Simulate Anything: A Generalized Social Simulation Framework Driven by LLM Agents Based on a Large-Scale Real-World User Pool

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 the scenarios and improve the simulation accuracy. However, they are faced with limitations such as rigid frameworks, small-scale simulations, and narrow evaluation criteria. To this end, we introduce ***Simulate Anything***: a generalized social simulation framework driven by LLM agents, which is composed of a 10-million-user real-world pool, a demographic distribution sampling strategy, and a unified simulation evaluation method. We evaluate the framework by conducting massive simulations under political, journalistic, and economic scenarios. The results prove that our framework can support diverse and trustworthy massive social simulations with a standard pipeline and minimal changes. *Upon acceptance, we will release all three simulations with the corresponding user pool.*

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 mathematics functions to simulate the actions

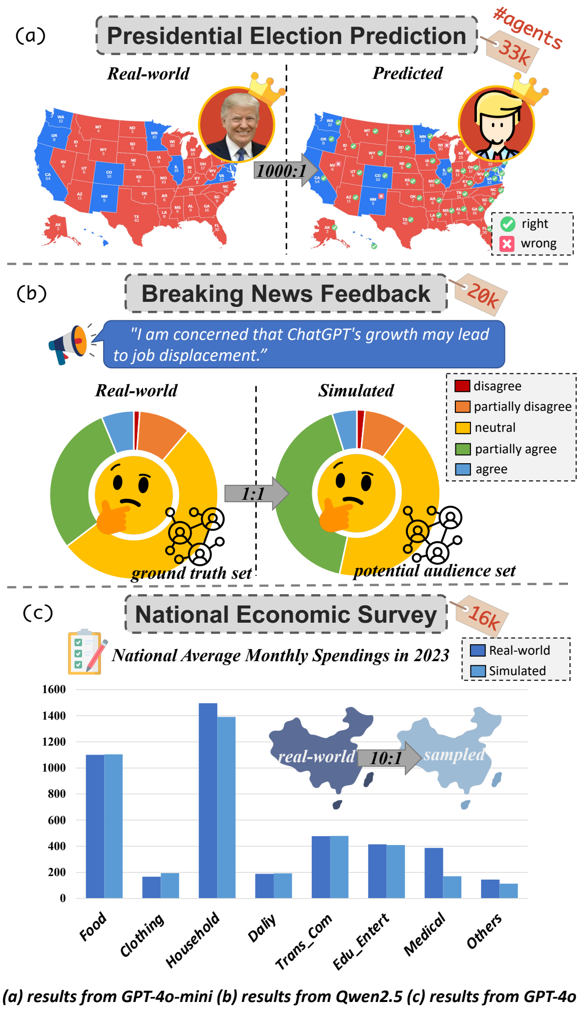

Figure 1: An illustration of the simulation results following the ***Simulate Anything*** framework in (a) presidential election prediction, (b) breaking news feedback, and (c) national economic survey scenarios.

of individuals (Tang, 2024), and then scales up these actions to forecast the collective result. With the rise of agent-based simulations powered by 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., 2024a), 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 ***Simulate Anything*** 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 Simulate Anything 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 Simulate Anything framework is of great help in constructing a standard and accurate massive social simulation. To conclude, contributions in this paper are as follows:

- **Simulate Anything**: a generalized social simulation framework driven by LLM agents based on large-scale real-world user pool, which allows for diverse simulating scenarios with high confidence by aligning 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 combing 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 Simulate Anything framework.

2 Related Works

2.1 Social Simulation Research

Traditional social simulation methods mainly rely on opinion polls, expert judgment, and statistical models (Erikson and Wlezien, 2014; Burnap et al., 2016; Bohannon, 2017). The ABM method provides a more objective and accurate

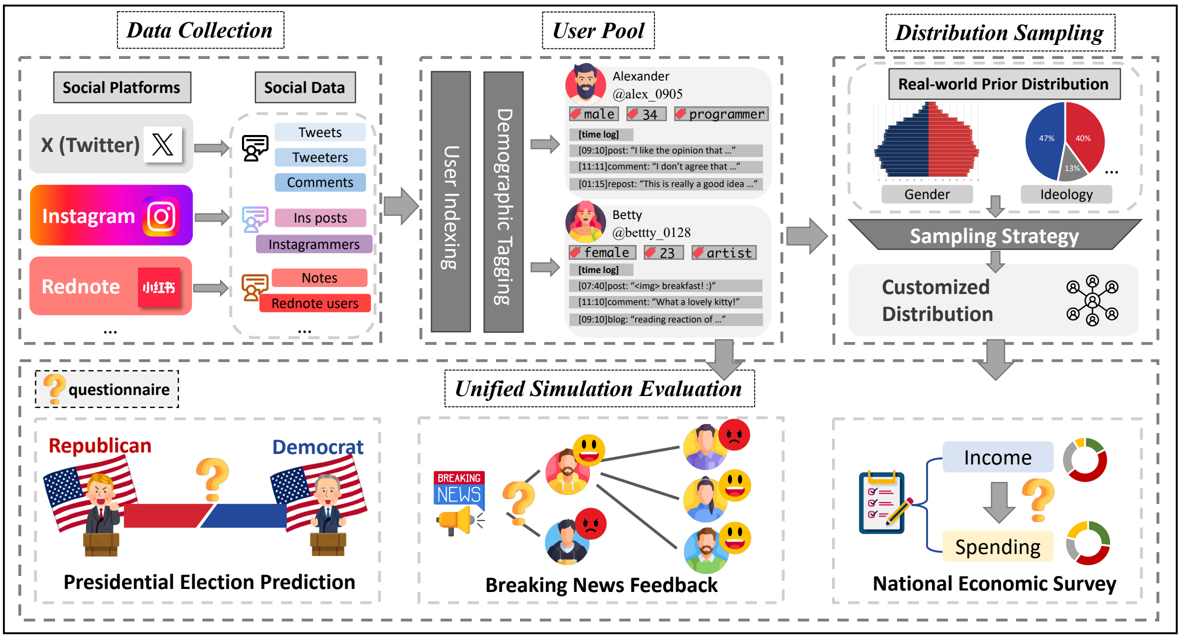

Figure 2: An illustration of ***Simulate Anything*** framework. We collect raw data from multiple social media platforms and then construct a 10M-size user pool through user indexing and annotation. For a customized simulation, a prior distribution is calculated first to sample target agents, then an adaptive simulation is conducted to generate the simulation result.

prediction method by simulating individual behavior, combining micro-individual characteristics and macro-socioeconomic factors (Qiu and Phang, 2020; Sobkowicz, 2016). In recent years, with the rapid development of LLM, researchers have discovered its potential to solve problems in the field of social science (Linegar et al., 2023; Gujral et al., 2024). Preliminary research has shown positive outcomes in domains including electoral prediction, policy evaluation, and the simulation of public sentiment (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 possibility that may shed light on solving general problems paradigm (Xi et al., 2023; Guo et al., 2024; Gao et al., 2024). While individual-level simulation (also known as role-playing agents) focuses on highly reliable and reproducible human-like behavior (Shao et al., 2023; Xie et al., 2024; Sun et al., 2024), 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). Task-level simulations also vary depending on different scenarios, wherein general-purpose scenarios highlight the intelligence within LLMs (Park et al., 2023; Yue et al., 2024; Mou et al., 2024b) while specific-domain scenarios emphasize the combination between workflows and domain specialization, like journalism (Liu et al., 2024; Li et al., 2024b), economy (Horton, 2023; Zhao et al., 2023), social media (Cai et al., 2024; Lyu et al., 2024), etc.

3 Simulate Anything

3.1 Overall Framework

The ***Simulate Anything*** framework follows a structured pipeline to achieve realistic social simulation results, 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 real-world demographic distributions. (4) Various large-scale social simulations are conducted using an adaptive simulation method. (5) The results closely resemble the real-world trends.

3.2 Data Collection

Data Source The data source comprises diverse social media platforms, including X[1], Insta-

[1] https://x.com/

Source	Format	# Users	# Posts
X	pure text	1,006,517	30,195,510
Instagram	text-image pair	33,935	10,180,500
Rednote	text-image pair	9,158,404	40,963,735

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

gram[2] (Kim et al., 2020), Rednote[3]. The diversity allows our user pool to encompass a broad distribution of user groups across different languages, cultures, and religions. Considering the structural differences and user preferences among these social media platforms, we collect only posts—such as tweets, weblogs, and notes—along with engagement data, including the number of likes, comments, and reposts. These posts provide rich textual and multimodal information from users. All collected data are *publicly available*, and our approach strictly adheres to platform privacy policies.

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 Constrution

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: $UserPool = \{U_i, P_i \mid i \in \mathbb{S}\}$, 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 demographic information is not directly accessible, we design a demographics annotation system to infer and tag user attributes. The process begins with multiple LLMs serving as initial annotators, classifying users across various demographic dimensions. Human annotators then evaluate and refine the LLM-generated labels, ensuring the reliability of the user tags dataset. The curated dataset is subsequently used to train demographic classifiers, enabling large-scale annotation in a cost-effective manner. 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.

[2] https://www.instagram.com/
[3] https://www.xiaohongshu.com/

3.4 Distribution Sampling Strategy

By constructing the 10M-size user pool, we enable the customization of any group distribution for specific social simulations. The large scale and diversity of the user pool ensure flexible sampling strategies. The sampling strategy can be formulated as $D_S = Sampler(UserPool, D_P(i))$, where D_S and $D_P(i)$ denote the sampled pool and prior distribution for the $i-th$ task, respectively. For simulation scenarios where only marginal demographic distributions (e.g., census data) are available, we apply iterative proportional fitting (IPF) to estimate the joint distribution from these marginals (Choupani and Mamdoohi, 2016). When the joint distribution is already known (e.g., online group distributions), identical distribution sampling (IDS) can be directly applied. Implementation details are specified 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 in collaboration with domain experts. For simulations that result in a discrete label space, such as representative election and attitudes simulation, the labels can be transformed into options in the questionnaire naturally. For simulations involving continuous results, such as financial and statistical events, the options are formulated into a range of numerical intervals. During the simulation, each agent is initialized with the corresponding user's demographic profile and historical posts from the user pool. The agent then responds to the questionnaire following predefined instructions. The answers are converted into quantitative metrics, which are compared against real-world data or computed ground truth for evaluation. We will detail the questionnaire in the following section.

4 Scenario Formulations

In this section, three massive social simulation scenarios are introduced following the ***Simulate Anything*** framework: 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 votes receives all the state's electoral votes. We predict the state-level election results in this task.

Prior Distribution Existing studies have conducted extensive research on the influence of demographics on elections (Major et al., 2018; Teixeira, 2009), which is considered as a significant role in U.S. elections. To accurately reflect the demographic and ideological makeup of U.S. citizens, we combine multiple datasets to construct the population distribution in our study. Specifically, 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). Demographics including *age*, *gender*, *race*, *income*, *education*, *area*, *region*, *employment*, *marital*, *religious*, *party*, and *ideology* are considered to construct the overall prior distribution, and iterative proportional fitting sampling is employed to sample target agents from the user pool as only marginal distributions are available.

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 elector options and optimize them into proper forms for LLM-based agents. The whole questionnaire can be found in Appendix D.1.

Comparision Metric Two metrics are used to comprehensively compare the simulated election results to the real-world results. (1) Accuracy rate (Acc) is measured by calculating the proportion of states for which the election simulation results align with the actual result, serving as a coarse-grained evaluation metric. (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, framing narratives, and influencing discourse through media coverage (van Dalen, 2024; Gómez-Calderón and Ceballos, 2024). Online social media platforms, as an emerging information consumption medium, 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 accuracy and foreseeability of public attitudes.

Prior Distribution We take all the users on the rednote in our user pool as the universal set and collect the users interested in the technology area as the **potential audience set** $\mathbb{P}$. We take the users who have talked about ChatGPT directly on the platform as the **ground truth set** $\mathbb{G}$ through keyword matching. It can be formulated that $\mathbb{G} \subset \mathbb{P} \subset UserPool$. During the simulation, the user content is cut off to the time before the news is released to prevent information leakage. The distribution of the potential audience set is regarded as the prior distribution and then we sample identical distribution agents from the user pool to carry out the simulation, i.e. $D_s = IDS(UserPool, \mathbb{P})$. During sampling, demographics like *gender*, *age*, *education*, and *consumption* are considered to contribute to the prior distribution.

Questionaire Design We design the public cognitive questionnaire based on the theory of the ABC model of attitude (Liu et al., 2021), which conceptualizes attitude as a combination of Affect, Behavior, and Cognition. It explains how attitudes form through a hierarchical process, where cognition influences emotions, which in turn shape behavior. 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

Scenario	# Agents	# Demographics	Type of demo	Sampling	Source	Language	# Questions	Ground truth
PresiElePred	33,182	12	label	IPF	X	EN	49	real world
BreakNewsFeed	20,000	7	label	IDS	rednote	ZH	18	calculated
NationEcoSur	16,000	9	label+number	IDS	rednote	ZH	17	real world

Table 2: Detail settings of three simulation scenarios, where PersiElePred, BreakNewsFeed, and NationEcoSur stand for three simulations mentioned in the paper respectively. # stands for *number of*.

public acceptance (PA). The whole questionnaire can be found in Appendix D.2.

Comparison Metric Agents from both the ground truth set and the potential audience set are required to answer the questionnaire to get pairwised answers. Then two evaluation dimensions are employed for breaking news feedback. (1) Normalized RMSE (NRMSE) is measured by calculating the answer points between simulated answers and ground truth answers in PC, PR, PB, TR, FA, and PA, serving as the value evaluation. (2) KL-divergence (KL-Div) is measured by taking the 6-dimension answer list as a distribution and calculating between the simulated distribution and ground truth distribution, serving as the distribution evaluation of the consistency.

4.3 National Economic Survey

Task Formulation Economic simulation is another 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). By integrating economic factors with social interactions, it helps predict systemic outcomes, guiding decision-making in areas such as governance, urban planning, and crisis management. We conduct a national economic survey by interviewing Chinese citizens on their monthly spending given the average salary of each province in China.

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). We sample nationwide agents from our user pool proportionally according to their *region* population and generate their *income* distribution according to the regional average income (NBS China, 2023a). 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 examples and uses segmented interval options in each question. The whole questionnaire can be referred to in Appendix D.3.

Comparison Metric Both value evaluation and distribution evaluation are involved in the national economic survey as well. (1) NRMSE of the nine categories is measured between the simulated results and official statistics. (2) KL-Div is measured by taking the 8-item spending as a distribution to evaluate the consistency between the simulation and the real world.

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), and DeepSeek-R1-671b (Guo et al., 2025). For commercial models, we select GPT-4o[4] (OpenAI, 2024b) and GPT-4o-mini[5] (OpenAI, 2024a).

Implementation Details We compare the settings of all three scenarios for better understanding, which is shown in Table 2. As the Presidential Election Prediction covers a 1/1000 sample of the U.S. citizen population, GPT-4o is not compared due to the cost restriction.

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 for all models to enable chain-of-thoughts during the generation and the temperature is set to 0.7 to encour-

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

Model	PresiElePred				BreakNewsFeed		NationEcoSur			
	Overall		*Battleground*				*Overall*		*1st-Region*	
	Acc↑	RMSE	Acc↑	RMSE	KL-Div	NRMSE	KL-Div	NRMSE	KL-Div	NRMSE
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
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 *1st-Region* indicates top-10 developed regions in China in terms of GDP.

age diversity. Implementation details for user pool construction and demographics annotation can be found in Appendix A and B.

5.2 Overall Results

The overall simulation results of the three scenarios are shown in Table 3. We report subset results for presidential election prediction and national economic survey additionally.

Presidential Election Prediction We report the overall results and the battleground states' results separately. The battleground states are complicated even in the real world and thus become the focus during the election process. According to the results, GPT-4o-mini and Qwen2.5-72b show competitive performance both in Acc and RMSE. Typically, according to the winner-takes-all rule, **all 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.

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 NRMSE respectively, and the following detailed analysis in §6.2 will reveal that **all public trends and opinions are consistently predicted**.

National Economic Survey We report the overall results and results of the top-10 developed regions in GDP (i.e. 1st-Region) separately. Generally, all the models perform closely to 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 1st-Region subset than overall. The results demonstrate that **individuals' spending habits can be accurately revived under the *Simulate Anything* framework, especially in developed regions**.

The overall results from both value evaluation and distribution evaluation of three simulations sufficiently prove that ***Simulate Anything*** can support diverse and accurate massive social simulations with a standard pipeline and minimal changes with human experts in the loop. However, the powered engine LLMs can impact the precision to some degree under different scenarios, which deserves further research.

6 Further Analysis

6.1 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 4, prior demographics distribution significantly improves the accuracy of the simulation in both Acc and RMSE compared to random demographics distribution. Additionally, past posts from users on social media platforms improve the fine-grained performance, especially for Llama3-70b in Acc and all the models in RMSE. We can tell from the ablation study that **both prior distribution and real-world knowledge in the *Simulate Anything* pipeline are significant during the simulation**.

6.2 Can Group Preference and Perspectives Be Well Reflected?

During the Breaking News Feedback simulation, the core concern is whether the preferences and perspectives of the target group are well captured and reflected in the results. We reformulate the

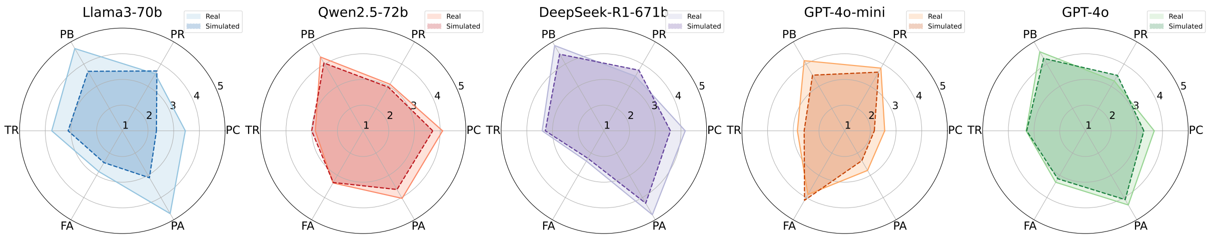

Figure 3: An illustration of the performances of the breaking news feedback simulation, where PC, PR, PB, TR, FA, and PA denote six dimensions from the Likert scale (see §4.2), with 1-point standing for totally disagree and 5-point for totally agree.

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

Table 4: Ablation experiment results on the presidential election prediction simulation, where -w/o Knowledge denotes *without real-world user knowledge* and -w/o Know & PriorDist denotes *without user knowledge and using random demographics distribution*.

Item	Llama3	Qwen2.5	4omini	4o	R1
Food	0.037	0.031	0.031	0.040	0.032
Clothing	0.012	0.015	0.019	0.015	0.015
Housing	0.052	0.110	0.107	0.120	0.102
Daily	**0.007**	**0.009***	**0.006**	**0.010***	**0.009**
Trans_Com	0.016	0.020	0.027	0.023	0.017
Edu_Entert	0.018	0.022	0.024	0.017	0.022
Healthcare	0.023	0.062	0.041	0.057	0.060
Others	**0.008***	**0.008**	**0.010***	**0.005**	**0.009**

Table 5: Detailed results on the national economic survey simulation reported in NRMSE, where the Item column indicates the components of spending (see §4.3). The best results are **bolded**, the second best results are bolded with a **star***, and the worst are underlined.

original questionnaire into the Likert 6-dimension scale ranging from 1 to 5 points, representing from totally disagree to totally agree. As the ground truth of the simulation is calculated by prompting LLM agents from the ground truth set, the *simulated* and *real* results are paired for each model, as shown in Figure 3. **All the models powered by potential audience set during the simulation tend to behave consistently with the ground truth users.** However, Llama3-70b perform poorly with a larger gap between the *simuated* and *real* results than other models. GPT-4o-mini shows different attitudes in the fairness (FA), and public acceptance (PA) dimensions, which may be because the news is related to OpenAI. Another trend indicates that, generally, **all the models perform more disagreeably in the *simulated* results than the *real* results**, which also underlines the potential risk of biases during the public opinion simulation.

6.3 In Which Domain Do LLMs Predict Better/Worse?

The simulation of the national economic survey covers 8 spending dimensions, as mentioned in 4.3. The overall results in Table 3 show the average performance of these dimensions, while model performances among these dimensions can also vary. We calculate the averaged NRMSE of 31 regions on each spending level, as shown in Table 5. It is worth mentioning that all the models show high consistency. Eliminating the *others* item, **all the models perform best on *daily necessities* spending planning and worst on *housing* spending**, which can reveal the LLM's preference on the economic decision-making and highlight the challenge in *housing* spending strategy.

7 Conclusion

In this paper, we introduce the ***Simulate Anything*** framework for massive social simulations using LLM agents. Our framework is featured with 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

Simulate Anything 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

During the collection of raw data, we strictly obey the privacy policy of each social media platform. Only post-related information is collected and the user IDs will be masked before the release of the user pool. Additionally, all the scenarios mentioned in the paper are only for research purposes and do not constitute any guidance for policy-making or political activities.

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.

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.

Jinyu Cai, Jialong Li, Mingyue Zhang, Munan Li, Chen-Shu Wang, and Kenji Tei. 2024. Language evolution for evading social media regulation via llm-based multi-agent simulation. *arXiv preprint arXiv:2405.02858*.

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*.

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*.

Robert S Erikson and Christopher Wlezien. 2014. Forecasting us presidential elections using economic and noneconomic fundamentals. *PS: Political Science & Politics*, 47(2):313–316.

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.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

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*.

Seungbae Kim, Jyun-Yu Jiang, Masaki Nakada, Jinyoung Han, and Wei Wang. 2020. Multimodal post attentive profiling for influencer marketing. In *Proceedings of The Web Conference 2020*, pages 2878–2884.

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. 2024a. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Xinyi Li, Yu Xu, Yongfeng Zhang, and Edward C Malthouse. 2024b. Large language model-driven multi-agent simulation for news diffusion under different network structures. *arXiv preprint arXiv:2410.13909*.

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.

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.

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.

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 6.

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

Table 6: 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 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[6], Claude3.5-Sonnet[7], and Gemini-1.5[8] are employed. For users 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 7.

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 7: Human annotators verification results. We report the consistency between humans and different LLMs.

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-

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

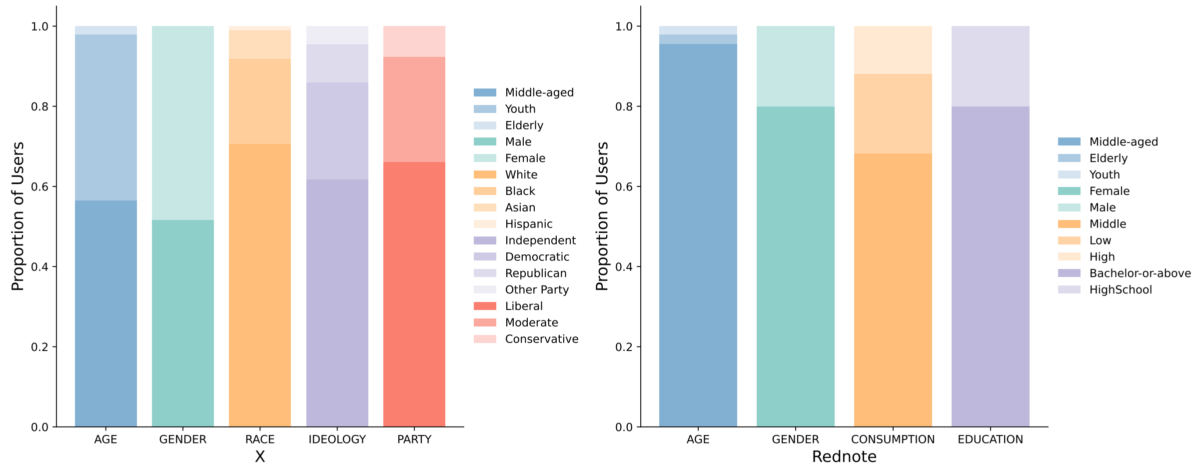

Figure 4: Demographic distribution on X and Rednote user pool.

chinese (Devlin et al., 2019) for Rednote. The implementation details are shown in Table 8.

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 8: Implementation details for demographic classifiers.

We report the performances of demographic classifiers on each demographic factor in Table 9.

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 9: 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 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 $\alpha > 1$:

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

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

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:

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

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 Questionnaire Design Details

We provide the questionnaires here for all three simulations.

D.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

D.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

D.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