

Article

Multi-Agent Social Simulation: Protocolizing LLM-Driven Agent-Based Modeling as a Quantitative Research Method [†]

Xiaoli Hu ^{1,*} and Yang Shen ^{1,2}

¹ School of Journalism and Communication, Tsinghua University, Beijing 100084, China; yshen1974@mail.tsinghua.edu.cn

² College of AI, Tsinghua University, Beijing 100084, China

* Correspondence: huxiaoli@mail.tsinghua.edu.cn

[†] This article is a revised and expanded version of the paper entitled “Multi-Agent Social Simulation: An Experimental Framework for Language-Native Social Experiments,” which was presented at the Open Conference of AI Agents for Science 2025, Virtual Conference, 22 October 2025.

Abstract

Social and behavioral research often needs to examine policy shocks, information interventions, platform-mediated attention, and governance feedback, but direct experiments on real populations are constrained by ethical risks, intervention costs, and limited repeatability. This study proposes Multi-Agent Social Simulation (MASS), a protocolized form of large language model-driven agent-based modeling (LLM-driven ABM) designed as a low-risk, repeatable, and auditable pre-experimental simulation method for quantitative research. MASS embeds LLMs in an agent-based modeling (ABM) framework and uses role settings, round-based scheduling, information control, background-rule control, structured outputs, harness checks, reason-action logs, and replication manifests to transform open-ended language generation into recordable, checkable, and statistically analyzable agent-round observations. The method is evaluated through the New Jersey–Pennsylvania minimum wage natural experiment, the 2016 UK Brexit digital campaigning context, and the 2023 Zibo barbecue tourism public-opinion event. Results show that protocolized LLM-driven ABM can generate analyzable and empirically assessable outputs across policy-shock, information-intervention, and governance-feedback scenarios. The strongest evidence concerns rule-shock identification, declining undecided share under targeting, and mechanism-chain consistency among governance response, public sentiment, and behavioral intention. MASS is not a substitute for real-world experiments or causal inference; it is a pre-experimental simulation method for mechanism rehearsal, risk identification, counterfactual comparison, and research design preparation.

Keywords: multi-agent social simulation; LLM-driven agent-based modeling; quantitative research methods; computational social science; model validation



Academic Editor: Miguel Angel Cazorla

Received: 3 June 2026

Revised: 12 July 2026

Accepted: 21 July 2026

Published: 24 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As exemplified by the Stanford Prison Experiment in 1973 [1,2], the social and behavioral sciences have long faced a methodological and ethical dilemma. Researchers need to understand how people respond to institutions, roles, incentives, information, and group environments, yet they cannot freely place real participants into manipulated social situations for repeated experimentation. The closer an experiment comes to real social processes, the more likely it is to produce large and unpredictable consequences. Wage policies, digital targeting, platform-based public-opinion governance, collective mobilization,

organizational coordination, and public-emotion interventions may all generate economic, psychological, or social consequences in real-world settings. Researchers can observe such consequences after they occur, but it is difficult to actively manipulate, regulate, and repeat them under controlled conditions in real populations.

The social sciences, therefore, need a simulation sandbox that can be run before real-world interventions. Such a sandbox would provide a low-risk, repeatable, and adjustable environment in which researchers can rehearse possible behavioral pathways, mechanism chains, and counterfactual outcomes before formal policies, communication interventions, or governance practices are implemented. For research on public policy, group communication, and social governance, such pre-experimental environments can supplement real-world evidence by helping researchers identify potential mechanisms, compare alternative conditions, and improve subsequent empirical research designs under lower-risk conditions.

To date, the social sciences have developed multiple alternative and simulation-based methods to address the ethical constraints, costs, and limited repeatability of real-world experiments. Natural experiments, quasi-experimental designs, wargaming, policy games, tabletop exercises, system dynamics, microsimulation, and agent-based modeling (ABM) each provide different ways of handling this problem.

Natural experiments and quasi-experimental designs provide social scientists with a real-world logic for causal identification. Researchers use policy changes, institutional differences, or external shocks that have already occurred to construct treatment and control groups and then evaluate intervention effects through methods such as difference-in-differences, pre-trend tests, and placebo tests [3–5]. For example, Card and Krueger's study [6,7] of the New Jersey–Pennsylvania minimum wage case used a real policy change to construct treatment and comparison conditions, allowing researchers to compare changes in wages, employment, and prices. The advantage of natural experiments lies in their proximity to real social processes, but their limitation is that researchers cannot actively set the timing, intensity, external disturbances, or repeated implementation of the intervention.

Wargaming, policy games, and tabletop exercises provide another process-oriented simulation pathway. These methods use structured scenarios, role assignment, multi-round actions, strategic interaction, and situational adjudication to explore complex social processes. Policy games, role-playing simulations, serious games, and participatory simulations extend this logic by allowing participants to act in different roles and make continuous decisions within constrained environments [8]. Such methods can represent incomplete information, conflicting goals, and strategic adaptation among multiple actors. However, they are often limited by weaker reproducibility, insufficient quantitative outputs, and a dependence on participant experience or expert adjudication.

Computational social simulation further transforms social processes into executable models. System dynamics is suitable for modeling macro-level feedback loops, microsimulation is suitable for projecting policy effects at the individual level, and ABM simulates macro-level emergence through heterogeneous agents, local interactions, and environmental rules [9–12]. ABM allows researchers to define agent types, behavioral rules, interaction structures, and environmental states, and to observe system behavior through repeated runs. However, the behavioral rules of agents in ABM are usually specified manually by researchers. When the research object involves linguistic action, role interpretation, institutional meaning, local information, and situational judgment, fixed functions and heuristic rules often struggle to capture the behavioral logic of real actors. Put simply, in traditional ABM, agents cannot use natural language to describe or explain their own behavior in detail.

LLM-driven ABM lies at the intersection of these methodological traditions. It absorbs the treatment-control and counterfactual logic of natural experiments, extends the multi-role, multi-round, and situational interaction logic of wargaming and policy games, preserves the agent modeling, state updating, and repeatable execution of ABM, and uses large language models (LLMs) to enhance linguistic, role-based, and context-sensitive behavioral generation. The introduction of LLMs gives ABM a more dynamic and language-sensitive form of behavioral simulation. Agents can generate actions and reasons based on role cards, local information, external rules, and current-round events, while researchers can transform these actions into structured variables for statistical testing, process auditing, and cross-case comparison.

1.1. The Rise of AI-Driven Social Simulation Platforms and the Methodological Gap

In recent years, several representative AI-driven social simulation systems have emerged. Generative Agents demonstrated the ability of LLM agents to generate memory, reflection, planning, and everyday interaction in a small sandbox society [13]. Agent Hospital embeds agents such as doctors, nurses, and patients into a virtual hospital workflow to simulate multi-stage medical behavior [14]. OASIS and AgentSociety further extend this line of work toward social media diffusion and large-scale social-life simulation, exploring information diffusion, group polarization, external shocks, and public policy [15,16]. These studies show that AI-driven social simulation has moved from proof of concept toward platformized practice, but existing platforms are still mostly organized around specific scenarios, fixed environments, or system capabilities.

Unlike AI social simulation platforms centered on fixed scenarios, this study focuses on a transferable research method: how to organize LLM-driven agent-based modeling (LLM-driven ABM) into a configurable, auditable, and empirically assessable simulation experiment. This study operationalizes the method as Multi-Agent Social Simulation (MASS), whose core lies in a reusable research protocol. Researchers need to define role types, temporal granularity, background rules, information boundaries, output structures, and validation indicators around specific research questions, while maintaining comparable experimental designs across treatment, control, and counterfactual conditions. This approach extends the ABM literature on model transparency, process description, and reproducibility. For example, the ODD protocol emphasizes that model purpose, entities, state variables, scales, processes, scheduling, initialization, and input data should be explicitly reported to improve the transparency and reproducibility of simulation models [17,18]. For LLM-driven ABM, this model-description requirement must be further extended to prompts, role cards, information boundaries, background rules, experimental conditions, output schemas, model versions, runtime configurations, and validation targets. Accordingly, prompt design, role cards, background rules and experimental conditions, output schemas, logging mechanisms, and validation matrices are not merely engineering configurations; they are components of the research design, corresponding respectively to behavioral generation, variable construction, experimental conditions, process auditing, and methodological validity assessment.

Protocolization is the key condition for transforming LLM-driven ABM from narrative role-playing into a reviewable research method. Naive LLM-agent simulations usually provide models with role prompts and ask them to continuously output actions or opinions. Such practices can rapidly generate seemingly plausible narrative results, but they are prone to temporal-order errors, role-identity deviation, data contamination, rule-or-condition errors, outputs without structured fields, numerical out-of-range errors, and difficulties in reproduction. Recent research on LLM-empowered ABM and generative social simulation has also noted that although LLMs enhance ABM in environmental perception,

heterogeneous-agent expression, reasoning and decision-making, and action generation, they also introduce new challenges related to output reliability, model-version dependence, action validity, and validation standards [19,20]. Therefore, LLM-driven ABM for social science research cannot remain at the level of free conversation or narrative rehearsal. It must integrate rounds, roles, information, rules, outputs, checks, and validation into an explicit protocol. Only in this way can the simulation process be recorded, the simulation results audited, and different runs, conditions, and cases form a basis for empirical comparison.

1.2. Research Questions and Contributions

Based on the methodological positioning above, this study proposes three research questions.

RQ1: Can LLM-driven ABM serve as a method for simulating social behavior in behavioral science?

RQ2: What methodological failures arise in naive LLM-agent simulation, and what forms of protocolized control are needed to mitigate them?

RQ3: After protocolized control is introduced, can LLM-driven ABM generate auditable, empirically assessable simulation outputs with evidence of cross-scenario applicability across different social-behavioral settings?

This study makes four main contributions in relation to these questions.

First, starting from the ethical boundaries of real-world social experimentation, this study proposes the value of LLM-driven ABM as a low-risk pre-experimental method. This method does not replace experiments with real participants or natural experiments. Rather, it provides a simulation space for mechanism rehearsal, risk identification, and counterfactual comparison.

Second, this study diagnoses the typical failure modes of naive LLM-agent simulation and translates temporal-order errors, role-identity deviation, data contamination, rule-or-condition errors, unstructured outputs, invalid numerical values, and replication difficulties into operational protocol-control items.

Third, this study proposes a protocolized LLM-driven ABM framework that incorporates prompts, role cards, background rules, experimental conditions, information boundaries, output schemas, harness checks, and validation targets into the research design. Through this framework, LLM-driven ABM can generate behavioral simulation outputs that are auditable, reproducible, and empirically comparable.

Fourth, this study conducts a multi-evidence evaluation through three social-behavioral cases: a minimum wage policy shock, a digital campaigning information intervention, and urban tourism public-opinion dynamics. These three cases jointly test the method's data analyzability, counterfactual comparability, group-ordering consistency, and process-explanation capacity.

2. Literature Review

To answer the research questions above, this study first situates LLM-driven ABM within the broader methodological landscape of social simulation and behavioral science. The key issue is not only what existing methods have already solved, but also where gaps remain in reproducibility, auditability, behavioral rule specification, and validity assessment. Clarifying these methodological foundations is necessary for explaining why protocolized LLM-driven ABM is needed and what forms of evaluation it should undergo.

This chapter proceeds in three steps. First, it returns to the problem of behavioral rule specification in ABM. It discusses how traditional ABM explains macro-level emergence through agents, environments, rules, and state updating, while also showing why hand-written rules and fixed functions face expressive pressure when the research object involves linguistic interaction, role identity, institutional interpretation, local information, and situa-

tional judgment. Second, it discusses replication and process auditing in LLM-driven ABM. LLMs can extend the behavioral generation capacity of ABM, but they also introduce new problems such as prompt sensitivity, role-identity deviation, data contamination, output instability, and model-version dependence. These problems call for protocolized controls such as prompt manifests, replication manifests, reason-action logs, harness checks, and structured outputs. Third, the chapter discusses validity assessment in generative social simulation. For LLM-driven ABM, the central issue is not whether the model can generate text that “looks plausible,” but whether it can produce simulation data that are structurally clear, process-auditable, and empirically comparable.

2.1. ABM, LLM-Driven ABM, and the Problem of Behavioral Rule Specification

ABM consists of agents, environments, state variables, behavioral rules, and interaction mechanisms. It is well suited for studying how local interactions among heterogeneous agents generate macro-level social outcomes. Its advantages lie in controllability, repeatability, counterfactual reasoning, and the ability to observe emergent structures and nonlinear processes [9,12]. In the social sciences, ABM has often been used to explain group segregation [21], norm diffusion, organizational behavior, market change, public-opinion dynamics, and policy outcomes.

A central challenge in traditional ABM is the specification of behavioral rules. Researchers usually need to define in advance how agents perceive the environment, choose actions, update states, and interact with other agents. These rules may be derived from rational choice, bounded rationality, empirical statistics, expert judgment, or existing theory. Simon’s discussion [22] of bounded rationality suggests that real-world actors often make satisficing decisions under incomplete information, limited cognitive capacity, and complex situations. In many social-behavioral settings, action is also shaped by role identity, institutional interpretation, linguistic framing, local information, social expectations, and situational meaning. When traditional ABM handles such behavior, complex action is often compressed into fixed functions or heuristic rules.

For social-behavioral simulation, the methodological significance of LLM-driven ABM first derives from the corpus-based foundation of large language models. LLMs do not generate actions from abstract rules in a vacuum; rather, they learn conditional relationships among roles, contexts, norms, attitudes, and response patterns from large-scale human text corpora. Existing research has shown that, under appropriate conditioning and clearly specified task constraints, language models can to some extent approximate the response distributions of particular human samples or subgroups. Argyle et al. introduced the concept of “algorithmic fidelity,” arguing that language models conditioned on demographic and social background information can generate attitude distributions corresponding to multiple human subgroups [23]. Aher et al. further proposed “Turing Experiments” to examine the ability of language models to replicate classic human-subject experiments, showing that LLMs can generate responses close to human experimental patterns in several tasks in economics, psycholinguistics, and social psychology [24]. AITurk-related research also suggests that ChatGPT 4.0 can imitate the psychological and behavioral responses of human participants at relatively high rates in several online crowdsourcing research contexts [25]. Taken together, these studies suggest that LLMs can be viewed as conditional behavior generators grounded in accumulated public language data, rather than as mere text-completion tools.

This provides a new possibility for specifying behavioral rules in ABM. Traditional ABM requires researchers to predefine how agents perceive environments, choose actions, and update states. This may compress complex role identities, institutional interpretations, local information, and social expectations into fixed functions or heuristics. LLM-driven

ABM does not remove the researcher's control over model structure. Rather, it extends behavioral rules from static handwritten rules to protocolized generative rules. Researchers remain responsible for defining role types, background rules, experimental conditions, information boundaries, state variables, and validation indicators. Within these constraints, LLMs generate concrete actions based on role cards, local information, and current-round events. In this way, situational judgment, linguistic expression, and role-based response in social behavior can be incorporated into a recordable and auditable simulation process.

However, the approximation of human responses by LLMs is not unconditional. Model outputs may be affected by training corpora, prompts, model versions, alignment mechanisms, and social desirability bias. They may also produce overly consistent, overconfident, or majority-oriented responses in particular tasks. Therefore, this study does not treat LLM agents as substitutes for real individuals or real samples. Instead, it embeds them into ABM as constrained behavior generators. Role cards determine the social position and behavioral preferences of agents; local information determines their visible environment; background rules and experimental conditions specify institutional constraints; round-based scheduling defines the boundary of state updating; structured outputs transform actions into analyzable variables; and harness checks and validation targets are used to examine role consistency, information boundaries, numerical validity, and result comparability. In this sense, the key to LLM-driven ABM is not to let models freely simulate human beings, but to transform human response patterns embedded in language corpora into reproducible, comparable, and empirically assessable social-simulation data through protocolized control.

2.2. Replication, Process Auditing, and Model Description

The credibility of social simulation models depends on reproducibility and auditability. The ABM field has long emphasized the importance of model description. The ODD protocol requires researchers to describe the model's purpose, agents, state variables, scales, processes, scheduling, initialization, input data, and submodels in a standardized manner, thereby improving model transparency, replicability, and comparability [17,18]. For LLM-driven ABM, model description must also cover prompts, role cards, information boundaries, background rules, experimental conditions, output schemas, model versions, randomness settings, retry policies, and validation targets.

In LLM-driven ABM, replication is more challenging than in traditional ABM. Technically, Transformer-based language models generate subsequent tokens through conditional probabilities. Vaswani et al. noted in the Transformer paper that decoder outputs are converted by softmax into predicted next-token probabilities [26]. In actual generation processes, outputs are also affected by decoding strategies, temperature, top_p, random seeds, context length, and server-side backend configurations. Holtzman et al.'s study of neural text degeneration shows that, even with exactly the same neural language model, different decoding strategies can significantly affect generated outputs [27]. The reproducible-output example in the OpenAI Cookbook also notes that the seed parameter can improve output consistency, but that its effect depends on identical request parameters and system_fingerprint and cannot guarantee absolute determinism [28]. Therefore, a single run of LLM-driven ABM should not be treated as a stable conclusion. Researchers need to reduce the influence of single-run stochastic output through multiple independent runs, directional consistency checks, and cross-condition comparisons.

To standardize replication, researchers need to build both a prompt manifest and a replication manifest. The prompt manifest records role prompts, background prompts, rule prompts, output prompts, forbidden information, and visible information boundaries. The replication manifest records model names, running time, parameter settings, data versions,

case configurations, round granularity, randomness controls, failure-retry mechanisms, and backend-configuration identifiers. Without these records, it is difficult for third parties to determine whether simulation results arise from social mechanisms, prompt induction, model priors, backend changes, or random output variation.

Process auditing is another key requirement for LLM-driven ABM. When LLMs generate behavior, researchers need not only structured variables but also auxiliary materials that can trace the behavior-generation process. A reason-action log can record how an agent understands local information in each round, how it chooses actions, and how those actions are translated into state changes. Its function is to serve as process-auditing material, recording behavioral generation pathways and checking output consistency. Researchers can use reason-action logs to examine whether actions conform to role boundaries, whether they reference forbidden information, whether hindsight contamination occurs, and whether they are consistent with the direction of numerical outputs. Structured fields provide the primary objects of analysis, while explanatory text provides auxiliary auditing material.

Output structure and runtime checking mechanisms determine whether simulation data can enter statistical analysis. Naive LLM-agent simulations often generate long natural language outputs, making it difficult for researchers to compare different runs, conditions, and cases. LLM-driven ABM for behavioral science research requires structured outputs, value ranges, state-update boundaries, and format checks so that simulation results can be statistically analyzed, visually compared, and checked for anomalies. Reproducibility, process auditing, and output structure therefore jointly determine whether LLM-driven ABM can move from narrative simulation to a verifiable research method. In Section 3, this study further operationalizes these principles as prompt manifests, replication manifests, harness checks, local checks, structured outputs, and full-log preservation.

2.3. *Validity, Reliability, and the Evaluation Gap in Generative Social Simulation*

The credibility of simulation models cannot rely solely on whether their results “look reasonable.” For social simulation, reliability and validity involve at least two levels. Reliability concerns whether a model can generate stable, reviewable, and repeatedly analyzable outputs under identical or similar conditions. Validity concerns whether model structure, agent behavior, process mechanisms, and outcome directions form an interpretable correspondence with the target social process. In other words, a social simulation model should not only generate an outcome; it should also explain how that outcome is produced by agents, rules, environments, and interaction processes, and whether that generative process can be recorded, reviewed, and compared.

Existing simulation research provides important theoretical foundations for this issue. Guala [29] argues that complex relationships exist among models, simulations, and experiments, and that simulations explore mechanisms and possibilities through artificially constructed systems. Morgan [30] discusses forms of experimentation without material intervention, showing that model experiments and virtual experiments can also play a role in scientific reasoning. Sugden [31] proposes the concept of “credible worlds,” emphasizing that the value of a model lies not in replicating every detail of the real world but in constructing an internally coherent artificial world that can carry theoretical mechanisms and support reasoning. For ABM, this means that model validity should not be judged only by endpoint outcomes. It also requires examining whether agent settings, behavioral rules, interaction structures, state updates, and macro-level outputs are internally consistent.

This issue becomes even more prominent in LLM-driven ABM. Recent reviews have pointed out that large language models strengthen ABM in environmental perception, heterogeneous agent representation, reasoning and decision-making, and action generation, while also introducing new challenges related to human alignment, action reliability,

evaluation standards, and cross-scenario validation [19]. Larooij and Törnberg [20] further argue that validation is the central challenge for generative social simulation: many studies still rely mainly on subjective judgments of whether simulations “appear credible,” while providing limited evidence that models are operationally valid. For this study, this means that LLM-driven ABM should not merely demonstrate narrative plausibility or role-playing capacity. It must provide reviewable structural evidence, process evidence, and outcome evidence.

Accordingly, this study designs validity and reliability assessment as a multi-evidence framework rather than relying on a single indicator. At the structural level, the model must specify whether agent types, state variables, temporal granularity, rule conditions, and information boundaries reasonably represent the target social process. At the behavioral level, the model must examine whether agent actions conform to role positions, local information, and institutional constraints. At the process level, the model must preserve reason-action logs, public messages, state changes, and harness results so that behavioral generation pathways can be audited. At the outcome level, the model must form comparable relationships with real-world references, natural-experiment findings, counterfactual conditions, known group orderings, or mechanism chains.

Based on this logic, Section 3 uses three complementary cases to examine validity and reliability. Case 1 uses the New Jersey–Pennsylvania minimum wage natural experiment, with treatment, comparison, and counterfactual groups, and uses DID and DDD to test whether the minimum wage rule shock can stably enter managerial decision-making. Case 2 compares targeting and no-targeting conditions to examine whether a digital targeting mechanism can generate distinguishable changes in group attitudes, and uses known group ordering to check whether role settings remain consistent. Case 3 compares active governance and passive governance conditions to examine whether platform attention, offline experience, and governance feedback form a directionally consistent mechanism chain. Together, the three cases cover rule-shock identification, information-intervention differentiation, and multi-agent interaction with process consistency in a public information environment, thereby providing multiple types of evidence for the reliability and validity of MASS.

This leads to the literature gap and research positioning of the present study. Natural experiments provide treatment-control logic and causal identification in the real world, but researchers cannot actively repeat or adjust intervention conditions. Wargaming, policy games, and role-playing simulations provide multi-role, multi-round interaction, but their reproducibility and quantitative outputs are limited. Traditional ABM provides a repeatable computational framework, but its behavioral rules are often static. LLM agents provide linguistic and role-based behavioral generation capacity, but are prone to prompt sensitivity, role-identity deviation, data contamination, and difficulties in replication. The protocolized LLM-driven ABM proposed in this study builds a bridge among these methodological capacities. It preserves the repeatability and counterfactual capacity of ABM, introduces the contextual behavioral generation capacity of LLMs, and uses structured outputs, process auditing, and a multi-evidence validation framework to examine whether it can serve as an auditable, reproducible, and empirically assessable method for social simulation research.

3. Materials and Methods

3.1. Naive LLM-Agent Baseline and Failure Modes

This study first constructs a naive LLM-agent baseline as a methodological reference to illustrate the problems that may arise when large language models are directly connected to agent simulation. The baseline follows a simple process: researchers write role prompts for each agent; in each round, the system provides the current situation to the model; the

model outputs an action or opinion; and the researchers feed the output into the next-round state. This process can quickly generate narrative simulation results, but it lacks strict temporal boundaries, role boundaries, information boundaries, and output constraints, making it difficult to use directly as a reviewable behavioral science research method. In this study, the baseline is the methodological starting point and the underlying foundation of the project: using LLMs to conduct social simulation through text as the medium.

The main problem with the naive baseline is that its generated behavior lacks stable experimental discipline. An agent may predict the future within the same round, step outside its own role identity to provide narrator-like explanations, or refer to researcher-side information or real final outcomes. Natural language outputs may also lack computable fields, making it difficult to compare different runs, conditions, and cases. Therefore, this study treats naive LLM-agent simulation as an initial form that must be protocolized and summarizes eight typical failure modes accordingly.

To connect the risks of naive LLM-agent simulation with MASS control mechanisms, Table 1 summarizes the eight failure modes identified in this study, their typical manifestations, their research implications, and the corresponding forms of protocolized control.

Table 1. Major failure modes of the naive LLM-agent baseline and MASS control mechanisms.

Failure Mode	Typical Manifestation	Research Implication	MASS Control Mechanism
Temporal-order errors	Predicting the future too early or repeatedly updating states	Unstable behavioral trajectories	Fixed round-based scheduling
Role-identity deviation	Shifting from the assigned role to another role	Distorted behavioral subject	Role cards and role-boundary checks
Data contamination	Accessing information that should not be available, leaking final outcomes, or leaking researcher-side information	Backward reasoning or misattribution	Information control, locally visible information, and case blinding
Rule-or-condition errors	Changing background settings, policy conditions, price constraints, or action ranges autonomously	Loss of experimental-condition control	Locked background rules and experimental conditions
Outputs without structured fields	Long text without stable fields	Difficult statistical analysis	Structured output schema
Numerical or logical errors	Out-of-range values or internal contradictions	Distorted state updating	Local numerical checks
Missing process records	Results without auditable reasons	Difficult process review	Reason-action logs
Missing reproducibility information	Missing prompts, model names, or configurations	Difficult third-party review	Prompt and replication manifests

The failure modes above provide the basis for the protocol design in this study. Protocolization aims to place the open-ended generative capacity of LLMs within a controllable ABM structure so that agent actions can retain linguistic, role-based, and context-sensitive features while also satisfying the requirements of state updating, output recording, and empirical validation.

3.2. Protocolized Control Scheme, System Architecture, and Operational Implementation

This study operationalizes a protocolized LLM-driven ABM as a MASS simulation workflow. Researchers first define the research question; specify the number, types, and relationships of agents; prepare background information; and define state variables, output fields, and experimental conditions. These elements are then recorded in a config file or configured through the MASS platform. At runtime, MASS uses the configuration to

schedule rounds, assemble contexts, call the model, parse structured outputs, run local and harness checks, update states, preserve logs, and export data for analysis.

Case design provides the methodological starting point for MASS. It specifies the social process to be simulated, the roles and experimental conditions, the information visible to agents, the variables to be analyzed, and the criteria for acceptable outputs. At runtime, MASS translates the design into executable role definitions, background rules, information boundaries, output schemas, checking rules, and logging procedures. This places LLM-generated language within the round structure, rules, states, and validation framework of ABM, producing agent-round observations that can be recorded, audited, reproduced, and statistically analyzed.

Figure 1 illustrates the researcher workflow for designing and configuring a MASS simulation. Researchers define a question; specify agent numbers, types, and functional roles; design inter-agent relationships; prepare the background information required by the LLM agents; and identify data fields and state variables. These specifications are consolidated into a config file or entered into the MASS platform. Pilot runs and preliminary analysis then test whether agent settings, background information, variables, and preliminary outputs serve the research purpose. Identified issues feed back into configuration; once the configuration is stable, formal simulation and validation analysis begin.

Researcher-side design and configuration workflow

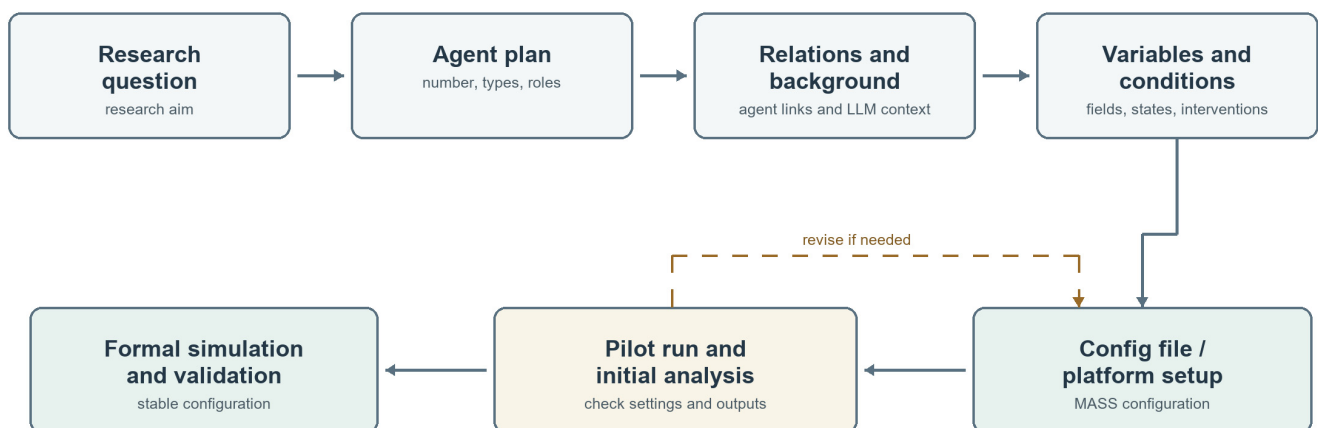


Figure 1. Researcher workflow for designing and configuring a MASS simulation.

3.2.1. Overall Architecture and Agent Scheduling Mechanism

After the case design and platform configuration shown in Figure 1 are completed, MASS converts the config file or platform inputs into a structured experimental configuration. The round scheduler then advances the simulation according to the preset temporal granularity. The case background, role structure, experimental conditions, variable fields, output formats, and validation targets specified during configuration serve as runtime constraints; agents cannot alter them during the simulation. In each round, the system assembles each agent's visible context, invokes the designated LLM, and passes the resulting output to downstream parsing, checking, state-updating, and log-preservation procedures.

MASS processes each agent output in a fixed sequence. In a given round, the output normally passes through six steps: first, the system assembles the prompt context from the role card, current-round events, background rules, visible information from the previous round, and the output schema; second, it calls the LLM and receives the raw output; third, the structured output parser and local format check assess whether the output conforms to the required structure; fourth, the case-specific harness check assesses whether it violates role boundaries, information boundaries, or case-specific rules; fifth, deterministic local

checks assess fields, formats, value ranges, and basic logical consistency; sixth, records that pass the checks enter state updating, the public board, and the formal analysis table. If any stage fails, the system triggers a correction retry according to the configuration. A record that still fails after the maximum number of retries remains in the full log and is not manually completed, imputed, or included in the formal statistical analysis.

The current-round context of each agent comprises five categories of information:

- A. Role settings, including identity, social position, action goals, decision concerns, and communication permissions;
- B. Current-round events and background rules;
- C. Locally visible information, including previous-round state summaries and public-board fields that the agent is permitted to read;
- D. The output schema, namely the structured fields and formats that the agent must return; and
- E. Formatting requirements, prohibitions, and boundary reminders specified by the case-specific harness or local checks.

MASS does not rely on model-side long-term memory. State preservation, public-board updating, log recording, and data export are all handled by MASS, and every round's visible information is assembled from researcher-defined context rules.

3.2.2. Role Settings

In MASS, an agent typically represents a social role or actor group rather than a single real individual. Researchers use role cards to define an agent's identity, social position, action goals, visible information, behavioral preferences, and communication permissions, allowing it to maintain a specific role identity throughout the simulation. Group-level agents are better suited to meso- and macro-level social processes than direct simulations of individuals and reduce interference from individual heterogeneity.

During runtime, the system constrains roles through role prompts, identity fields, visible-information fields, and the output schema. In each round, an agent may generate actions or state updates only from its own role, local information, and current-round rules; it cannot access the complete experimental design, hidden validation targets, or final outcomes. If an output clearly exceeds the role boundary, for example by making decisions for another role, summarizing final outcomes from the researcher's perspective, or citing information unavailable to the current role, the harness check flags a role-boundary or information-boundary violation and triggers a correction retry. Role settings therefore both support behavior generation and delimit role boundaries.

3.2.3. Agent Interaction Mechanism

MASS organizes interactions among agents through information boundaries and a public board. The public board is a system-maintained shared information area that records information agents are permitted to disclose and makes it available in later rounds to agents with reading permission. Here, public denotes visibility within the simulated environment; it is unrelated to actual public disclosure or real-platform data.

Agent communication is implemented through prompt-mediated interaction. Public fields written by one agent in the current round are filtered by the system and injected into the prompt contexts of relevant agents in the next round. The main communication fields used in this study are `action_summary`, `public_message`, `hierarchical report`, and `broadcast`; `action_summary` is a brief structured summary of an agent's action in the current round; `public_message` is information written to the public board for authorized agents to read in subsequent rounds; `hierarchical report` is a structured report transmitted by organizational or hierarchical agents to superiors, subordinates, or coordinating agents; and `broadcast` is

an event announcement preconfigured by researchers or produced by a rule-based process to inject exogenous events, policy changes, platform rules, or governance requirements into relevant agents at specified rounds.

The scheduler mediates agent interaction, which is governed by round order, role permissions, and field-level access ranges. Role permissions define the types of information that each agent class may send, receive, or read under the case design. A field-level access range is a whitelist of fields that an agent class may read and that may therefore enter its next-round prompt. To prevent data contamination, the public board exposes only explicitly permitted fields, such as `agent_id`, `round_index`, `event_id`, `action_summary`, and `public_message`. Hidden fields, including `state_before`, `state_delta`, `state_after`, `confidence`, `evidence_basis`, `validation_targets`, `researcher notes`, `official outcome summary`, and `final real-world results`, do not enter agent-visible contexts. Field-level information boundaries separate simulation-internal public information from researcher-side validation information and reduce the risk that agents infer hidden states or final outcomes.

This study uses the public board, hierarchical report, broadcast, and prompt-context injection to manage agent interaction and does not introduce direct communication protocols such as A2A or MCP. The emphasis is on information boundaries, process auditing, and reproducibility control within a single simulation environment; these mechanisms provide the interaction functions required here while reducing additional external dependencies and uncontrolled information flows.

3.2.4. Background Rules and Experimental Condition Control

MASS uses pre-specified background rules and experimental conditions to control common constraints and between-group differences. In the case design, researchers specify policy changes, information interventions, platform rules, governance responses, or event shocks, together with their timing, target population, and scope. At the relevant stage, the system writes the applicable rules or conditions into agent-visible contexts so that agents can adjust their actions under the specified conditions.

Differences among treatment, control, and counterfactual groups are specified at the research-design stage. Agents may adjust their own actions under the assigned rules, but they cannot alter policies, platform mechanisms, governance conditions, or event shocks. For example, in the minimum-wage case, the New Jersey wage increase is an exogenous rule specified in the case design: agents may adjust starting wages, employment, or prices in response, but cannot decide whether the policy occurs. In the digital campaign case, targeting and no-targeting are researcher-defined conditions that agents cannot change during runtime. In the public-opinion governance case, active governance and passive governance are preset counterfactual conditions to which agents respond. These arrangements preserve comparability across experimental conditions and reduce the risk of uncontrolled simulations caused by agents rewriting the conditions.

3.2.5. Structured Outputs, State Updating, and Formal Analysis Tables

Drawing on failures observed in prior runs, this study requires agents to return structured fields under a predefined schema. Structured outputs convert natural-language actions into analyzable variables and facilitate format checks, field checks, and subsequent statistical analysis. Different research settings may use different state variables, but every variable must be defined in the experimental configuration in advance, including its field name, type, value range, update method, and inclusion in the formal statistical analysis.

In MASS, state variables are updated from structured state fields returned by an agent in each round and passed through the checks. Agents submit their current-round actions, explanations, and state variables; the system accepts only fields that pass the parser, local

checks, and harness checks. Accepted fields are recorded as the agent's state in that round and may become part of the next round's historical summary, public-board information, or formal analysis data.

State updating follows the principle of round boundaries. An agent outputs only current-round actions and state changes, and state variables are updated after system parsing and checking. For direct state-value variables, such as `leave_support`, `undecided_share`, `positive_sentiment`, `visit_intent`, and `trust_in_official_response`, the checked output value in the current round is recorded as the agent-round state value. For settings with `state_before`, `state_delta`, and `state_after`, the system requires all fields and checks whether the relation between `state_before`, `state_delta`, and `state_after` is consistent. If `state_after` clearly conflicts with `state_before + state_delta`, or a state variable falls outside its preset range, the record is treated as a state-update error and triggers a retry or enters the full log.

In Case 1, `starting_wage`, `target_fte`, and `price_full_meal_after_tax` are closer to managerial decision variables. Each agent returns its planned weekly value; after checking field format, numerical range, and case rules, the system records the value as that agent-round's decision state. Valid wage ranges mainly derive from policy rules and minimum-wage constraints; for example, `starting_wage` should not fall below the `minimum_wage_floor` applicable to the current round. Price checks focus on numerical plausibility and abnormal changes; for example, negative, missing, or extreme values are not permitted. The sources of these checks differ: wage ranges derive primarily from institutional rules, whereas price ranges derive from researcher-defined variable ranges and data-cleaning rules in the case design.

The formal analysis table contains only agent-round observations that are successfully parsed and retained. Failed calls, timeouts, invalid outputs, missing fields, format errors, and out-of-range values enter the full log or audit log and are not manually repaired or imputed before formal statistical analysis. Formal analysis therefore uses simulation records that have undergone structured parsing and runtime checking, while the complete logs preserve failures and anomalies for process auditing.

3.2.6. Output Validation Mechanism

MASS validates output through local checks and case-specific harness checks, which serve distinct functions. Local checks address structural, formatting, and deterministic hard-rule requirements, including whether an output can be parsed, conforms to the JSON or key-value capsule structure, contains required fields, uses correct field types, converts numerical fields to numbers, and places variables within their preset ranges. The Structured Output Parser is the entry point to local checks: it first determines whether the raw LLM output can be parsed into structured fields that the system can process. When the structure cannot be parsed, a required field is missing, or a field type is clearly incorrect, the record triggers a format-correction retry and does not enter state updating.

Harness checks address more complex case-specific review. Each harness is generated before runtime for a specific case. The researcher-defined case background, role structure, experimental conditions, information boundaries, output schema, and background rules provide the source material. The system uses an LLM to extract invalid-output rules from these materials and forms case-specific harness rules. These rules are then applied during runtime to identify outputs that clearly violate the case setting, role boundary, information boundary, or background rules. Thus, the harness is used for runtime invalid-output detection, rather than for outcome-level validity assessment.

Harness checks examine whether an agent output violates role boundaries, information boundaries, or case-specific rules. For example, the system checks whether an agent cites final real-world results or the official outcome summary; uses researcher-side infor-

mation such as `validation_targets` or researcher notes; writes hidden state fields such as `state_after` or `state_delta` into a `public_message`; makes decisions for another role; changes a preset policy condition, governance condition, or platform rule; or uses future information that should not be visible in the current round. Such issues usually fall outside ordinary format checking and therefore require case-specific harness review.

In this study, invalid output includes JSON format errors, missing fields, incorrect field types, out-of-range values, inconsistent state-update relations, hidden-field leakage, role-boundary violations, information-boundary violations, and agent-initiated rewrites of experimental conditions. For example, an unparseable state field or the absence of a schema-required field is a structural error; a clear inconsistency between `state_after` and `state_before + state_delta` is a state-update contradiction; writing `validation_targets`, final real-world results, or researcher notes into a `public_message` is hidden-information leakage; and independently declaring a policy change or making a decision for another role is a rule or role-boundary error.

When an output fails local checks or harness checks, the system triggers a correction retry according to the configuration. A retry does not alter the experimental conditions or manually supply content for an agent; the model must return an output that satisfies the schema and boundary requirements under the same role, rules, and visible information. If the maximum retry count is reached without success, the record is marked and retained in the full log but does not enter the formal analysis table. For repeated failures of the same key rule, MASS stops the current run so that the researcher can revise the case design, background materials, output schema, or variable-range settings before initiating a new run. MASS thereby distinguishes analyzable records, audit records, and run-level configuration issues through a retry, marking, log retention, formal analysis exclusion, and run-stopping procedure.

Runtime checking and outcome validation operate at different levels. Local checks and harness checks determine whether an individual agent-round observation has the correct format, complete fields, compliant boundaries, and basic logical usability. DID, DDD, group ordering, directional consistency, and mechanism-chain tests evaluate the overall results after simulation and assess whether they provide evidence for the reliability and validity of MASS.

Figure 2 summarizes the MASS runtime workflow. After researchers complete the case design and platform configuration, the system translates them into a scenario configuration and the round scheduler advances each round. Each agent-round observation proceeds through agent-context building, LLM request-response, structured-output parsing, local-format checking, case-specific harness checking, local deterministic checking, state updating/public-board updating/logging, and data export. Records that pass the checks enter the formal analysis table, whereas records that still fail after retries remain in the full audit log.

Within this architecture, the case design determines the simulation object, experimental conditions, and validation targets; role settings, information boundaries, and background rules delimit the action space of agents; the structured output parser and local checks handle parsing, formatting, and basic numerical checks; and case-specific harness checks address role boundaries, information boundaries, and complex case rules. The formal analysis table receives only agent-round observations that pass parsing and checking, while failed or anomalous records remain in the full logs. MASS outputs therefore retain explicit role, temporal, conditional, action, and state information for subsequent statistical analysis, process auditing, and reliability and validity assessment.

MASS runtime workflow and validation sequence

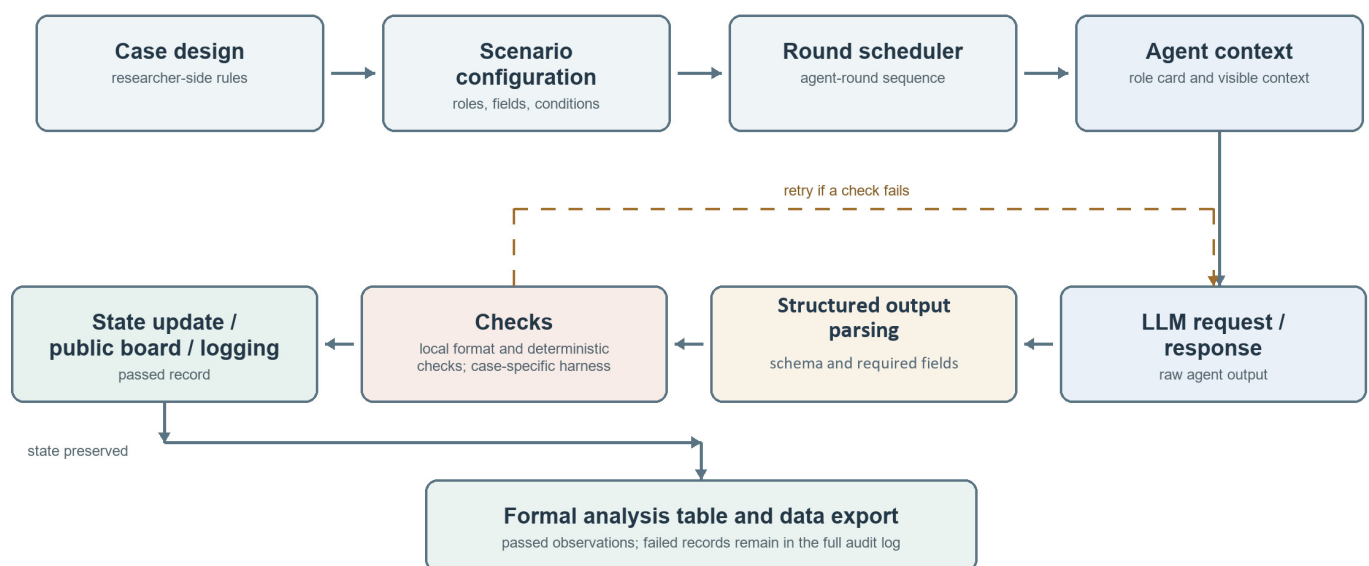


Figure 2. MASS architecture and runtime workflow.

3.3. Three Behavioral Simulation Cases

As shown in Table 2, this study selects three differentiated social-behavioral scenarios for methodological testing. The three cases correspond respectively to a policy shock, an information intervention, and platform-governance feedback. Case selection follows two principles: first, each case has a real-world reference and can be compared with existing empirical data or real-world processes; second, the cases involve different mechanisms, enabling tests of the applicability of protocolized LLM-driven ABM across different social-behavioral scenarios.

Table 2. Design of the three behavioral simulation cases.

Case	Setting	Behavior	Source	Mechanism	Validation Target
1	Managerial decisions under rules	Economic policy	Natural/quasi experiment	Minimum-wage shock in wage, employment, and price decisions	Policy-shock identification
2	Group attitudes under differential information exposure	Political attitudes	Scenario simulation/information-intervention counterfactual	Digital targeting shapes issue salience and attitudes	Information-intervention differentiation
3	Collective behavior under platform attention and governance feedback	Public-opinion event	Open public-opinion simulation/mechanism-chain test	Platform amplification, offline experience, and governance response shape sentiment and intention	Mechanism-chain consistency

The first case is based on the New Jersey–Pennsylvania minimum wage natural experiment [6,7]. The simulated actors are fast-food store-manager groups. Each agent represents several store managers and makes weekly decisions regarding the minimum wage rule, profit pressure, employment constraints, and price adjustment. The main analysis variables retain starting_wage, target_fte, and price_full_meal_after_tax. This case is used to test whether the model can generate comparable policy-shock responses across treatment, control, and counterfactual groups.

The second case uses the Cambridge Analytica controversy and the 2016 UK Brexit referendum digital campaigning context as real-world references [32,33]. It operationalizes data analytics and targeted advertising delivery as simulation conditions and compares differences in voter-group attitudes between the targeting condition and the no-targeting control condition. The main analysis variables include *leave_support*, *remain_support*, *undecided_share*, *turnout_intent*, *trust_in_experts*, and *issue_salience*. This case is used to test whether the model can generate distinguishable group-attitude changes between conditions with and without a digital targeting mechanism.

The third case uses the 2023 Zibo barbecue tourism public-opinion event as a real-world reference to examine the capacity of MASS to simulate an open-ended public-opinion event. The distinctive feature of this case is that complete posts, interaction networks, recommendation mechanisms, tourist-flow data, and consumption microdata from real social media platforms are difficult to obtain in full and difficult for third parties to reproduce. Therefore, this study does not expect the case to directly fit or replicate real platform data, but instead treats it as a public-opinion process simulation case with a publicly available real-world reference. This study uses the Baidu Index search trend for the keyword “Zibo barbecue” as an external reference [34] to assess whether the simulation window covers the broad real-world process from low-level warm-up, rapid diffusion, and peak attention to post-peak decline. It should be emphasized that the Baidu Index does not enter the DID, DDD, or mechanism-correlation models, and it is not used as real-world causal validation data. The main evidence in Case 3 comes from the comparison between the active governance condition and the passive governance counterfactual condition, as well as from whether governance response, public sentiment, trust, and behavioral intention form theoretically expected relationships. In this case, Groups A-F constitute independent runs under the active-governance condition, and Group G is the passive-governance counterfactual condition. Variables are divided into organizational-response and public-response categories. Accordingly, Case 3 should be understood as a case of public-opinion process simulation and mechanism-chain testing, rather than as a direct reproduction of real social media data or real tourism-consumption outcomes. Table 2 above summarizes the differences among the three cases in social setting, mechanism type, and validation target.

3.4. Validation Strategy

This study adopts a multi-evidence validation strategy. For generative simulation methods such as LLM-driven ABM, validation must simultaneously consider data completeness, behavioral boundaries, counterfactual differences, directional consistency, and mechanism transmission.

The study uses difference-in-differences (DID) to compare pre-post changes between treatment and control groups. Its basic form is: $DID = (\text{treatment group post-period} - \text{treatment group pre-period}) - (\text{control group post-period} - \text{control group pre-period})$. In Case 1, DID is used to compare differences between NJ and PA before and after the minimum wage policy shock. In Case 2 and Case 3, DID is used to compare pre-post differences between treatment conditions and control or counterfactual conditions.

The study also introduces difference-in-difference-in-differences (DDD), which adds a third comparison dimension to DID and is used to compare differences in DID estimates across policy conditions or experimental conditions. Its basic form is: $DDD = DID_{\text{policy condition}} - DID_{\text{counterfactual condition}}$. In Case 1, DDD is used to compare the NJ-PA DID difference between the A/B/D policy-shock groups and the C counterfactual group, in order to determine whether the minimum wage rule shock comes from the researcher-

specified policy condition rather than from a historical prior automatically generated by the model.

In addition to DID and DDD, this study uses the multi-evidence validation strategies shown in Table 3 to avoid judging the reliability and validity of MASS on the basis of a single indicator. Different cases use different evidence combinations according to their social mechanisms and data structures: Case 1 focuses on rule-shock identification, Case 2 on information-intervention differentiation and group ordering, and Case 3 on governance feedback and mechanism-chain consistency.

Table 3. Validation strategies and applicable cases.

Validation Evidence	Main Role	Applicable Focus
Data completeness and protocol compliance	Checks whether outputs are analyzable	All three cases
Counterfactual contrast	Compares treatment and control groups	Case 1, Case 2, Case 3
Placebo and pre-trend tests	Rules out false pre-intervention differences	Case 1, Case 2
Repeated runs and directional consistency	Checks result stability	All three cases
Known group ordering	Checks whether role differences are reasonable	Case 2
Mechanism-chain tests	Checks whether variable transmission is consistent	Case 3

3.5. Reproducibility Materials and Data Processing

The data analyzed in this study were generated during the continuous development and protocolized refinement of the MASS platform, and they do not all come from the same version. To avoid misinterpreting differences in platform version, model setting, or group purpose as differences in simulation results, this section distinguishes the data source, runtime condition, model name, and analytical use of each case, as shown in Table 4.

Table 4. Case groups, model versions, and analytical uses.

Case	Groups	Condition	MASS Version Stage	LLM Setting	Analytical Role
Case 1	A, B, D	NJ minimum wage shock	Agent 4 Science version	GPT-4o	Main DID/DDD policy-shock groups
Case 1	C	No NJ minimum wage shock	Agent 4 Science version	GPT-4o	Main DID/DDD counterfactual group
Case 1	E	NJ-PA benchmark scenario with post-harness checks	Harness MASS version	DeepSeek-R1-0528	Supplementary harness-check run only; not included in the main DDD estimates
Case 2	A–F	Targeting condition	Harness MASS version	DeepSeek-R1-0528	Targeting treatment runs
Case 2	G–H	No-targeting condition	Harness MASS version	DeepSeek-R1-0528	No-targeting control runs
Case 3	A–F	Active governance condition	Harness MASS version	DeepSeek-R1-0528	Active-governance treatment runs
Case 3	G	Passive governance condition	Harness MASS version	DeepSeek-R1-0528	Passive-governance counterfactual run

It should be noted that the number of runs in the three cases is not fully balanced. The main goal of this study is not to precisely reproduce a particular real-world case or estimate an average treatment effect under a strictly balanced experiment, but to test the basic reliability and validity of MASS as a protocolized LLM-driven ABM platform and research protocol. The three cases are used to examine whether MASS can generate parseable, auditable, and comparable agent-round observations across different social processes, and whether it can form multiple types of validation evidence through counterfactual

differentiation, directional consistency, known group ordering, and mechanism-chain tests. The larger number of runs under treatment conditions is mainly used to observe mechanism unfolding and output stability, while control or counterfactual groups provide baseline comparisons. Given this data structure, this study confines its results to a multi-evidence test of MASS reliability and validity rather than a strict average treatment effect estimate.

Groups A-D in Case 1 belong to the same batch of historical benchmark data as the NJ-PA minimum wage simulation data used in the authors' earlier MASS methodological exploration paper presented at the Open Conference of AI Agents for Science 2025 [35]. This study does not present these data as newly generated data. Instead, it incorporates them as an early-version historical benchmark into a unified cross-case methodological validation framework and re-conducts data cleaning, DID/DDD calculation, and result interpretation. Group E in Case 1 and all runs in Cases 2 and 3 come from the subsequent post-harness stage and are used to further test the functioning of later MASS builds in structured output, runtime checking, log preservation, and data export.

Groups A-D in Case 1 are the historical benchmark from the initial conference-version implementation, which used GPT-4o. From Group E in Case 1 onward, all subsequent runs use DeepSeek-R1-0528 through the same provider. Neither the model nor the provider changes across these runs.

Subsequent debugging and optimization changes to MASS do not alter its main architecture. Later versions retain the core workflow of front-end experimental setup, structured runtime configuration, role configuration, round scheduling, OpenAI-compatible API calls, structured outputs, harness checks, local numerical checks, log preservation, and data export.

LLM calls are completed through an OpenAI-compatible API and do not rely on vendor-side conversational memory. In each round, for each agent, the MASS scheduler reassembles messages and requests the model once. Agent history enters the next round in the form of structured summaries or restricted fields; the relevant information-boundary and hidden-field controls are described in Section 3.2.

Data processing follows the principle of separating the formal analysis table from full logs. The formal analysis table retains only final successful and parseable agent-round observations for statistical analysis, while the full log retains prompts, model responses, failure reasons, harness results, and structured fields for process auditing. If a model call fails, the output is invalid, required fields are missing, or numerical values are out of range, the system triggers a retry according to local validation rules or retains the record in the full log. Records that do not successfully enter the formal analysis table are not manually rewritten or imputed. The numbers of formal analysis records and supplementary records for each case are reported in the data-completeness table in Section 4.

Statistical analysis is conducted in Python 3.12 in Google Colab, mainly using pandas, statsmodels, and openpyxl. The specific tests for each case follow the validation strategy described in Section 3.4, and all confidence intervals and p values are generated from the Colab statistical scripts.

Scenario configurations, exported data, statistical scripts, and necessary audit materials related to reproducibility will be described in the Supplementary Materials or the Data Availability Statement. Materials involving API keys, account information, platform access credentials, and security-sensitive runtime identifiers are not publicly released.

4. Results

4.1. Data Completeness and Protocolized-Control Effects

This section analyzes the formally exported datasets from the three cases. Unless otherwise specified, all statistical results in this section are based on MASS simulation

exports. External real-world data are used only as case references, event-stage anchors, and validation background. They are not merged with simulation records into the same statistical model. All statistical tables were generated from exported CSV/XLSX files in Google Colab. GPT-assisted analysis was used only for exploratory checking, code debugging, and interpretation support, and was not used as the source of formal statistical results.

This section uses only parseable agent-round observations that enter the formal analysis table. Failed calls, timeouts, invalid outputs, or missing records are retained in the full log or audit records. They are not manually repaired or imputed and do not enter the formal statistical analysis. Table 5 summarizes the formal analysis records, supplementary records, and analytical uses for the three cases. Group E in Case 1 is used only as a supplementary post-harness verification run and is not included in the main DID/DDD estimates for Case 1.

Table 5. Data completeness across the three cases.

Case	Conditions/Runs	Rounds	Agents per Run	Formal Analysis Records	Supplementary Records	Main Use in Analysis
Case 1	4 main + 1 supplementary	46	20	3680	918	Groups A-D are used for the main DID/DDD analysis; Group E is used only as a supplementary harness-check run
Case 2	8	18	11/10	1548	0	Groups A-F are targeting runs; Groups G-H are no-targeting controls
Case 3	7	12	24	2016	0	Groups A-F are active-governance runs; Group G is the passive-governance counterfactual

Table 5 shows that all three cases produced parseable formal analysis data, and the subsequent statistical analyses are based on these exported datasets.

4.2. Case 1: Managerial Decision-Making Under Rule Constraints

Case 1 uses the NJ-PA minimum wage natural experiment as its empirical reference and examines managerial decision-making under a minimum wage rule shock. Groups A, B, and D are policy-shock groups. After Round 9, NJ agents are subject to the minimum wage increase, while PA agents remain under the original wage rule. Group C is the counterfactual group, in which the NJ minimum wage increase is not implemented. This study defines Rounds 1–8 as the pre-intervention period and Rounds 9–46 as the post-intervention period, and compares changes between NJ and PA before and after the intervention. Group E is reported only as a supplementary verification run for the later protocolized version and is not included in the main DDD estimates in Table 6.

Table 6 shows that the DDD estimate for the A/B/D policy-shock groups relative to the C counterfactual group is +0.753, $p < 0.001$, indicating that the minimum wage rule consistently influences NJ managers' starting-wage decisions, whereas the same wage-increase pattern does not automatically appear in Group C when the policy shock is not implemented. The price variable also shows a significant but weaker pass-through result, indicating that the wage-rule shock can be partially reflected in price adjustments. By contrast, the DDD result for target_fte does not reach conventional significance and therefore should not be interpreted as a stable reproduction of real employment effects. Accordingly, the main evidence in Case 1 is rule-shock identification rather than precise prediction of real-world employment effects from the minimum wage.

Table 6. DID and DDD results for the main analysis groups in Case 1.

Variable	A DID	B DID	D DID	Mean DID of A/B/D	C Counter-factual DID	DDD A/B/D vs. C	95% CI for DDD	p	E Supplementary DID
starting_wage	0.734	0.641	0.638	0.671	−0.079	0.753	[0.649, 0.857]	<0.001	−0.078
target_fte	19.679	−10.936	0.695	3.146	−28.979	32.296	[−10.938, 75.530]	0.143	2.572
price_full_meal_after_tax	0.092	−0.009	0.147	0.077	−0.261	0.337	[0.030, 0.644]	0.031	−0.171

Note 1. Groups A, B, and D are policy-shock runs. Group C is the counterfactual run without the NJ minimum wage shock. Group E is reported only as a supplementary harness-check run and is not included in the main DDD estimates. Formal analysis records refer to parseable agent-round records. DID/DDD estimates in the table are calculated using non-missing observations for the corresponding variables.

Overall, Case 1 shows that MASS can identify explicit institutional rule changes and translate them into structured managerial decision outputs. The wage variable provides the strongest evidence, the price variable provides auxiliary evidence, and the employment variable indicates that MASS is better suited as a mechanism-rehearsal and pre-experimental simulation tool than as a precise prediction model for real economic effects.

4.3. Case 2: Group Attitudes Under Differentiated Information Exposure

Case 2 uses the Cambridge Analytica controversy and the 2016 UK Brexit referendum digital campaigning context as empirical references to examine whether a targeting mechanism can generate distinguishable group-attitude changes under simulation conditions. This case does not claim that Cambridge Analytica, AggregateIQ, or digital targeting determined the referendum result, nor does it use Cambridge Analytica internal data or real-world polling microdata for causal estimation [36,37]. Real-world materials are used mainly to construct the digital campaigning context, define data analytics and targeted-delivery mechanisms, set the campaign time window, and provide external reference points for voter-cluster ordering. Table 7 provides a descriptive comparison between the official referendum result and the MASS post-period means. The DID results in Table 8 are based on MASS simulation outputs, not on a re-estimation of real referendum survey data.

Table 7. Descriptive comparison between the real referendum result and MASS post-period means in Case 2.

Comparison Indicator	Real Reference Data	MASS Targeting Post-Period Mean	MASS No-Targeting Control Post-Period Mean
Leave share among decided voters	51.90%	51.90%	51.80%
Remain share among decided voters	48.10%	48.10%	48.20%
Turnout/turnout intention	72.20%	70.30%	69.30%
Undecided share	Not applicable	21.80%	26.80%

Note 2. Real reference data are from the official 2016 UK referendum results published by the UK Electoral Commission [38,39]. Leave received 17,410,742 votes, Remain received 16,141,241 votes, and the official turnout was 72.2%. The real Leave/Remain shares are calculated among valid votes as Leave/(Leave + Remain) and Remain/(Leave + Remain). MASS targeting and no-targeting control post-period means are based on the exported Case 2 simulation data. In MASS, the Leave/Remain proportions are re-normalized among decided attitudes as leave_support/(leave_support + remain_support) and remain_support/(leave_support + remain_support). turnout_intent is a simulated voting-intention variable and is not equivalent to real turnout. undecided_share has no corresponding ballot category in the final referendum result. This table is used only for descriptive external alignment and not as evidence of real-world causal estimation or true vote prediction.

Groups A–F are targeting conditions that include data analytics and targeted-communication agents. Groups G and H are no-targeting control conditions used to compare attitude evolution without a precision delivery mechanism. This study defines Round 10 as the beginning of the post-period after the strengthening of targeting and compares Rounds 1–9 with Rounds 10–18.

Table 8. DID results for targeting and no-targeting conditions in Case 2.

Variable	Targeting Pre	Targeting Post	Control Pre	Control Post	DID	95% CI	<i>p</i>
leave_support	0.371	0.404	0.368	0.379	0.022	[−0.005, 0.049]	0.105
remain_support	0.339	0.374	0.317	0.353	−0.001	[−0.034, 0.033]	0.967
undecided_share	0.286	0.218	0.292	0.268	−0.044	[−0.082, −0.005]	0.028
turnout_intent	0.676	0.703	0.667	0.693	0.001	[−0.035, 0.037]	0.944
trust_in_experts	0.496	0.496	0.487	0.505	−0.017	[−0.053, 0.019]	0.352
issue_salience_immigration	0.574	0.6	0.585	0.598	0.014	[−0.048, 0.075]	0.664
issue_salience_economy	0.627	0.661	0.593	0.659	−0.032	[−0.086, 0.023]	0.253

Table 7 shows that the Leave/Remain proportions in the MASS post-period means are descriptively close to the official referendum result. This comparison is only an auxiliary reference. It does not show that MASS predicted the real referendum result, nor does it provide causal evidence that Cambridge Analytica, AggregatIQ, or digital targeting affected the real referendum result. The main methodological evidence in Case 2 still comes from the comparison between targeting and no-targeting conditions.

In Case 2, the most stable condition difference is the decline in undecided_share under the targeting condition. Table 8 shows that the DID estimate for undecided_share is −0.044, $p = 0.028$, indicating that, compared with the no-targeting control, the targeting mechanism more clearly reduces the undecided-group share. The direction of change in leave_support is consistent with expectations, but it does not reach conventional significance. Variables such as remain_support, turnout_intent, trust_in_experts, and issue_salience also do not form stable significant differences. Therefore, the validation evidence for Case 2 should not be described as targeting significantly increasing Leave support. It should instead be understood as “a decline in undecided share plus stable group ordering.”

Known group ordering further supports the behavioral validity of Case 2. Leave-leaning or swing voter groups consistently show higher leave_support than young urban Remain voters, indicating that MASS does not merely generate aggregate average changes but also preserves the group differences expected from the role settings. Together with the significant decline in undecided_share, this result constitutes the main validation evidence for Case 2.

Accordingly, Case 2 shows that MASS can generate distinguishable group-attitude evolution between conditions with and without a digital targeting mechanism. The main evidence in this case comes from the significant decline in undecided_share and the stability of voter-group ordering, rather than from a significant change in any single support-rate variable.

4.4. Case 3: Collective Behavior Under Platform Attention and Governance Feedback

Case 3 uses the public diffusion process of Zibo barbecue on Chinese social media platforms in 2023 as a real-world reference and focuses on whether MASS can simulate the development of platform public opinion under multi-agent interaction. This case does not use real Weibo posts, platform recommendation logs, real individual tourist data, or consumption microdata for model training or causal estimation, and it does not claim to reproduce the real social media network structure. Real-world materials are used mainly to construct the event background, divide attention stages, and set the simulation window. Relevant real-world background can be found in Fang [40], Jiang et al. [41], and Zhang et al. [42]. The Baidu Index search trend for the keyword “Zibo barbecue” [34] is used as an external attention-stage anchor, with the query set to nationwide, PC+mobile, from 1 March to 24 May 2023. The index is used only to show that the simulation rounds cover the main attention cycle of the real event, from low-level warm-up and rapid diffusion to the peak around the May Day holiday and subsequent post-peak decline. It is not directly

numerically fitted to MASS governance variables, sentiment variables, or behavioral-intention variables, and it does not enter the DID, DDD, or mechanism-correlation models.

Groups A-F are independent runs under the active governance condition, and Group G is the passive governance counterfactual condition. Case 3 uses more active-governance runs because, compared with the first two cases, its real platform data are harder to obtain in full, the public-opinion process is more open-ended, the influencing factors are more complex, and the simulation granularity is coarser. Groups A-E were first used to examine whether MASS could produce stable directions in the main mechanism variables and a stage pattern broadly aligned with the Baidu Index attention stages in Table 9. Group F was then added as an additional stability-check run; after its inclusion, the later-period cumulative means and directional relationships of the main mechanism variables did not change substantively. Group G was subsequently used as the passive-governance counterfactual reference. In combination with the external stage alignment in Table 9, Rounds 1–3 are treated as the low-level warm-up and early diffusion stage, while Rounds 4–12 are treated as the later stage from diffusion initiation to governance response and continued diffusion. The analysis focuses on organizational-response variables, public-sentiment variables, visit-intention variables, and directional relationships among them.

Table 9. Correspondence between MASS rounds and Baidu Index attention stages in Case 3.

MASS Rounds	Approximate Dates	Baidu Index Attention Stage	Use in This Study
R1–R3	1 March–21 March 2023	Low baseline and early attention	Initial diffusion stage
R4–R6	22 March–11 April 2023	Pre-surge transition	Transition before large-scale diffusion
R7–R8	12 April–25 April 2023	Rapid rise and high attention	Diffusion acceleration and service-pressure buildup
R9–R10	26 April–9 May 2023	Peak and immediate post-peak decline	Peak attention and governance-response pressure
R11–R12	10 May–24 May 2023	Post-peak decline and long-tail attention	Continued response and stabilization

Note 3. Baidu Index provides an external reference for search-attention trends and helps confirm that the simulation window for Case 3 is broadly aligned with the real-world attention cycle. It does not enter the DID, DDD, or mechanism-correlation models and is not used as real-world causal-test data.

The DID results for Case 3 are shown in Table 10. Overall, compared with the passive governance condition, the active governance condition produces clearer directional differences in organizational response, public sentiment, behavioral intention, and trust in governance. Active governance increases *resource_level*, *channel_effectiveness*, *positive_sentiment*, *visit_intent*, *recommend_intent*, and *trust_in_official_response*, and decreases *compliance_risk* and *negative_sentiment*. The direction for *message_coherence* is positive but does not reach conventional significance; therefore, it is not treated as strong evidence in this case. Accordingly, the main result of Case 3 is not a change in any single variable, but an overall consistent direction of change among organizational response, public sentiment, and behavioral intention under the active governance condition.

Table 11 further shows that some variables in Case 3 form correlations consistent with theoretical expectations. Specifically, *compliance_risk* is significantly positively correlated with *negative_sentiment*, and *positive_sentiment* also shows a significant or near-significant positive relationship with *visit_intent*. The directions for *resource_level* with *trust_in_official_response* and *channel_effectiveness* with *positive_sentiment* are positive, but they do not reach conventional significance. These results indicate mechanism-chain consistency within the simulation process and should not be interpreted as causal estimates of real public-opinion diffusion or tourism-consumption outcomes.

Table 10. DID results for active governance and passive governance conditions in Case 3.

Variable	Active Early	Active Later	Passive Early	Passive Later	DID	95% CI	<i>p</i>	Round 12 Difference
message_coherence	0.615	0.740	0.599	0.699	0.025	[−0.014, 0.065]	0.208	0.043
resource_level	0.567	0.569	0.548	0.506	0.044	[0.017, 0.071]	0.001	0.089
channel_effectiveness	0.612	0.760	0.594	0.671	0.071	[0.038, 0.104]	<0.001	0.110
compliance_risk	0.275	0.288	0.297	0.351	−0.041	[−0.072, −0.011]	0.008	−0.065
positive_sentiment	0.446	0.465	0.433	0.336	0.115	[0.033, 0.198]	0.006	0.182
negative_sentiment	0.258	0.341	0.273	0.456	−0.100	[−0.159, −0.042]	0.001	−0.174
visit_intent	0.498	0.488	0.507	0.371	0.127	[0.091, 0.164]	<0.001	0.170
recommend_intent	0.487	0.487	0.483	0.355	0.127	[0.088, 0.167]	<0.001	0.188
trust_in_official_response	0.462	0.507	0.449	0.328	0.166	[0.113, 0.219]	<0.001	0.262

Table 11. Mechanism-chain correlations in Case 3 active-governance runs.

Mechanism Pair	n Rounds	Pearson <i>r</i>	<i>p</i>	Spearman rho	<i>p</i>
compliance_risk → negative_sentiment	12	0.766	0.004	0.762	0.004
positive_sentiment → visit_intent	12	0.669	0.017	0.559	0.059
resource_level → trust_in_official_response	12	0.483	0.112	0.573	0.051
channel_effectiveness → positive_sentiment	12	0.415	0.179	0.455	0.138

Note 4. Correlations are based on round-level aggregate data from active governance runs. These results are used to interpret mechanism consistency in the simulation process and are not real-world causal estimates.

Overall, Case 3 shows that MASS can generate public-opinion process records of round-by-round interaction among multiple types of agents in a public-information environment. Under the active governance condition, organizational response, public sentiment, trust in governance, and behavioral intention form a relatively clear directional chain. This indicates that MASS can use the public board and structured outputs to capture processual outcomes generated through multi-agent interaction. This result should still be limited to evidence for public-opinion process simulation and mechanism-chain testing, rather than a direct reproduction of real social media data.

4.5. Cross-Case Validation Summary

The three cases provide different types of evidence for MASS reliability and validity. Case 1 mainly supports rule-shock identification: the minimum wage rule can enter managers' starting-wage decisions, while the counterfactual group without the policy shock does not automatically generate the same change pattern. Case 2 mainly supports information-intervention differentiation: the targeting condition significantly reduces undecided_share and preserves voter-group ordering. Case 3 mainly supports multi-agent process simulation: active governance and passive governance form distinguishable directional differences in organizational response, public sentiment, trust in governance, and behavioral intention. Overall, the three cases show that protocolized LLM-driven ABM can generate auditable, comparable, and empirically assessable simulation outputs in policy-shock, information-intervention, and public-opinion scenarios. However, this evidence remains pre-experimental simulation evidence and should not be interpreted as precise prediction of real-world outcomes or an estimation of real causal effects.

5. Discussion

5.1. LLM-Driven ABM as a Quantitative Research Method

The first methodological contribution of this study is to organize LLM-driven ABM as a social simulation method that can facilitate quantitative analysis. Naive LLM-agent simulation usually produces natural-language narratives. Although such narratives can

present role actions and situational judgments, they are difficult to use directly for statistical comparison, process auditing, and reliability-and-validity assessment. The basic idea of MASS is to transform each agent's response in each round into an agent-round observation so that simulation outputs simultaneously contain role, time, condition, action, public information, state variable, and validation information.

The foundation of this method is to embed the language-generation capacity of LLMs into the experimental structure of ABM. An LLM is not treated as a substitute for a real individual or real sample, but as a behavior generator constrained by role, information boundaries, background rules, and output schema. Through role settings, round scheduling, rule control, and structured outputs, researchers fix model-generated actions as recordable, comparable, and reviewable observation units. As a result, LLM-agent outputs are no longer merely text; they can enter DID, DDD, pre-post comparisons, group ordering, correlation analysis, and mechanism-chain tests.

The three cases demonstrate different forms of this quantitative transformation. Case 1 uses agent-round observations to test changes in managerial decision-making under a rule shock. Case 2 compares group-attitude differences between targeting and no-targeting conditions. Case 3 compares organizational response and public response under active governance and passive governance conditions. These analyses show that the methodological value of MASS does not lie in directly equating LLM outputs with real human behavior, but in transforming a constrained behavioral generation process into auditable, reproducible, and empirically assessable simulation data. Thus, MASS is better suited as a tool for mechanism rehearsal, counterfactual comparison, risk identification, and research-design preparation than as a substitute for real experiments, natural experiments, surveys, interviews, or platform-data analysis.

5.2. Methodological Contribution: From Narrative Role-Playing to Protocolized Evidence

The second methodological contribution of this study is to transform common failure modes in naive LLM-agent simulation into executable research-control procedures. Naive simulation is prone to temporal-order errors, role-identity deviation, data contamination, rule-or-condition errors, outputs without structured fields, invalid numerical values, and reproduction difficulties. The significance of MASS does not lie in simply increasing the number of agents or expanding the complexity of fixed scenarios, but in translating these risks into control items in the research design and constraining them through the scheduling, role, information, rule, output, check, and log mechanisms described in Section 3.2.

This point also distinguishes MASS from existing AI social simulation platforms. Existing platforms have demonstrated the capabilities of LLM agents in small-scale social interaction, medical workflows, social media diffusion, and large-scale social-life simulation. This study does not focus on a specific platform scenario itself, but on a transferable research protocol: how researchers can set roles, rules, information boundaries, experimental conditions, counterfactual conditions, and validation indicators around specific research questions, and how they can transform LLM-agent outputs into auditable, reproducible, and empirically assessable data.

Therefore, the validity of MASS does not automatically come from the LLM itself; it depends on how researchers design and disclose the simulation protocol. Role settings affect behavioral boundaries, information boundaries affect contamination risk, background rules affect variable control, output schemas affect statistical analysis, and logs and reproducibility materials affect third-party review. Only when these elements are clearly recorded and incorporated into the research design can LLM-driven ABM move from narrative role-playing to a methodologically meaningful tool for social simulation research.

5.3. Cross-Case Evidence and Boundary Conditions

The three cases jointly address RQ3, but they provide different types of reliability-and-validity evidence. The strength of Case 1 lies in rule-shock identification, showing that explicit institutional rules can enter agents' structured decision outputs; its boundary is that the employment variable is unstable and cannot be interpreted as a reproduction of real employment effects. The strength of Case 2 lies in information-intervention differentiation, showing that the targeting condition can reduce undecided_share and preserve group ordering; its boundary is that leave_support and most issue-salience variables do not form stable significant results and cannot be interpreted as evidence that digital targeting changed the real referendum result. The strength of Case 3 lies in multi-agent process simulation, showing that MASS can generate directional relationships among governance response, public sentiment, trust, and behavioral intention in a public-information environment; its boundary is that the Baidu Index is only a public attention reference, and correlation analysis cannot replace causal validation with real platform data.

Accordingly, the three cases provide a "limited but complementary" multi-evidence test of reliability and validity. Case 1 examines whether rules can stably enter decision-making, Case 2 examines whether information conditions can distinguish group-attitude changes, and Case 3 examines whether multi-agent interaction can form an interpretable process chain. Together, the three cases show that protocolized LLM-driven ABM can generate auditable, comparable, and empirically assessable simulation outputs under different social mechanisms. However, these outputs should still be understood as pre-experimental simulation evidence and should not be interpreted as estimates of real-world causal effects or predictions of real-world outcomes.

5.4. Limitations, Ethical Boundaries, and Future Research

This study still has several limitations. First, LLM-driven ABM is highly sensitive to prompts, role settings, model versions, and parameter settings. The granularity of role cards, the way information is presented, the completeness and executability of background rules, output schemas, temperature settings, and model-provider environments may all affect simulation trajectories. This study uses prompt manifests, replication manifests, local information boundaries, harness checks, and local checks to reduce these risks, but it cannot claim to have fully eliminated prompt sensitivity or model-version dependence. Future research should systematically conduct prompt robustness tests, sensitivity analyses, cross-model replication, and cross-vendor comparisons to examine the stability of results across different prompts, models, and runtime environments.

Second, the number of independent runs in this study remains limited. The main analysis of Case 1 relies on the historical benchmark data from the four early Groups A-D. Although Case 2 and Case 3 already form comparisons between treatment and counterfactual conditions, the estimation precision for weaker or more volatile variables such as leave_support, issue_salience, employment, and message_coherence may still be affected by the number of independent replications, the number of agents, model-output variance, and the structure of condition groups. This study does not interpret non-significant results as "simulation failure," nor does it simply attribute them to an insufficient number of runs. A more cautious interpretation is that the current results already support several directional and mechanism-based judgments, but the stability of weak-effect variables still requires more independent runs for further testing. Future research can increase independent replications and conduct bootstrap analysis, multi-run aggregation, cross-model replication, and sensitivity analysis to further assess whether these variables have stable directions and reviewable effect ranges.

Third, model outputs require continuous calibration. MASS is not a model that naturally approaches reality after a single setup. The fluctuation of the employment variable in Case 1, the lack of significance for most attitude variables in Case 2, and the simplified platform mechanism and public-information diffusion rules in Case 3 all show that external data, expert judgment, preliminary trial runs, and error analysis remain important components of model calibration. Future research should further introduce more fine-grained real-world data, platform data, interview materials, expert coding, and event-timeline materials to calibrate agent profiles, information-exposure conditions, state-variable ranges, and mechanism-chain assumptions, thereby improving the external validity and structural validity of the model.

Fourth, the risks of training-corpus priors and case contamination remain difficult to eliminate completely. The NJ-PA minimum wage case, Brexit, and Zibo barbecue may all exist in model training corpora or public knowledge. This study reduces this risk through local information boundaries, case blinding, counterfactual groups, hidden-field restrictions, final-outcome shielding, and contamination checks, but it cannot claim to fully eliminate model priors. Especially for classic cases, an LLM may already have learned relevant public narratives. Future research can use synthetic cases, background rewriting, case blinding, event renaming, cross-model replication, and counterfactual perturbation to further test whether simulation results depend on existing case memory rather than being generated under role, rule, and local-information constraints.

Fifth, the evidence type of MASS should still be limited to pre-experimental simulation evidence. The DID, DDD, group ordering, and mechanism-chain tests in this study are all based on synthetic agent-round observations generated by MASS, rather than on direct causal estimation using real-population data. The results of Case 1 cannot be interpreted as a precise reproduction of real employment effects, the results of Case 2 cannot be interpreted as the causal impact of Cambridge Analytica or AggregateIQ on the real referendum result, and the results of Case 3 cannot be interpreted as numerical fitting to the Baidu Index or real tourism-consumption outcomes. The main value of MASS lies in mechanism rehearsal, counterfactual comparison, risk identification, and research-design assistance, rather than in replacing real experiments, natural experiments, surveys, interviews, or platform-data analysis.

The real platform data for Case 3 are difficult to reproduce fully and publicly. Therefore, this study uses it only as a public-opinion process simulation case with a publicly available event reference. Its results are mainly used to test the counterfactual difference between active governance and passive governance, and the directional relationships among organizational response, public sentiment, trust, and behavioral intention. They cannot be interpreted as causal estimates of real platform-diffusion effects or tourism-consumption outcomes.

Finally, the use of MASS requires clear ethical boundaries. This study positions MASS as a pre-experimental simulation method rather than a real-world predictor or automated decision system. Simulation outputs should not be directly used for real policy adjudication, voter manipulation, individual profiling, commercial discrimination, public-opinion manipulation, or unvalidated population interventions. Especially in political communication, public governance, and platform public-opinion scenarios, researchers need to disclose model versions, prompt designs, data sources, validation strategies, and limitations, and avoid packaging simulation results as real-world causal conclusions. The responsible use of MASS is to treat it as a tool for mechanism exploration, risk anticipation, and research-design preparation, and to interpret its results together with real-world evidence, expert review, and public deliberation.

6. Conclusions

This study proposes and evaluates Multi-Agent Social Simulation as an operational form of protocolized LLM-driven ABM. Through role settings, round-based scheduling, information boundaries, background-rule control, structured outputs, harness checks, reason-action logs, and replication manifests, MASS transforms the open-ended generative capacity of LLM agents into recordable, auditable, and statistically analyzable agent-round observations.

With respect to RQ1, this study argues that LLM-driven ABM can serve as a pre-experimental social-behavioral simulation method in behavioral science. However, its methodological position should be mechanism rehearsal, counterfactual comparison, risk identification, and research-design assistance, rather than a replacement for real-world experiments or causal identification. With respect to RQ2, this study shows that naive LLM-agent simulation is prone to temporal-order errors, role-identity deviation, data contamination, rule-or-condition errors, outputs without analyzable fields, invalid numerical values, and replication difficulties. These problems require protocolized control through role cards, information boundaries, background rules, experimental conditions, output schemas, harness checks, local checks, and replication manifests. With respect to RQ3, the three cases provide preliminary cross-scenario evidence: Case 1 supports rule-shock identification, Case 2 supports information-intervention differentiation, and Case 3 supports public-opinion evolution simulation under multi-agent interaction in a public-information environment.

Overall, the value of MASS does not lie in claiming that it can precisely predict real-world outcomes. Rather, it lies in embedding the contextual generative capacity of LLM agents into a controllable, reviewable, and statistically analyzable ABM structure. The three cases show that protocolized LLM-driven ABM can generate empirically assessable simulation outputs in policy-shock, information-intervention, and public-opinion scenarios. Future research should continue to strengthen prompt robustness, independent-run replication, external-data calibration, cross-model comparison, case blinding, and open reproducibility so that LLM-driven ABM can move from an exploratory tool toward a more mature research method for the social and behavioral sciences.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/ai7080279/s1>. The Supplementary Materials include the versioned scenario configuration files for the three simulation cases, the prompt manifest, the replication manifest, exported CSV/XLSX datasets, Python 3.12 statistical analysis scripts, non-sensitive audit summaries, and Supplementary Figure S1. Supplementary Figure S1 documents the Baidu Index search trend for the keyword “Zibo barbecue” under nationwide, PC+mobile, 1 March–24 May 2023 conditions. This figure is used only to document the external search-attention trend and event-stage alignment for Case 3 and is not included in DID, DDD, or mechanism-correlation analyses. Materials involving API keys, account information, platform access credentials, non-public runtime identifiers, or information that may expose security-sensitive runtime environments are not publicly released.

Author Contributions: X.H.: conceptualization, methodology development, research design, MASS system requirement design and functional planning, scenario configuration and case design, data curation, formal analysis, result interpretation, writing—original draft preparation, writing—review and editing, and visualization. Y.S.: research supervision, research framework review, manuscript review, revision suggestions, and academic oversight. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve real human-subject experiments, human biological materials, animal experiments, or direct interventions on

real-world populations. The real-world reference materials used in this study were mainly derived from published literature, official reports, public platform indices, or publicly available aggregated data. The formal statistical analyses were based on synthetic simulation outputs generated by the MASS system.

Informed Consent Statement: Not applicable. This study did not involve real human participants and did not collect individual-level survey, interview, experimental, or personally identifiable information.

Data Availability Statement: The simulation datasets, scenario configuration files, prompt manifest, replication manifest, and statistical analysis scripts supporting the reported results will be made available as Supplementary Materials or through an external public repository upon publication, where permitted. The Case 1 Groups A–D data are historical benchmark simulation outputs generated by an early version of MASS and were previously used in the authors' earlier conference paper; their version provenance and reuse relationship will be identified in the replication manifest and Supplementary Materials. Some runtime logs are not publicly released because they may contain API configuration details, account-related information, platform access credentials, non-public runtime identifiers, or other sensitive operational information. For peer review or replication purposes, restricted materials may be made available from the corresponding author upon reasonable request after API keys, account credentials, and security-sensitive information have been removed.

Acknowledgments: This article is a revised and expanded version of the paper entitled “Multi-Agent Social Simulation: An Experimental Framework for Language-Native Social Experiments” [35], which was presented at the Open Conference of AI Agents for Science 2025, Virtual Conference, 22 October 2025. During the research and writing process, the authors used generative AI tools to assist with language polishing, structural checking, English translation, code debugging, expression refinement, and discussion of presentation clarity. Large language models were also used as constrained behavior generators inside the MASS simulation environment, as described in Section 3. The project requirements, functional planning, case design, running rules, validation logic, and statistical analysis plan of the MASS system were proposed and approved by the first author. Some code writing, code refactoring, error diagnosis, and debugging processes were completed with the assistance of generative AI tools and coding agents. The first author was responsible for specifying the functional requirements of AI-generated or AI-assisted code, testing its operation, verifying results, exporting data, conducting statistical analyses, and taking responsibility for the final code use, simulation design, data processing, and result interpretation. All research designs, scenario configurations, statistical analyses, result interpretations, manuscript structure, and final manuscript content were reviewed and verified by the authors, who take full responsibility for the work. Generative AI tools were not listed as authors and do not bear research responsibility, authorship responsibility, or academic responsibility.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Haney, C.; Banks, W.C.; Zimbardo, P.G. Interpersonal dynamics in a simulated prison. *Int. J. Criminol. Penol.* **1973**, *1*, 69–97.
2. Zimbardo, P.G. *The Lucifer Effect: Understanding How Good People Turn Evil*; Random House: New York, NY, USA, 2007.
3. Angrist, J.D.; Pischke, J.-S. *Mostly Harmless Econometrics: An Empiricist's Companion*; Princeton University Press: Princeton, NJ, USA, 2009.
4. Dunning, T. *Natural Experiments in the Social Sciences: A Design-Based Approach*; Cambridge University Press: New York, NY, USA, 2012.
5. Imbens, G.W.; Rubin, D.B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*; Cambridge University Press: New York, NY, USA, 2015.
6. Card, D.; Krueger, A.B. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *Am. Econ. Rev.* **1994**, *84*, 772–793.
7. Card, D.; Krueger, A.B. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Reply. *Am. Econ. Rev.* **2000**, *90*, 1397–1420. [[CrossRef](#)]
8. Mayer, I.S. The gaming of policy and the politics of gaming: A review. *Simul. Gaming* **2009**, *40*, 825–862. [[CrossRef](#)]

9. Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7280–7287. [CrossRef] [PubMed]
10. Epstein, J.M. Agent-based computational models and generative social science. *Complexity* **1999**, *4*, 41–60. [CrossRef]
11. Gilbert, N.; Troitzsch, K.G. *Simulation for the Social Scientist*, 2nd ed.; Open University Press: Maidenhead, UK, 2005.
12. Railsback, S.F.; Grimm, V. *Agent-Based and Individual-Based Modeling: A Practical Introduction*, 2nd ed.; Princeton University Press: Princeton, NY, USA, 2019.
13. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023; pp. 1–22. [CrossRef]
14. Li, J.; Lai, Y.; Li, W.; Ren, J.; Zhang, M.; Kang, X.; Wang, S.; Li, P.; Zhang, Y.-Q.; Ma, W.; et al. Agent Hospital: A simulacrum of hospital with evolvable medical agents. *arXiv* **2024**, arXiv:2405.02957. [CrossRef]
15. Yang, Z.; Zhang, Z.; Zheng, Z.; Jiang, Y.; Gan, Z.; Wang, Z.; Ling, Z.; Chen, J.; Ma, M.; Dong, B.; et al. OASIS: Open agent social interaction simulations with one million agents. *arXiv* **2024**, arXiv:2411.11581. [CrossRef]
16. Piao, J.; Yan, Y.; Zhang, J.; Li, N.; Yan, J.; Lan, X.; Lu, Z.; Zheng, Z.; Wang, J.Y.; Zhou, D.; et al. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv* **2025**, arXiv:2502.08691. [CrossRef]
17. Grimm, V.; Berger, U.; Bastiansen, F.; Eliassen, S.; Ginot, V.; Giske, J.; Goss-Custard, J.; Grand, T.; Heinz, S.K.; Huse, G.; et al. A standard protocol for describing individual-based and agent-based models. *Ecol. Model.* **2006**, *198*, 115–126. [CrossRef]
18. Grimm, V.; Railsback, S.F.; Vincenot, C.E.; Berger, U.; Gallagher, C.; DeAngelis, D.L.; Edmonds, B.; Ge, J.; Giske, J.; Groeneveld, J.; et al. The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *J. Artif. Soc. Soc. Simul.* **2020**, *23*, 7. [CrossRef]
19. Gao, C.; Lan, X.; Li, N.; Yuan, Y.; Ding, J.; Zhou, Z.; Xu, F.; Li, Y. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1259. [CrossRef]
20. Larooij, M.; Törnberg, P. Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. *Artif. Intell. Rev.* **2026**, *59*, 15. [CrossRef] [PubMed]
21. Schelling, T.C. Dynamic models of segregation. *J. Math. Sociol.* **1971**, *1*, 143–186. [CrossRef]
22. Simon, H.A. A behavioral model of rational choice. *Q. J. Econ.* **1955**, *69*, 99–118. [CrossRef]
23. Argyle, L.P.; Busby, E.C.; Fulda, N.; Gubler, J.; Rytting, C.; Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Anal.* **2023**, *31*, 337–351. [CrossRef]
24. Aher, G.V.; Arriaga, R.I.; Kalai, A.T. Using large language models to simulate multiple humans and replicate human subject studies. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 337–371.
25. Qin, X.; Huang, M.; Ding, J. AITurk: Using ChatGPT for social science research. *SSRN* **2024**. [CrossRef]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; p. 30.
27. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
28. Anadkat, S. How to Make Your Completions Outputs Consistent with the New Seed Parameter. OpenAI Cookbook. 2023. Available online: https://developers.openai.com/cookbook/examples/reproducible_outputs_with_the_seed_parameter/ (accessed on 20 July 2026).
29. Guala, F. Models, simulations, and experiments. In *Model-Based Reasoning: Science, Technology, Values*; Magnani, L., Nersessian, N.J., Eds.; Kluwer Academic/Plenum Publishers: New York, NY, USA, 2002; pp. 59–74.
30. Morgan, M.S. Experiments without material intervention: Model experiments, virtual experiments, and virtually experiments. In *The Philosophy of Scientific Experimentation*; Radder, H., Ed.; University of Pittsburgh Press: Pittsburgh, PA, USA, 2003; pp. 216–235.
31. Sugden, R. Credible worlds: The status of theoretical models in economics. *J. Econ. Methodol.* **2000**, *7*, 1–31. [CrossRef]
32. House of Commons Digital, Culture, Media and Sport Committee. Disinformation and “Fake News”: Final Report. UK Parliament. 2019. Available online: <https://publications.parliament.uk/pa/cm201719/cmselect/cmcomeds/1791/1791.pdf> (accessed on 20 July 2026).
33. Information Commissioner’s Office. Investigation Into the Use of Data Analytics in Political Campaigns: A Report to Parliament. 2018. Available online: <https://ico.org.uk/media2/migrated/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf> (accessed on 20 July 2026).
34. Baidu Index. Search Index Data for the Keyword “Zibo Barbecue”, 1 March–24 May 2023, Nationwide, PC+Mobile [Data Set]. Available online: <https://index.baidu.com/> (accessed on 27 May 2026).
35. Hu, X.; Shen, Y.; You, K. Multi-Agent Social Simulation: An experimental framework for language-native social experiments. In *Proceedings of the Open Conference of AI Agents for Science 2025, Virtual*, 22 October 2025; OpenReview: Amherst, NH, USA, 2025. Available online: <https://openreview.net/forum?id=emsnDnmtYP> (accessed on 20 July 2026).

36. Kalla, J.L.; Broockman, D.E. The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *Am. Political Sci. Rev.* **2018**, *112*, 148–166. [CrossRef]
37. Rone, J. Beyond Brexit? Public participation in decision-making on campaign data during and after referendum campaigns. *Media Commun.* **2023**, *11*, 6200. [CrossRef]
38. Electoral Commission. Report on the 23 June 2016 Referendum on The UK's Membership of the European Union. 2016. Available online: <https://www.electoralcommission.org.uk/research-reports-and-data/our-reports-and-data-past-elections-and-referendums/report-23-june-2016-referendum-uks-membership-european-union> (accessed on 20 July 2026).
39. Electoral Commission. Results and Turnout at the Eu Referendum. Available online: <https://www.electoralcommission.org.uk/research-reports-and-data/our-reports-and-data-past-elections-and-referendums/results-and-turnout-eu-referendum> (accessed on 27 May 2026).
40. Fang, Z. What Changed and What Remained in Zibo After the Traffic Surge. *China City News*, 22 May 2023. Available online: https://paper.people.com.cn/zgcsb/html/2023-05/22/content_25987426.htm (accessed on 20 July 2026).
41. Jiang, D.; Li, N.; Jia, X. Zibo Barbecue, Full of the Flavor of Rule of Law. *Rule of Law Daily*, 9 May 2023. Available online: https://www.moj.gov.cn/pub/sfbgw/jgsz/jgszjgtj/jgtjfdy/fzdyjtjxw/202305/t20230509_478371.html (accessed on 20 July 2026).
42. Zhang, Y.; Lu, Z.; Li, Z.; Tan, J.; Hou, K. Electronic tourism transformation and its coordination with locality: The case of Zibo barbecue triggered by social media. *Hum. Geogr.* **2025**, *40*, 184–192. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.