Times, 2024a), energy (The New York Times, 2024c), and international affairs (The New York Times, 2024b) to verify the generalization ability in this simulation. We employed GPT-4o to answer two questions for each news item. The distribution sampling and simulation processes follow the same pipeline described in the paper, and the ground truths were directly scraped from comments under the news and annotated by GPT-4o.

Prompt 1: News Questionnaire

```
Question 1: Given the news, what's your
    sentiment towards the topic? First
    give your reason and then choose the
    answer.
Options: A. Positive  B. Neutral   C.
    Negative

Question 2: Given the news, what's your
    stance towards the topic? First give
    your reason and then choose the
    answer.
Options: A. Support  B. Neutral   C.
    Against
```

The results are shown in Table 14, which demonstrates that **the news feedback can be easily gen-**

eralized to different news feedback simulations and the results are unified and comparable if more LLMs are tested.

	Size	Sentiment	Stance
Politics	525	0.324	0.034
Energy	253	0.103	0.531
International Affairs	131	0.039	0.437

Table 14: Additional breaking news feedback simulation results, reported in KL-divergence.

E.2.2 Other Sources of Ground Truth for Attitudes toward ChatGPT

To verify whether the generated ground-truth set in the main experiment is trustworthy, we conduct an additional simulation based on the Pew Research dataset (Tyson and Kikuchi, 2023). The corresponding questions and ground truth responses can be found here (Pew Research Center, 2023). Since the original survey participants are primarily American adults, we adopt the U.S. population distribution as the prior and sample 11,201 agents in the simulation, matching the official dataset size. The performance of all models on each question, along with the average performance, is presented in Table 15, which is consistent with the main results.

	Llama3	Qwen2.5	DS-V3	4o	4o-mini
Avg	**1.28**	1.52	4.95	1.97	5.10
q_0	2.01	2.06	**2.00**	2.66	4.01
q_1	0.55	**0.39**	0.94	0.50	0.68
q_2	**0.27**	0.28	2.54	1.42	3.64
q_3	**1.34**	1.40	3.96	2.24	4.08
q_4	**2.85**	6.13	10.03	3.27	13.46
q_5	1.88	**1.72**	8.64	2.24	9.62
q_6	**1.56**	1.75	7.91	3.23	8.33
q_7	0.65	**0.56**	1.15	1.15	1.21
q_8	0.55	**0.23**	2.76	1.27	1.88
q_9	1.10	**0.64**	9.56	1.75	4.06

Table 15: Pew Research simulation on attitudes toward AI, reported in KL-divergence.

E.3 Experiment Setting for Wisdorm of Partisan Crowds

To verify the generalization ability in the multi-round interactive simulation scenario. We implement the wisdom of partisan crowds in our framework following the previous work (Chuang et al., 2023), where detailed task formulation and prompts can be found. Briefly speaking, the task aims to simulate the partisan group's response towards 8 political issues after 3 rounds of interaction within their party. The metrics used in §6.1 are detailed as follows:

- **Wisdom of Crowds ($\overline{\Delta\varepsilon}$)**: The normalized group error ε is calculated by averaging the normalized error among all 35 agents within a group for Democrats and Republicans, respectively. $\Delta\varepsilon$ is the difference of ε between $t = 3$ and $t = 1$. $\overline{\Delta\varepsilon}$ is then calculated by averaging $\Delta\varepsilon$ across all questions and both parties.
- **Partisan Bias ($\overline{\beta_{PB}}$)**: The normalized group mean η is calculated by averaging the normalized mean among all 35 agents within a group for Democrats and Republicans, respectively. β_{PB} is the difference between the average η of Republicans and the average η of Democrats for all time steps, and then times a coefficient $sign(h_q)$. $\overline{\beta_{PB}}$ is then calculated by averaging β_{PB} across all questions.
- **Human Likeness Index (HLI)**: HLI is calculated as $\overline{\beta_{PB}} - \overline{\Delta\varepsilon}$.
- **Extreme Values (Ext.%)**: The Ext.% metric evaluates the proportion of LLM agent responses that are unrealistic, based on established criteria (Becker et al., 2019).

The simulation includes 12 independent groups for each party, and the averaged metrics are reported in Table 4.

E.4 Supplementary Infromation for Figure 3

For improved clarity, the results of Figure 3 are reformulated in Table 16. We report the value of Real minus Simulated here, demonstrating that (1) all the models tend to behave consistently between real and simulated sets; (2) all the models perform more conservatively in the simulated results than the real results (most results are above 0).

Dimension	Llama3	Qwen2.5	DS-R1	4o-mini	4o
PublicCognition	1.12	0.38	0.58	0.40	0.40
PubilicRisks	-0.13	0.13	-0.27	0.18	-0.25
PerceivedBenefits	1.02	0.27	0.39	0.65	0.30
TRust	0.64	-0.14	0.12	0.27	0.02
FAirness	0.40	0.02	0.12	-0.26	0.16
PublicAcceptance	1.61	0.40	0.52	0.44	0.24

Table 16: Detailed performances of the breaking news feedback simulation, reported in (Real minus Simulated).

F Prompt Library

Prompt 2: Prompt for Presidential Election Prediction

```
It's 2024, and you're being surveyed for the 2024 American National Election Studies. You are
    a real person living in {state} with the following personal information. Please answer
    the following question as best as you can. You should act consistently with the role. Do
    not refuse to answer.

Some of your historical comments on social media platforms: {sampled historical comments
    before 2024.9}

Personal information: {personal demographics info}

Candidate Information: Donald Trump, the former President of the United States and a prominent
     figure in the Republican Party, is running for the 2024 Presidential Election. Known for
     his assertive communication style and stringent immigration policies, Trump has promised
     to implement even more restrictive measures against illegal immigration if re-elected.
    He is also advocating for comprehensive tariffs on foreign goods, aiming to protect
    American industries, although this could lead to increased consumer prices. Trump's
    campaign emphasizes economic nationalism and a return to his "America First" approach to
    governance. Kamala Harris, the current Vice President of the United States and a key
    member of the Democratic Party, is seeking the presidency in the upcoming election. As a
    former prosecutor and attorney general, Harris brings a strong background in law and
    justice to her campaign. She is focusing on issues such as reducing child poverty,
    supporting labor unions, ensuring affordable healthcare, and advocating for paid family
    leave. Harris is also a proponent of voting rights legislation, gun control measures, and
     reproductive rights. Her vice presidency has seen her engage with voting reforms,
    immigration policies, and efforts to protect and expand access to abortion services.

Question: {question}
Options: {options}

You should give your answer (you only need to answer the option letter number) in JSON format
    as example below:
```json
{"answer": "xxx"}
```
```

Prompt 3: Prompt for Breaking News Feedback

```
You are a {age}-year-old {gender}, with a {education} degree, living in {location}, and your
    level of consumption is {consumption}.
Here is some content you previously posted on social media:
{sampled_historical_post}

Here is a piece of news that has just been released:
{news}

You are now receiving a survey about {news_topic}. Please fill it out carefully.

Please rate the following statements based on your actual feelings, using a scale from A to E:
(A. Strongly Disagree; B. Somewhat Disagree; C. Neutral; D. Somewhat Agree; E. Strongly Agree)

{statements}

You should give your answer (you only need to answer the option letter number) in JSON format
    as example below:
```json
{"answer": "xxx"}
```
```

Prompt 4: Prompt for National Economic Survey

```
You are a {age}-year-old Chinese internet user with a monthly income of {income} yuan. Based
    on your historical posts, please complete a questionnaire on consumption behavior. The
    questionnaire includes 8 categories: food, tobacco and alcohol; clothing; housing;
    household goods; transportation and communication; education, culture, and entertainment;
     healthcare; and other goods and services.

Please allocate your expenditures across these categories reasonably according to your income
    and spending budget. Try not to exceed the budget. Answer the following question in the
    specified format.

Historical posts: "{sampled_historical_posts}"

Question: {question_title}
{choices}

**Formatting requirements**: Please output your answer in the following JSON format:

```json
{"answer": "#only_answer", "explanation": #one_sentence_only}
```

* The value of the `answer` key should be your answer to the question. For multiple choice
    questions, only include the letter option. For fill-in-the-blank questions, provide only
    the answer.
* The value of the `explanation` key should be a one-sentence brief explanation of why you
    chose this answer, given your identity, income, and historical posts.

(Below Only For DeepSeek-R1)
Important: Provide your answer with a simple thought process within 500 characters. Avoid
    overthinking or lengthy reasoning.
```

G Questionnaire Design Details

We provide the questionnaires here for all three simulations.

G.1 Questionnaire for Presidential Election Prediction

Q01	**Voting Behavior**
Question	ORDER OF MAJOR PARTY CANDIDATE NAMES
Value Labels	1. Democrat first / Republican second 2. Republican first / Democrat second

Q02	**Social Security**
Question	Next I am going to read you a list of federal programs. For each one, I would like you to tell me whether you would like to see spending increased, decreased, or kept the same. What about Social Security? Should federal spending on Social Security be increased, decreased, or kept the same?
Value Labels	-2. DK/RF 1. Increased 2. Decreased 3. Kept the same

Q03	**Education**
Question	What about public schools? Should federal spending on public schools be increased, decreased, or kept the same?
Value Labels	-2. DK/RF 1. Increased 2. Decreased 3. Kept the same

Q04	**Immigration**
Question	What about tightening border security to prevent illegal immigration? Should federal spending on tightening border security to prevent illegal immigration be increased, decreased, or kept the same?
Value Labels	-2. DK/RF 1. Increased 2. Decreased 3. Kept the same

Q05	**Criminal Justice**
Question	What about dealing with crime? Should federal spending on dealing with crime be increased, decreased, or kept the same?
Value Labels	-2. DK/RF 1. Increased 2. Decreased 3. Kept the same

Q06	**Social Welfare**
Question	What about welfare programs? Should federal spending on welfare programs be increased, decreased, or kept the same?

Value Labels	-2. DK/RF 1. Increased 2. Decreased 3. Kept the same

Q07	**Infrastructure**
Question	What about building and repairing highways? Should federal spending on building and repairing highways be increased, decreased, or kept the same?
Value Labels	-2. DK/RF 1. Increased 2. Decreased 3. Kept the same

Q08	**Aid to Poor**
Question	What about aid to the poor? Should federal spending on aid to the poor be increased, decreased, or kept the same?
Value Labels	-2. DK/RF 1. Increased 2. Decreased 3. Kept the same

Q09	**Environment**
Question	What about protecting the environment? Should federal spending on protecting the environment be increased, decreased, or kept the same?
Value Labels	-2. DK/RF 1. Increased 2. Decreased 3. Kept the same

Q10	**Government**
Question	How much do you feel that having elections makes the government pay attention to what the people think?
Value Labels	-2. DK/RF 1. A good deal 2. Some 3. Not much

Q11	**Economy**
Question	Which party do you think would do a better job of handling the nation's economy?
Value Labels	-2. DK/RF 1. Democrats would do a better job 2. Not much difference between them 3. Republicans would do a better job

Q12	**Health Care**
Question	Which party do you think would do a better job of handling health care?

Value Labels	-2. DK/RF 1. Democrats would do a better job 2. Not much difference between them 3. Republicans would do a better job

Q13	**Immigration**
Question	Which party do you think would do a better job of handling immigration?
Value Labels	-2. DK/RF 1. Democrats would do a better job 2. Not much difference between them 3. Republicans would do a better job

Q14	**Taxes**
Question	Which party do you think would do a better job of handling taxes?
Value Labels	-2. DK/RF 1. Democrats would do a better job 2. Not much difference between them 3. Republicans would do a better job

Q15	**Environment**
Question	Which party do you think would do a better job of handling the environment?
Value Labels	-2. DK/RF 1. Democrats would do a better job 2. Not much difference between them 3. Republicans would do a better job

Q16	**Education**
Question	Some people think the government should provide fewer services even in areas such as health and education in order to reduce spending. Other people feel it is important for the government to provide many more services even if it means an increase in spending. And, of course, some people have a neutral position. Which of the following best describes your view?
Value Labels	-2. DK/RF 1. Government should provide fewer services 2. Neutral 3. Government should provide more services

Q17	**Defense**
Question	Some people believe that we should spend less money for defense. Others feel that defense spending should be increased. And, of course, some people have a neutral position. Which of the following best describes your view?
Value Labels	-2. DK/RF 1. Decrease defense spending 2. Neutral 3. Increase defense spending

Q18	**Health Care**

Question	There is much concern about the rapid rise in medical and hospital costs. Some people feel there should be a government insurance plan which would cover all medical and hospital expenses for everyone. Others feel that all medical expenses should be paid by individuals through private insurance plans like Blue Cross or other company paid plans. And, of course, some people have a neutral position. Which of the following best describes your view?
Value Labels	-2. DK/RF 1. Government insurance plan 2. Neutral 3. Private insurance plan

Q19	**Social Welfare**
Question	Some people feel the government in Washington should see to it that every person has a job and a good standard of living. Others think the government should just let each person get ahead on their own. And, of course, some people have a neutral position. Which of the following best describes your view?
Value Labels	-2. DK/RF 1. Government should see to jobs and standard of living 2. Neutral 3. Government should let each person get ahead on own

Q20	**Aid to Blacks**
Question	Some people feel that the government in Washington should make every effort to improve the social and economic position of blacks. Others feel that the government should not make any special effort to help blacks because they should help themselves. And, of course, some people have a neutral position. Which of the following best describes your view?
Value Labels	-2. DK/RF 1. Government should help blacks 2. Neutral 3. Blacks should help themselves

Q21	**Environment**
Question	Some people think we need much tougher government regulations on business in order to protect the environment. Others think that current regulations to protect the environment are already too much of a burden on business. And, of course, some people have a neutral position. Which of the following best describes your view?
Value Labels	-2. DK/RF 1. Tougher regulations on business needed to protect environment 2. Neutral 3. Regulations to protect environment already too much a burden on business

Q22	**Abortion**
Question	Would you be pleased, upset, or neither pleased nor upset if the Supreme Court reduced abortion rights?

Value Labels	-2. DK/RF 1. Pleased 2. Upset 3. Neither pleased nor upset

Q23	**Criminal Justice**
Question	Do you favor or oppose the death penalty for persons convicted of murder?
Value Labels	-2. DK/RF 1. Favor 2. Oppose

Q24	**US Position in World**
Question	Do you agree or disagree with this statement: 'This country would be better off if we just stayed home and did not concern ourselves with problems in other parts of the world.'
Value Labels	-2. DK/RF 1. Agree 2. Disagree

Q25	**US Position in World**
Question	How willing should the United States be to use military force to solve international problems?
Value Labels	-2. DK/RF 1. Willing 2. Moderately willing 3. Not willing

Q26	**Inequality**
Question	Do you think the difference in incomes between rich people and poor people in the United States today is larger, smaller, or about the same as it was 20 years ago?
Value Labels	-2. DK/RF 1. Larger 2. Smaller 3. About the same

Q27	**Environment**
Question	Do you think the federal government should be doing more about rising temperatures, should be doing less, or is it currently doing the right amount?
Value Labels	-2. DK/RF 1. Should be doing more 2. Should be doing less 3. Is currently doing the right amount

Q28	**Parental Leave**
Question	Do you favor, oppose, or neither favor nor oppose requiring employers to offer paid leave to parents of new children?

Value Labels	-2. DK/RF 1. Favor 2. Oppose 3. Neither favor nor oppose

Q29	**LGBTQ+ Rights**
Question	Do you think business owners who provide wedding-related services should be allowed to refuse services to same-sex couples if same-sex marriage violates their religious beliefs, or do you think business owners should be required to provide services regardless of a couple's sexual orientation?
Value Labels	-2. DK/RF 1. Should be allowed to refuse 2. Should be required to provide services

Q30	**LGBTQ+ Rights**
Question	Should transgender people - that is, people who identify themselves as the sex or gender different from the one they were born as - have to use the bathrooms of the gender they were born as, or should they be allowed to use the bathrooms of their identified gender?
Value Labels	-2. DK/RF 1. Have to use the bathrooms of the gender they were born as 2. Be allowed to use the bathrooms of their identified gender

Q31	**LGBTQ+ Rights**
Question	Do you favor or oppose laws to protect gays and lesbians against job discrimination?
Value Labels	-2. DK/RF 1. Favor 2. Oppose

Q32	**LGBTQ+ Rights**
Question	Do you think gay or lesbian couples should be legally permitted to adopt children?
Value Labels	-2. DK/RF 1. Yes 2. No

Q33	**LGBTQ+ Rights**
Question	Which comes closest to your view? You can just tell me the number of your choice.
Value Labels	-2. DK/RF 1. Gay and lesbian couples should be allowed to legally marry 2. Gay and lesbian couples should be allowed to form civil unions but not legally marry 3. There should be no legal recognition of gay or lesbian couples' relationship

Q34	**Immigration**
Question	Some people have proposed that the U.S. Constitution should be changed so that the children of unauthorized immigrants do not automatically get citizenship if they are born in this country. Do you favor, oppose, or neither favor nor oppose this proposal?

Value Labels	-2. DK/RF 1. Favor 2. Oppose 3. Neither favor nor oppose

Q35	**Immigration**
Question	What should happen to immigrants who were brought to the U.S. illegally as children and have lived here for at least 10 years and graduated high school here? Should they be sent back where they came from, or should they be allowed to live and work in the United States?
Value Labels	-2. DK/RF 1. Should be sent back where they came from 2. Should be allowed to live and work in the US

Q36	**Immigration**
Question	Do you favor, oppose, or neither favor nor oppose building a wall on the U.S. border with Mexico?
Value Labels	-2. DK/RF 1. Favor 2. Oppose 3. Neither favor nor oppose

Q37	**Unrest**
Question	During the past few months, would you say that most of the actions taken by protestors to get the things they want have been violent, or have most of these actions by protesters been peaceful, or have these actions been equally violent and peaceful?
Value Labels	-2. DK/RF 1. Mostly violent 2. Mostly peaceful 3. Equally violent and peaceful

Q38	**Government**
Question	Do you think it is better when one party controls both the presidency and Congress, better when control is split between the Democrats and Republicans, or doesn't it matter?
Value Labels	-2. DK/RF 1. Better when one party controls both 2. Better when control is split 3. It doesn't matter

Q39	**Government**
Question	Would you say the government is pretty much run by a few big interests looking out for themselves or that it is run for the benefit of all the people?
Value Labels	-2. DK/RF 1. Run by a few big interests 2. For the benefit of all the people

Q40	**Government**
Question	Do you think that people in government waste a lot of the money we pay in taxes, waste some of it, or don't waste very much of it?

Value Labels	-2. DK/RF 1. Waste a lot 2. Waste some 3. Don't waste very much

Q41	**Election Integrity**
Question	Do you favor, oppose, or neither favor nor oppose allowing convicted felons to vote once they complete their sentence?
Value Labels	-2. DK/RF 1. Favor 2. Oppose 3. Neither favor nor oppose

Q42	**Democratic Norms**
Question	How important is it that news organizations are free to criticize political leaders?
Value Labels	-2. DK/RF 1. Not important 2. Moderately important 3. Important

Q43	**Democratic Norms**
Question	How important is it that the executive, legislative, and judicial branches of government keep one another from having too much power?
Value Labels	-2. DK/RF 1. Not important 2. Moderately important 3. Important

Q44	**Democratic Norms**
Question	How important is it that elected officials face serious consequences if they engage in misconduct?
Value Labels	-2. DK/RF 1. Not important 2. Moderately important 3. Important

Q45	**Democratic Norms**
Question	How important is it that people agree on basic facts even if they disagree politically?
Value Labels	-2. DK/RF 1. Not important 2. Moderately important 3. Important

Q46	**Democratic Norms**
Question	Would it be helpful, harmful, or neither helpful nor harmful if U.S. presidents could work on the country's problems without paying attention to what Congress and the courts say?

Value Labels	-2. DK/RF 1. Helpful 2. Harmful 3. Neither helpful nor harmful

Q47	**Democratic Norms**
Question	Do you favor, oppose, or neither favor nor oppose elected officials restricting journalists' access to information about government decision-making?
Value Labels	-2. DK/RF 1. Favor 2. Oppose 3. Neither favor nor oppose

Q48	**Gender Resentment**
Question	'Many women interpret innocent remarks or acts as being sexist.' Do you agree, neither agree nor disagree, or disagree with this statement?
Value Labels	-2. DK/RF/technical error 1. Agree 2. Neither agree nor disagree 3. Disagree

Q49	**Gender Resentment**
Question	'Women seek to gain power by getting control over men.' Do you agree, neither agree nor disagree, or disagree with this statement?
Value Labels	-2. DK/RF/technical error 1. Agree 2. Neither agree nor disagree 3. Disagree

G.2 Questionnaire for Breaking News Feedback

Q01	**Public Cognition (PC)**
Question	I have heard of ChatGPT.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree

Q02	**Public Cognition (PC)**
Question	Many people around me use ChatGPT.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree

Q03	**Public Cognition (PC)**
Question	I have a deep understanding of ChatGPT's functions and applications.

Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q04	**Perceived Risks (PR)**
Question	ChatGPT may lead to the widespread dissemination of false information.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q05	**Perceived Risks (PR)**
Question	ChatGPT may reduce human thinking ability and creativity.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q06	**Perceived Risks (PR)**
Question	The development of ChatGPT may replace certain jobs, and I am deeply concerned about this.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q07	**Perceived Benefits (PB)**
Question	ChatGPT will definitely improve my work and study efficiency.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q08	**Perceived Benefits (PB)**
Question	ChatGPT helps broaden my knowledge and provides me with new perspectives and ideas.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q09	**Perceived Benefits (PB)**
Question	ChatGPT promotes technological innovation and development in related fields.

Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q10	**Trust (TR)**
Question	I fully trust the team developing ChatGPT to manage and guide its development responsibly.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q11	**Trust (TR)**
Question	I have strong confidence in the accuracy and reliability of the information generated by ChatGPT.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q12	**Trust (TR)**
Question	I believe that the future application of ChatGPT will be effectively regulated.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q13	**Fairness (FA)**
Question	The opportunities to use ChatGPT are distributed fairly among different groups of people.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q14	**Fairness (FA)**
Question	I find the distribution of benefits brought by ChatGPT to be fair.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree
Q15	**Fairness (FA)**

Question	I believe that the decision-making process for the development and promotion of ChatGPT is fully transparent and adequately reflects public interests.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree

Q16	**Public Acceptance (PA)**
Question	Overall, I strongly welcome the emergence of ChatGPT.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree

Q17	**Public Acceptance (PA)**
Question	I am definitely willing to use ChatGPT in my work or studies.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree

Q18	**Public Acceptance (PA)**
Question	I strongly support increased investment in the research and development of AI technologies like ChatGPT.
Value Labels	1. Disagree 2. Partially disagree 3. Neutral 4. Partially agree 5. Agree

G.3 Questionnaire for National Economic Survey

Q01	**Food**
Question	What is your average monthly expenditure on food (including dining out)? (Unit: CNY)
Value Labels	A. Below 500 CNY B. 501-650 CNY C. 651-800 CNY D. 801-1000 CNY E. Above 1000 CNY

Q02	**Food**
Question	Do you think your current spending on food, tobacco, and alcohol is too high relative to your income?

Value Labels	A. Yes B. No C. Acceptable
Q03	**Clothing**
Question	What is your average monthly expenditure on clothing (including apparel, shoes, and accessories)? (Unit: CNY)
Value Labels	A. Below 50 CNY B. 51-100 CNY C. 101-150 CNY D. 151-200 CNY E. Above 200 CNY
Q04	**Clothing**
Question	How much economic pressure do you feel from clothing expenses?
Value Labels	A. Very low, almost no pressure B. Moderate, some pressure but manageable C. High, requires careful spending D. Very high, affects spending in other areas
Q05	**Household**
Question	What is your average monthly housing expenditure? (Including rent, mortgage, property fees, maintenance, etc.) (Unit: CNY)
Value Labels	A. Below 200 CNY B. 201-500 CNY C. 501-800 CNY D. 801-1200 CNY E. Above 1200 CNY
Q06	**Household**
Question	What percentage of your monthly income is spent on housing? (Including rent, mortgage, property fees, maintenance, etc.)
Value Labels	A. Below 10% B. 10%-20% C. 21%-30% D. 31%-40% E. Above 40%
Q07	**Daily Service**
Question	What is your average monthly expenditure on daily necessities (personal care, household items, cleaning supplies, etc.) and services (housekeeping, repairs, beauty, pet services, etc.)? (Unit: CNY)
Value Labels	A. Below 80 CNY B. 81-120 CNY C. 121-160 CNY D. 161-200 CNY E. Above 200 CNY
Q08	**Transportation & Communication**

Question	What is your average monthly expenditure on transportation (public transport, taxis, fuel, parking, etc.) and communication (mobile and internet fees)? (Unit: CNY)
Value Labels	A. Below 200 CNY B. 201-300 CNY C. 301-400 CNY D. 401-500 CNY E. Above 500 CNY
Q09	**Education & Entertainment**
Question	What is your average monthly expenditure on education (tuition, training, books, etc.) and cultural entertainment (movies, performances, games, fitness, cultural activities, etc.)? (Unit: CNY)
Value Labels	A. Below 100 CNY B. 101-200 CNY C. 201-300 CNY D. 301-400 CNY E. Above 400 CNY
Q10	**Education & Entertainment**
Question	Can you easily afford your current education, cultural, and entertainment expenses?
Value Labels	A. Yes, spending does not affect other areas B. Barely, needs some control C. Not really, affects other expenditures D. No, it creates significant financial pressure
Q11	**Medical**
Question	What is your average monthly expenditure on healthcare (medications, medical services, health management, etc.)? (Unit: CNY)
Value Labels	A. Below 100 CNY B. 101-200 CNY C. 201-300 CNY D. 301-400 CNY E. Above 400 CNY
Q12	**Medical**
Question	Have you purchased private medical or health insurance for yourself or your family?
Value Labels	A. Yes B. Not yet, but planning to C. No, and no plans to
Q13	**Others**
Question	Besides food, clothing, housing, daily necessities and services, transportation, education, culture, and healthcare, what is your average monthly expenditure on other areas (e.g., hobbies, charitable donations, investment, etc.)? (Unit: CNY)

Value Labels	A. Below 30 CNY B. 31-60 CNY C. 61-90 CNY D. 91-120 CNY E. Above 120 CNY

Q14	**Overall**
Question	How would you evaluate the impact of your current consumption level on your household (or personal) financial situation?
Value Labels	A. Comfortable, can moderately increase spending B. Average, can maintain current spending C. Tight, need to control or reduce spending D. Very tight, affects quality of life

Q15	**Overall**
Question	Do you feel that your consumption pressure is too high relative to your income level?
Value Labels	A. Yes B. No C. Not sure

Q16	**Overall**
Question	If your income increases, which consumption areas would you most like to expand or improve? (Multiple choices allowed)
Value Labels	A. Food and alcohol B. Clothing C. Housing D. Daily necessities and services E. Transportation and communication F. Education, culture, and entertainment G. Healthcare H. Other goods and services

Q17	**Overall**
Question	What is your consumption expectation for the next six months to a year?
Value Labels	A. Will continue to increase B. Will remain roughly the same C. Will moderately decrease D. Uncertain