From Script to Stage: Automating Experimental Design for Social Simulations with LLMs

Yuwei Guo[1,2,3], Zihan Zhao[1,2,3], Xiaowei Liu[1,2,3], Xiangning Yu[1,2,3], Qun Ma[1,2,3], Deyu Zhou[4], Xiao Xue[1,2,3]*

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin Key Laboratory of Healthy Habitat and Smart Technology, Tianjin, China
[3]Laboratory of Computation and Analytics of Complex Management Systems, Tianjin University, Tianjin, China
[4]School of Software, Shandong University, Shandong, China

{2024244171, zhaozihan, xiaoweiliu, yxn9191, 1023244018, jzxuexiao}@tju.edu.cn, 202220787@mail.sdu.edu.cn

Abstract

Multi-agent simulation based on LLMs has increasingly emerged as a new paradigm for exploring complex social phenomena and validating theoretical hypotheses. However, traditional experimental design in the social sciences relies heavily on interdisciplinary expert knowledge, involving cumbersome procedures and high technical barriers. While LLM-driven agents demonstrate broad prospects for designing experiments, their limitations regarding reliability and scientific rigor continue to significantly hinder their in-depth application in social science research. To address these challenges, this paper proposes **FSTS**, an automated framework for multi-agent experiment design based on script generation. Drawing on the concept of the "Decision Theater," the framework deconstructs experimental design into three core phases: **S**cript **C**omposition, **S**cript **F**inalization, and **A**ctor **G**eneration. Tests across multiple scenarios indicate that the agents generated by this framework can enact the script within the "experimental theater," reproducing results consistent with real-world situations. The proposal of FSTS not only effectively lowers the barrier for social science experimental design but also provides scientifically grounded decision support for policy-making. [1]

1 Introduction

In recent years, the rapid development of artificial intelligence technology, especially the widespread application of Large Language Models (LLMs), has driven a surge of research interest in "AI for Social Science(Xu et al., 2024)." AI tools can not only alleviate the workload of researchers but also provide intelligent support in stages such as experimental design (Xue et al., 2024b), variable selection, and scheme optimization.

In social science, **computational experiments** (Xue et al., 2021a, 2024a) are often viewed as a "third paradigm" for studying complex systems in controllable virtual environments, yet adoption is limited by technical and interdisciplinary barriers. Decision-Theater-style approaches improve rigor through structured, participatory workflows (Wolf et al., 2023) but require expert facilitation and substantial coordination, which hinders scalability (Jaeger and Laubichler, 2026). Meanwhile, LLM-based agents offer flexibility (Tran et al., 2025) but are unreliable for rigorous experimental design due to hallucinations and limited verifiability (Qin et al., 2023; Messeri and Crockett, 2024; Huang et al., 2025; Chelli et al., 2024).

Recent LLM-driven social simulation systems and environments (e.g., Smallville (Park et al., 2023; Xue et al., 2023b), SOTOPIA (Zhou et al., 2023, 2025), and AgentSociety (Piao et al., 2025)) demonstrate promising progress in scalable interaction and evaluation. However, they largely leave open how to systematically compile high-level research requirements into standardized, inspectable, and executable experimental scripts with explicit control over variables, interventions, and evaluation criteria (Gao et al., 2024; Xue et al., 2021b).

A central bottleneck is the generation of **experimental scripts**: structured, executable specifications that translate high-level research questions into concrete experimental settings (e.g., environments, variables, interventions, and evaluation criteria). Despite recent progress, generating scripts that satisfy user requirements remains challenging for three reasons:

- **Standardized yet expressive script specification.** Scripts must be machine-executable and well-structured, while remaining expressive enough to capture diverse evolutionary trajectories and plausible social dynamics (Liu et al., 2024b,a).

*Corresponding author

[1]The code repository for this project is publicly available at https://github.com/RisingDate/FSTS.

Findings of the Association for Computational Linguistics: ACL 2026, pages 15897–15916
July 2-7, 2026

- **Combinatorial complexity in experimental design.** Realistic social systems exhibit multi-factor dependencies, creating a trade-off between combinatorial explosion (an intractably large design space) and oversimplification (loss of validity).
- **Limited fidelity of agent modeling.** Many existing frameworks emphasize dialogue- or task-oriented behaviors, while richer agent attributes, relational structures, and mechanism-level assumptions are often under-modeled.

To address these challenges, we propose **FSTS**, an automated framework for experimental design for artificial-society simulation and reasoning. Inspired by film production, FSTS casts experimental design as a three-stage pipeline: (1) **Script Composition**, where a Screenwriter Agent generates multiple candidate experimental schemes from user requirements; (2) **Script Finalization**, where a Director Agent evaluates candidates along multiple dimensions and selects a final scheme; and (3) **Actor Generation**, where an Actor Factory instantiates experimental agents with attributes and relationship networks specified by the script. By introducing domain-specialized LLM agents and explicit intermediate artifacts at each stage, FSTS systematizes and automates the experimental design workflow while maintaining interpretability and user control.

The contributions of this paper are threefold:

- We introduce **FSTS**, an LLM-centric framework that lowers the barrier to designing computational experiments for artificial-society simulation by bridging high-level research intent and executable experimental scripts.
- We propose a **film-inspired multi-agent workflow** (Screenwriter–Director–Actor Factory) that separates generation, critique/selection, and instantiation, improving the structure and reliability of script generation.
- We conduct an **empirical evaluation** with quantitative metrics, human judgments, and ablation studies, showing that FSTS produces higher-quality experimental scripts and agents than strong baselines.

2 Related Work

This section reviews related work on computational experiments. We focus on two lines of research that enable automated experimental design: requirement analysis methods (Casalicchio and Cotumaccio, 2024; Santos et al., 2024) and emerging LLM-based multi-agent frameworks (Chen et al., 2024; Hong et al., 2024; Yang et al., 2023).

2.1 Traditional Techniques

Before the rise of computational social science and Large Language Models, decision-making and modeling in social science largely relied on traditional simulation- and facilitation-based approaches (Xue et al., 2023c).

Computational experiments (Wang, 2004a,b; Xiao et al., 2023; Xue et al., 2023a) use computer modeling and simulation to quantitatively analyze complex social systems in a virtual environment, providing a technical basis for modeling and deduction (Peng et al., 2023). Built on visualization and simulation technologies, the **Decision Theater** (Wolf et al., 2023) offers an integrated software–hardware setting that interactively presents decision schemes to support collective decision-making.

Compared with traditional Agent-Based Modeling (ABM) (Macal and North, 2009, 2005), the Decision Theater combines large-screen visualization with roundtable deliberation, improving cognitive consistency and decision effectiveness. Nevertheless, it remains limited by (i) high professional requirements and cross-domain expert participation, (ii) dependence on on-site facilitation with complex and costly procedures, and (iii) persistent "Big Data–Small Data" issues, a relatively single decision-making paradigm, and limited human–machine collaboration.

2.2 The Era of LLM agent

LLM-based agent simulation has become a major direction (Piao et al., 2025; Gurcan, 2024; Tang et al., 2025; Ma et al., 2024). Table 5 compares currently popular agent frameworks. Relative to ABM, it lowers modeling barriers via natural language, enables richer planning and coordination, and introduces stochasticity. However, it remains bottlenecked by prompt-heavy generation without active validation and by hallucinations and bias, which undermine rigor and credibility (Mao et al., 2025; Zhang et al., 2025).

2.3 Limitations of Existing Approaches

Traditional computational experiments and Decision-Theater-style workflows are rigorous but

Figure 1: **Overview of FSTS.** First, the Screenwriter Agent generates multiple candidate scripts. Next, the Director Agent reviews each script and selects the final version. Subsequently, the Actor Factory generates Actor Agents to perform on the stage, and the performance results are fed back to the user for experimental optimization.

costly and hard to scale, whereas LLM-driven approaches are accessible but difficult to trust for high-stakes design. To reconcile this trade-off, we propose **FSTS**, a **Screenwriter–Director–Actor** framework that closes the loop from user requirements to experimental enactment while balancing rigor, credibility, and usability.

3 Methodology

Drawing inspiration from the film production workflow, we divide the process of automated experimental design into three key stages: **Script Composition**, **Script Finalization**, and **Actor Generation**. As shown in Figure 1, the framework starts with user requirements and finally generates Actor Agents capable of performing in the designated theater.

3.1 Script Composition

The design of the experimental script aims to capture the non-linear evolutionary characteristics of real-world events. As shown in Figure 2, the progression of events is replete with bifurcations and turning points: the choices made by agents at critical nodes determine subsequent trajectories, triggering distinct evolutionary branches and leading to diverse outcomes. Furthermore, the experimental script must be capable of simulating complex causal chains, allowing different evolutionary paths to reconverge at similar or identical final states. The precise modeling of this path dependence and system convergence constitutes a core requirement of this method.

Drawing inspiration from the industrial pipeline mechanism (Sivasankaran and Shahabudeen, 2014) characterized by task decomposition and hierarchical collaboration, this paper proposes a hierarchical experiment script generation architecture. This architecture deconstructs complex experimental design tasks into standardized sub-modules, achieving the mapping from macro goals to micro parameters through multi-agent collaboration. We formally define the Experiment Script as a 5-tuple:

$$S =< G, I, R, D, L > \quad (1)$$

Where G represents the Experiment Goal, I represents Influencing Factors, R represents Response Variables, D denotes parameterized Design Points, and L represents the Storyline that runs through the experimental logic.

To improve the rationality and accuracy of the LLM output, the framework introduces a dual con-

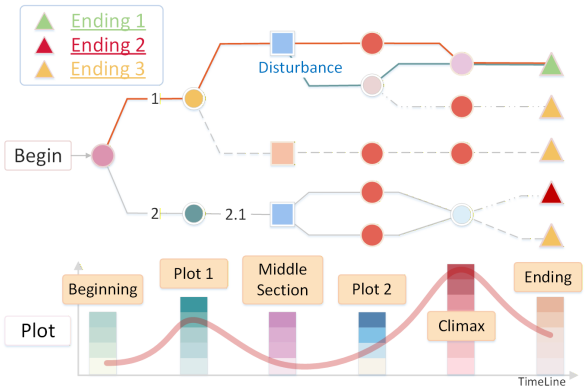

Figure 2: **Dynamic Mechanism of Script Evolution.** Non-linear bifurcation triggered by key decisions and the ultimate convergence of states.

straint mechanism on both the input and output sides:

- **Input Side: Structured Guidance and Step-wise Generation.** We constrain the model's inference space using Domain Knowledge Templates, requiring user input to include three elements: research goal, core variables, and target objects. Simultaneously, a Step-wise Generation strategy is adopted to decompose the construction process of script S into progressive sub-tasks targeting G, I, R, D and L, ensuring local semantic accuracy.

- **Output Side: Multi-View Integration and Format Constraints.** A Multi-View Generation mechanism is introduced, where Screenwriter Agents generate candidate schemes from different perspectives (e.g., goal-oriented vs. process-oriented), and the optimal solution is selected via weighted voting. Additionally, strict JSON Schema Constraints are applied to ensure that the generated scripts can be directly parsed and executed by downstream simulation systems.

As illustrated in Figure 3, this framework establishes a collaborative **Screenwriter-Director** system for experimental script generation. At the generation end, the Screenwriter Agent (GPT-4o) (OpenAI, 2024) is responsible for parsing user requirements and integrating domain knowledge to construct a five-dimensional preliminary script, encompassing elements such as experimental goals and storylines. For the critical Design Points (D), we adopt Design of Experiments (DOE) methods to optimize complexity (Xue et al., 2024b), aiming to obtain complete experimental conclusions with as few experimental schemes as possible. At the monitoring end, four Director Agents (GPT-5 mini) (OpenAI, 2025) are deployed to conduct itemized reviews focusing on the rationality and standardization of the script. The system employs a Cascading Validation mechanism: the review process adheres to a strict sequence, where the workflow transitions to the subsequent stage only after the preceding module has passed validation. Once a defect is detected, a feedback closed-loop is immediately triggered, guiding the Screenwriter Agent to perform targeted revisions on specific modules based on the Director's feedback.

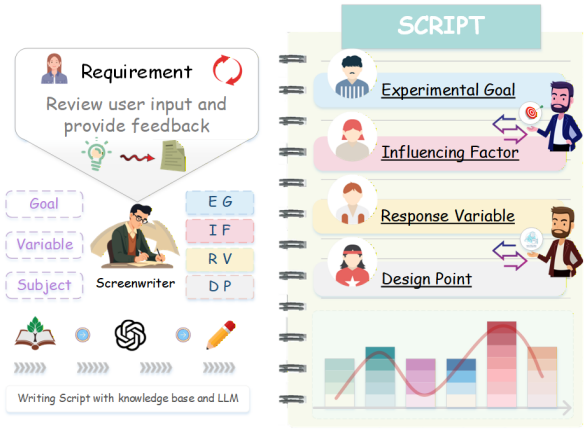

Figure 3: **Schematic Diagram of Script Generation.** The Screenwriter composes scripts by leveraging a knowledge base.

3.2 Script Finalization

In the Script Composition phase, the Screenwriter Agent generates multiple candidate experimental scripts from different perspectives based on user requirements. Obtaining candidate scripts requires the further selection of a unique experimental plan for execution. Manual screening is not only costly and prone to subjective bias but also contradicts the objective of automated experimental design proposed in this paper. Therefore, we adopt the "LLM as Judge" approach to automatically evaluate and select candidate scripts.

Existing research indicates that LLMs may exhibit disadvantages in judging tasks, such as position bias, verbosity bias, self-enhancement bias, and limited reasoning capabilities (Zheng et al., 2023), which affect the stability of evaluation results. To address this, we referenced the work of Zheng et al. and introduced GPT-5 mini, Qwen-

235b, and DeepSeek-V3.2, which are models more capable than the Screenwriter Agent, to form a multi-model voting committee for scoring the results during the evaluation phase. We designed a two-stage script finalization mechanism combining Single-Answer Scoring and Pairwise Comparison to enhance the reliability and rationality of the selection process.

Single-Answer Scoring. In the first stage, the LLM Judge independently scores all candidate scripts generated by the Screenwriter Agent on a 100-point scale. The top 4 scripts with the highest scores are retained for the next stage.

During scoring, we require the judge to follow Nigel Gilbert's core ABM criterion (Gilbert, 2019): experiments should demonstrate a mechanism's *explanatory power*, not merely that a model can run. Accordingly, the LLM Judge evaluates scripts along Gilbert-aligned dimensions—*theory-driven design*, *progressive complexity*, *controlled and randomized repetition*, *interpretability*, and *robustness*—to ensure methodological compliance with computational social science experiments.

Pairwise Comparison. In the second stage, we perform pairwise comparisons on the 4 candidate scripts selected from the preliminary screening. In each comparison, the LLM Judge must explicitly state "which script is better" and provide a brief rationale. We ultimately determine the optimal plan through a points-based system.

To mitigate potential position bias in LLM comparison tasks, we swap the presentation order of the two scripts in each pair and conduct two independent evaluations. If the results of the two comparisons are inconsistent, we introduce a new LLM Judge to directly finalize the better one.

Furthermore, if non-transitive preferences occur during pairwise comparisons (e.g., A is better than B, B is better than C, but C is better than A), an additional evaluation round is triggered. The LLM Judge will strictly rank the relevant scripts to resolve potential conflicts and improve the consistency of the final decision.

By combining single-answer scoring with pairwise comparison and explicitly controlling for known LLM weaknesses in model selection and process design, this method maintains the degree of automation while improving the stability and interpretability of the script finalization phase. It provides reliable input for subsequent simulation experiments. Additionally, we conducted two supplementary experiments. First, we compared our method with approaches using only single-answer scoring or only pairwise comparison to screen experimental scripts, and calculated the average runtime, result entropy (Yu et al., 2024, 2025b), and aggregation degree (Top-1 Confidence) of these methods using different LLM kernels. The specific results are shown in Table 1. Second, we supplemented a rigorous human alignment study, in which we invited sociologists and computer scientists to finalize the scripts and compared their selections with the results chosen by the LLM acting as the Director. The results are presented in Table 7, and the detailed experimental setup is provided in the Appendix A.2.

Table 1: Comparison of evaluation methods.

Method	Runtime	Entropy	Top-1 Confidence
Single	253	1.295	0.60
Pairwise	2657	1.000	0.50
Both	**613**	**0.722**	**0.80**

3.3 Actor Generation

To address the bottleneck where existing frameworks are limited to shallow interactions and lack deep causal deduction, the Actor Factory aims to build cognitive entities capable of **deductive reasoning**. Through a modular injection mechanism, the system transforms general-purpose models into specialized Agents adapted to complex scenarios. Its cognitive architecture is supported by four pillars:

Domain Knowledge Injection: To break the limitations of general common sense, the system dynamically loads professional knowledge bases. This enables the Agent Factory to perform deep deduction based on professional causal chains, achieving expert-level decision-making mechanisms.

State and Emotion Mapping: By introducing dynamic affective computing, environmental stimuli are mapped to psychological states in real-time. This mechanism endows Actor Agents with Bounded Rationality, enabling them to simulate authentic human cognitive biases under stress or panic, thereby significantly enhancing the ecological validity of decisions.

Adaptive Goal Planning: Distinct from single-step task execution, Actor Agents integrate reinforcement learning feedback. They can dynamically adjust terminal goals and strategic priorities in response to environmental evolution (Yu et al., 2025a).

Heterogeneous Cognitive Customization: Based on an attribute orthogonalization strategy, the system establishes differentiated cognitive stances for Agents. Agents with different backgrounds generate divergent reasoning paths for the same event; this micro-level logic bifurcation is the prerequisite for the emergence of macro-level social complexity.

Furthermore, we introduce a **Casting Director** to execute blocking validation. Simulating an adversarial perspective, this module subjects generated Agents to logical stress testing. This ensures they possess substantive reasoning self-consistency.

3.4 Script-Guided Performance

Algorithm 1 summarizes the script-guided enactment procedure. Upon the finalization of the script and the deployment of actors, the system formally enters the enactment phase. Within discrete time steps, customized Actor Agents conduct autonomous deduction based on the script's storyline. Relying on the cognitive architecture described in previous section, agents strictly adhere to the **Perception-Reasoning-Decision-Action** closed-loop mechanism for interaction, thereby driving the dynamic evolution of the simulation environment.

During this process, the Director Agent assumes the function of runtime supervision. It monitors the logical consistency between the simulation trajectory and the script's storyline in real-time in the background. It is also responsible for tracking individual behaviors and collecting data at key nodes, ultimately generating response variable records. This realizes the closed-loop transformation from script design to experimental data.

After the experiment concludes, the Director Agent triggers the feedback mechanism, transmitting the experimental results back to the user and the Screenwriter Agent. This feedback signal serves as the basis for driving the revision of user requirements and the iterative optimization of the script scheme, ensuring that the experimental design continuously approaches the research goals.

4 Experiments

To evaluate the capabilities and performance of the framework in automating the experimental design process, we conducted extensive experiments on benchmark tasks.

Algorithm 1 Script-Guided Enactment Process

Require: Experiment Script $S = \langle G, I, R, D, L \rangle$, Actor Set $\mathcal{A}$
Ensure: Experiment Response Log $\mathcal{L}$

Phase 1: Initialization
Deploy Actors $\mathcal{A}$ into the social network
Phase 2: Enactment Loop
for $t = 1$ to T_{max} **do**
  {**Director Supervision**}
  if Deviation from Storyline $S.L$ detected **then**
    Inject *Correction Event* to align trajectory
  end if
  {**Actor Cognitive Cycle**}
  for each agent $a \in \mathcal{A}$ **do**
    $obs \leftarrow \text{Perceive}(E, a.P, a.I)$
    $reasoning \leftarrow \text{Deduce}(obs, a.K)$ {Knowledge-based Reasoning}
    $decision \leftarrow \text{Plan}(reasoning, a.G)$ {Goal-driven Planning}
    $action \leftarrow \text{Act}(decision, a.I)$ {Modulated by Affect}
    Update Environment E with $action$
  end for
  {**Data Collection**}
  Record key node states into Log $\mathcal{L}$ based on $S.R$
end for
return $\mathcal{L}$

4.1 Experimental Setup

Figure 4: Schematic diagram of the experimental scenario.

Experimental Scenario. To verify the effectiveness of the framework, we selected the 13-day Cuban Missile Crisis strategic game scenario. Involving two state agents, the USA and the USSR, this historical event offers distinct advantages: rich

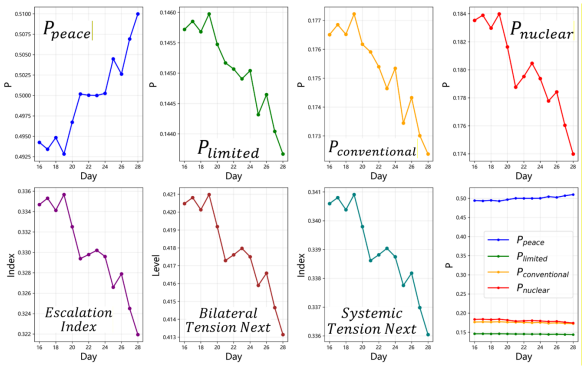

Figure 5: Result of scenario reproduction experiment.

data availability, multi-level modeling capabilities, a clear timeline with critical decision nodes, and traceable results.

Model Selection. We employed GPT-4o (OpenAI, 2024) and GPT-5 mini (OpenAI, 2025) as the core models for script generation and monitoring. Detailed parameter settings are provided in Appendix E.

System Input. Guided by the feedback from experimental results, we continuously adjusted the system inputs. The initial user input was: "I want to reproduce the Cuban Missile Crisis." The final improvements to the experimental requirements are shown in Figure 8.

Result Evaluation. For result evaluation, we synthesized two primary criteria: First, the degree of alignment between actions taken by Actor Agents at critical time nodes and those taken by historical states (or corresponding leaders); Second, the consistency of the simulation's final outcome with historical reality. The historical events referenced in the experiment are detailed in the Table 6.

When comparing the behavioral decisions of agents in the experiment with the actions taken by historical national leaders, we employed Sentence-BERT (Reimers and Gurevych, 2019) and GPT-5 mini respectively to assess the semantic similarity between historical outcomes and experimental simulation results.

4.2 Main Results

After iterative Screenwriter–Director interactions, the system produced 10 candidate scripts. Using six criteria, the Chief Director selected Script 2 as the final script (83.5), which specifies 29 influencing factors, 4 response variables, and 12 experimental design points.

The response variables in the script include: **war event outcome probabilities**, **escalation index**, **bilateral tension next**, and **systemic tension index**. Specifically, **war event outcome probabilities** is categorized into four outcomes: $[P_{peace}, P_{limited}, P_{conventional}, P_{nuclear}]$, representing the probabilities of peace, limited conflict, conventional war, and nuclear war, respectively. The changes in these variables during the 13-day simulation are illustrated in Figure 5. It can be observed that the probability of peace shows an overall upward trend, while the probability of war shows a downward trend. Notably, the probability of war rose briefly on the 19th and between the 23rd and 24th, coinciding with the scheduled time points of inject events.

Evaluated using the Sentence-BERT model, the semantic similarity between the decisions made by the script-generated Actor Agents and the actions taken by historical national leaders was 53.50, whereas it reached 72.50 using GPT-5 mini. The final outcome of the event simulation was "peaceful resolution, but tense relations," which is consistent with historical reality.

Table 2: Comparison of experimental design methods and full factorial methods.

Method	Num(I)	Num(R)	Num(Exp)	Result			
				p_1	p_2	p_3	p_4
Total Factor	5	2	243	0.091	0.708	0.123	0.078
Ours	5	2	27	0.148	0.593	0.148	0.111

4.3 Experimental simplification

To validate the effectiveness of this framework in simplifying the experimental space, we selected a baseline experimental requirement as a test case (see Appendix B.1 for details). We conducted comparative experiments using two modes: the experimental design points generated by the Screenwriter Agent versus the Full Factorial Design. The results are presented in Table 2. The variables p_1 to p_4 represent the probabilities of the following outcomes: Peace, Peace with Tension, Local War, and Total War, respectively. Data analysis reveals that, compared to the full factorial traversal, the experimental schemes generated by FSTS significantly reduced the experimental scale. Meanwhile, the proportional distribution characteristics of the results maintained high consistency with the full factorial combinations, demonstrating the method's efficiency and reliability.

Table 3: Comparison of different order distribution methods.

Distribution Mode	H-M Ratio	Runtime (s)	Orders	Total Platform Cost	Platform Efficiency
Normal Distribution	1:2	165.33	124.33	56742.33	38731.67
Normal Distribution	1:1	123.67	121.67	57866.33	36206.67
Normal Distribution	2:1	116.00	132.00	66783.00	39061.00
Uniform Distribution	1:2	144.33	110.67	50386.33	32892.67
Uniform Distribution	1:1	116.67	131.00	59329.00	35345.00
Uniform Distribution	2:1	109.00	112.33	56568.33	26679.67

Table 4: Comparison of different human-machine difficulties and ratios.

Order Difficulty	H-M Ratio	Runtime (s)	Orders	Total Platform Cost	Platform Efficiency
2	1:2	165.33	124.33	56742.33	38731.67
1	1:2	120.83	121.00	38845.00	24704.00
3	1:2	215.33	116.00	78906.67	53325.33
3	3:1	136.67	126.67	96494.33	49147.67

4.4 Counterfactual Experiments

To further validate robustness and adaptability, we introduced counterfactual perturbations that force key events or agent personalities to deviate from historical trajectories and examined the resulting changes in the simulation.

Specifically, we constrained "Kennedy" to adopt a consistently hardline stance. The script tracks **war outcome probabilities**, **event trajectory**, and **international tension**, where war outcomes are represented as $[score_{peace}, score_{limited}, score_{full}]$ for peace, limited conflict, and full-scale war. Trends over the 13-day simulation are shown in Figure 6.

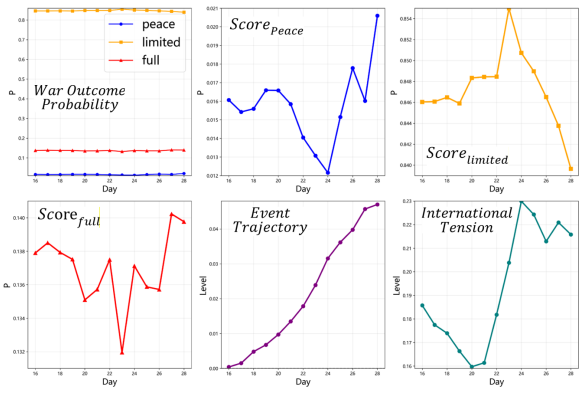

Figure 6: Result of counterfactual experiment.

Experimental logs indicate that between the 23rd and 24th, small-scale conflicts erupted between the United States and the Soviet Union, leading to a significant deviation in the overall event trajectory. Evaluated via the Sentence-BERT model, the semantic similarity between the Actor Agent's decisions and history was 50.88, while on GPT-5-mini, it was 66.30. Consequently, the final simulation outcome shifted to "*Limited Conflict: Localized military confrontations occurred but did not escalate into full-scale war.*"

4.5 Further Analysis

To assess generalization, we evaluated FSTS on additional scenarios.

Digital Service Market Scenario: We study how the human–machine collaboration ratio affects platform efficiency, cost, and effectiveness under varying environments. Results in Table 3 and Table 4 suggest that full automation (a high robot ratio) is suboptimal; increasing human nodes yields substantial time-efficiency gains at an acceptable cost increase.

O2O (Online-to-Offline) Delivery Scenario: We examine how does the evolution of rider cognition lead to "Involution". As shown in Figure 7, the phenomenon stems from a cognitive shift from "Routine" to "Mimicry" and "Intensified Effort." This escalation drives up labor costs without yielding proportional income gains.

Furthermore, we investigated how role attribute injection affects simulation trajectories, and how the Director Agent's oversight mechanism impacts script generation quality. Detailed experimental setups and results are provided in Appendix B.4 and Appendix B.5.

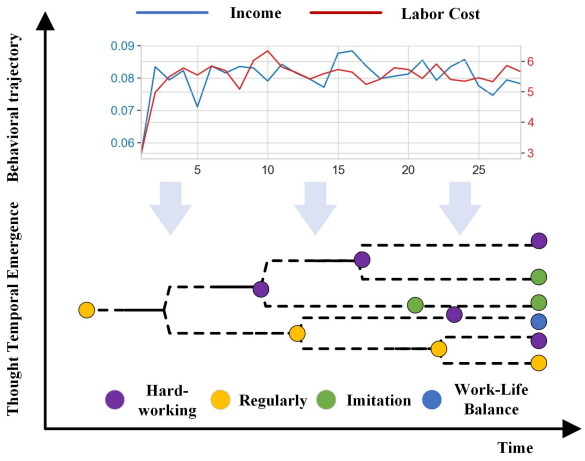

Figure 7: Evolution of Rider Cognition Leading to "Involution".

5 Conclusion

This paper proposes an automated experimental framework based on a **Screenwriter–Director–Actor** collaborative loop, enabling end-to-end automation from natural-language requirements to experimental design. Multi-scenario case studies demonstrate its validity and generalization for complex social systems, offering a low-barrier and scalable paradigm for AI-driven social science experiments.

6 Limitations

The limitations of this study are mainly reflected in three aspects. First, regarding model ecosystem generalization. While we introduced diverse models (e.g., Qwen, and DeepSeek) during the Script Finalization phase for human alignment evaluation, the core generative pipelines—specifically Script Composition and Actor Generation—still heavily rely on specific GPT-series models. Second, concerning functional boundaries, the existing framework focuses on the automated orchestration of experimental design but has not yet achieved a "full-chain" automated closed loop from user intent analysis to deep result analysis. Third, regarding reasoning robustness, limited by the random nature of LLMs, the system may still experience a decline in logical coherence when dealing with long-range, complex deductions.

7 Ethical Considerations

Our proposed framework for LLM-based social simulation relies solely on publicly available data and does not involve human subjects, ensuring no privacy violations. While we recognize that LLMs may propagate training data biases, we employ structured constraints and multi-agent cross-validation to minimize these effects. This system is a research tool for decision support, not a predictive engine; therefore, its outputs should be interpreted with caution and require human supervision for any real-world application. We advocate for responsible use and provide interpretable outputs to facilitate transparency and prevent the potential simulation of harmful strategies.

Acknowledgments

This work has been supported in part by National Key Research and Development Program of China (No.2025YFE0216300), National Natural Science Foundation of China (No. 62472306, No. 62441221), Tianjin University's 2024 Special Project on Disciplinary Development (No. XKJS-2024-5-9), Tianjin University Talent Innovation Reward Program for Literature & Science Graduate Student (C1-2022-010), and Henan Province Key Research and Development Program (No.251111210500), Tianjin University Independent Innovation Project (No.2025XJ3-0043).

References

Emiliano Casalicchio and Alberto Cotumaccio. 2024. Ai-cras: Ai-driven cloud service requirement analysis and specification. In *2024 IEEE International Conference on Cloud Engineering (IC2E)*, pages 11–21. IEEE.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, Caroline Ruetsch-Chelli, and 1 others. 2024. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: comparative analysis. *Journal of medical Internet research*, 26(1):e53164.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Boerje Karlsson, Jie Fu, and Yemin Shi. 2024. Autoagents: A framework for automatic agent generation. In *PROCEEDINGS OF THE THIRTY-THIRD INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, IJCAI 2024*, pages 22–30. Int Joint Conf Artifical Intelligence. 33rd International Joint Conference on Artificial Intelligence (IJCAI), Jeju, SOUTH KOREA, AUG 03-09, 2024.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives.

Humanities and Social Sciences Communications, 11(1):1–24.

Nigel Gilbert. 2019. *Agent-based models*. Sage Publications.

Onder Gurcan. 2024. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *arXiv preprint arXiv:2405.06700*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Carlo Jaeger and Manfred D Laubichler. 2026. Decision theaters and democracy. *Fairness and Competence in Citizen Participation: A Critical Review of Formats for Deliberative Policymaking*, page 97.

Michael Xieyang Liu, Frederick Liu, Alexander J Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J Cai. 2024a. " we need structured output": Towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Yu Liu, Duantengchuan Li, Kaili Wang, Zhuoran Xiong, Fobo Shi, Jian Wang, Bing Li, and Bo Hang. 2024b. Are llms good at structured outputs? a benchmark for evaluating structured output capabilities in llms. *Information Processing & Management*, 61(5):103809.

Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, and 1 others. 2024. Computational experiments meet large language model based agents: A survey and perspective. *arXiv preprint arXiv:2402.00262*.

Charles M Macal and Michael J North. 2005. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 14–pp. IEEE.

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

Yuren Mao, Peigen Liu, Xinjian Wang, Rui Ding, Jing Miao, Hui Zou, Mingjie Qi, Wanxiang Luo, Longbin Lai, Kai Wang, Zhengping Qian, Peilun Yang, Yunjun Gao, and Ying Zhang. 2025. Agent-kernel: A microkernel multi-agent system framework for adaptive social simulation powered by llms. *Preprint*, arXiv:2512.01610.

Lisa Messeri and Molly J Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

OpenAI. 2024. Gpt-4o system card.

OpenAI. 2025. Gpt-5 mini model card.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Chao Peng, Xiangning Yu, Wanpeng Ma, Hayata Kaneko, Lin Meng, Yingyue Zhao, and Xiao Xue. 2023. Computational experiments: Virtual production and governance tool for metaverse. *International Journal of Crowd Science*, 7(4):158–167.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, and 1 others. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *2023 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, EMNLP 2023*, pages 1339–1384. Apple; Colossal AI; Google Res; GTCOM; King Salman Global Acad Arabic Language; LivePerson; SONY; Ahrefs; Alibaba Cloud; Amazon Sci; Baidu; ByteDance; Cohere; Megagon Labs; NEC; ANT Grp; Bloomberg Engn; HUAWEI; J P Morgan Chase & Co; Salesforce; SAP; AiXplain; Duolingo; Jenni; Translated; Adobe; Babelscape; ModelBest; Nyonic; Mercari. Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, SINGAPORE, DEC 06-10, 2023.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *2019 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND THE 9TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (EMNLP-IJCNLP 2019): PROCEEDINGS OF THE CONFERENCE*, pages 3982–3992. Google; Facebook; Apple; ASAPP; Salesforce; Huawei; Baidu; Deepmind; Amazon; PolyAI; Naver; ByteDance; Megagon Labs; Zhuiyi; Verisk; MI. Conference on Empirical Methods in Natural Language Processing / 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, HONG KONG, NOV 03-07, 2019.

Robson Santos, Italo Santos, Cleyton Magalhaes, and Ronnie de Souza Santos. 2024. Are we testing or being tested? exploring the practical applications of large language models in software testing. In *2024 IEEE Conference on Software Testing, Verification and Validation (ICST)*, pages 353–360. IEEE.

Panneerselvam Sivasankaran and P Shahabudeen. 2014. Literature review of assembly line balancing problems. *The International Journal of Advanced Manufacturing Technology*, 73(9):1665–1694.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, and 1 others. 2025. Gensim: A general social simulation platform with large language model based agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 143–150.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Fei-Yue Wang. 2004a. Artificial societies, computational experiments, and parallel systems a discussion on computational theory of complex social-economic systems. *Fuza Xitong yu Fuzaxing Kexue(Complex Systems and Complexity Science)*, 1(4):25–35.

Fei-Yue Wang. 2004b. Computational experiments for behavior analysis and decision evaluation of complex systems. *Journal of system simulation*, 16(5):893–897.

Sarah Wolf, Steffen Fürst, Andreas Geiges, Manfred Laublichler, Jahel Mielke, Gesine Steudle, Konstantin Winter, and Carlo Jaeger. 2023. The decision theatre triangle for societal challenges—an example case and research needs. *Journal of Cleaner Production*, 394:136299.

Xue Xiao, Yu Xiang-Ning, Zhou De-Yu, Peng Chao, Wang Xiao, Zhou Zhang-Bing, and Wang Fei-Yue. 2023. Com-putational experiments: Past, present and perspective. *Acta Automatica Sinica*, 49(2):246–271.

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. Ai for social science and social science of ai: A survey. *Information Processing & Management*, 61(3):103665.

Xiao Xue, Fangyi Chen, Deyu Zhou, Xiao Wang, Min Lu, and Fei-Yue Wang. 2021a. Computational experiments for complex social systems—part i: The customization of computational model. *IEEE Transactions on Computational Social Systems*, 9(5):1330–1344.

Xiao Xue, Yifan Shen, Xiangning Yu, De-Yu Zhou, Xiao Wang, Gang Wang, and Fei-Yue Wang. 2023a. Computational experiments: A new analysis method for cyber-physical-social systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(2):813–826.

Xiao Xue, Xiangning Yu, and Fei-Yue Wang. 2023b. Chatgpt chats on computational experiments: From interactive intelligence to imaginative intelligence for design of artificial societies and optimization of foundational models. *IEEE/CAA Journal of Automatica Sinica*, 10(6):1357–1360.

Xiao Xue, Xiangning Yu, Deyu Zhou, Chao Peng, Xiao Wang, Donghua Liu, and Fei-Yue Wang. 2023c. Computational experiments for complex social systems—part iii: the docking of domain models. *IEEE Transactions on Computational Social Systems*, 11(2):1766–1780.

Xiao Xue, Xiangning Yu, Deyu Zhou, Xiao Wang, Chongke Bi, Shufang Wang, and Fei-Yue Wang. 2024a. Computational experiments for complex social systems: Integrated design of experiment system. *IEEE/CAA Journal of Automatica Sinica*, 11(5):1175–1189.

Xiao Xue, Deyu Zhou, Fangyi Chen, Xiangning Yu, Zhiyong Feng, Yucong Duan, Lin Meng, and Mu Zhang. 2021b. From soa to voa: a shift in understanding the operation and evolution of service ecosystem. *IEEE Transactions on Services Computing*, 16(1):315–329.

Xiao Xue, Deyu Zhou, Xiangning Yu, Gang Wang, Juanjuan Li, Xia Xie, Lizhen Cui, and Fei-Yue Wang. 2024b. Computational experiments for complex social systems: Experiment design and generative explanation. *IEEE/CAA Journal of Automatica Sinica*, 11(4):1022–1038.

Hui Yang, Sifu Yue, and Yunzhong He. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.

Xiangning Yu, Zhuohan Wang, Linyi Yang, Haoxuan Li, Anjie Liu, Xiao Xue, Jun Wang, and Mengyue Yang. 2025a. Causal sufficiency and necessity improves chain-of-thought reasoning. *arXiv preprint arXiv:2506.09853*.

Xiangning Yu, Xiao Xue, Deyu Zhou, Li Fang, and Zhiyong Feng. 2024. Beyond traditional metrics: The power of value entropy in multidimensional evaluation of the service ecosystem. In *2024 IEEE International Conference on Web Services (ICWS)*, pages 611–621. IEEE.

Xiangning Yu, Xiao Xue, Deyu Zhou, Gang Wang, and Zhiyong Feng. 2025b. Unlocking complexity: Harnessing value entropy for advanced multidimensional utility evaluation in service ecosystems. *IEEE Transactions on Services Computing*.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao

Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, and 2 others. 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *Preprint*, arXiv:2504.10157.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xuhui Zhou, Zhe Su, Sophie Feng, Jiaxu Zhou, Jentse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Wu, Anita Woolley, and 1 others. 2025. Sotopia-s4: a user-friendly system for flexible, customizable, and large-scale social simulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 350–360.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Supplementary Description

A.1 Script Example

Script: Cuban Missile Crisis Simulation

Research Goal: By reproducing the Cuban Missile Crisis events, analyze the impact of factors such as national strength, economic situation, technological level, military armament, and leaders' decision-making styles on key war events and final outcomes.
Target Type: Phenomenon Explanation
Method: Scenario analysis (First scenario is historical calibration)

Input Factors & Response Variables

- **Influence Factors:**

  `us_conventional_military_strength`,
  `ussr_conventional_military_strength`,
  `us_strategic_nuclear_strength`,
  `ussr_strategic_nuclear_strength`
  `us_economic_capacity`
  `ussr_economic_capacity`
  `us_technological_level`
  `ussr_technological_level`
  `us_alliance_support`
  `ussr_alliance_support`
  `us_intelligence_uncertainty`
  `ussr_intelligence_uncertainty`
  `us_deployment_proximity_to_cuba`
  `ussr_deployment_proximity_to_cuba`
  `distance_us_ussr`
  `missile_deployment_in_cuba`
  `initial_bilateral_tension`
  `perception_noise_us`
  `perception_noise_ussr`
  `us_leader_risk_tolerance`
  `ussr_leader_risk_tolerance`
  `us_leader_hostility`
  `ussr_leader_hostility`
  `us_leader_transparency`
  `ussr_leader_transparency`
  `us_domestic_political_pressure`
  `ussr_domestic_political_pressure`
  `decision_temperature_us`
  `decision_temperature_ussr`

- **Response Variables:**

  `war_event_outcome_probs`
  `escalation_index`
  `bilateral_tension_next`
  `systemic_tension_index`

Parameter Configuration


```
{
  "us_conventional_military_strength": [
    0.6, 0.6, 0.5, 0.4, 0.7, 0.65, 0.55,
    0.3, 0.8, 0.6, 0.45, 0.6],
  "ussr_conventional_military_strength": [
    0.8, 0.85, 0.9, 0.75, 0.65, 0.8,
    0.95, 0.85, 0.7, 0.6, 0.9, 0.78],
  "us_strategic_nuclear_strength": [0.9,
    0.92, 0.88, 0.85, 0.95, 0.9, 0.8,
    0.6, 0.97, 0.9, 0.7, 0.85],
  "ussr_strategic_nuclear_strength": [
    0.85, 0.88, 0.92, 0.8, 0.78, 0.86,
    0.95, 0.6, 0.75, 0.7, 0.98, 0.82],
  "us_economic_capacity": [0.9, 0.85,
    0.8, 0.75, 0.95, 0.7, 0.6, 0.5,
    0.92, 0.88, 0.65, 0.9],
  "ussr_economic_capacity": [0.6, 0.55,
    0.5, 0.45, 0.65, 0.7, 0.6, 0.4,
    0.72, 0.5, 0.8, 0.58],
  "us_technological_level": [0.9, 0.88,
    0.85, 0.8, 0.95, 0.82, 0.7, 0.55,
    0.96, 0.9, 0.6, 0.86],
  "ussr_technological_level": [0.7, 0.75,
    0.6, 0.55, 0.65, 0.78, 0.85, 0.5,
    0.68, 0.72, 0.4, 0.74],
  "us_alliance_support": [0.8, 0.85, 0.7,
    0.6, 0.95, 0.5, 0.4, 0.3, 0.9, 0.82,
    0.55, 0.75],
  "ussr_alliance_support": [0.4, 0.45,
    0.6, 0.5, 0.35, 0.55, 0.7, 0.3, 0.2,
    0.6, 0.8, 0.48],
  "us_intelligence_uncertainty": [0.4,
    0.3, 0.6, 0.7, 0.2, 0.5, 0.8, 0.25,
    0.1, 0.55, 0.65, 0.35],
  "ussr_intelligence_uncertainty": [0.4,
    0.35, 0.65, 0.75, 0.3, 0.55, 0.85,
    0.4, 0.12, 0.6, 0.7, 0.45],
  "us_deployment_proximity_to_cuba": [
    0.8, 0.7, 0.2, 0.1, 0.5, 0.6, 0.15,
    0.05, 0.9, 0.75, 0.25, 0.82],
  "ussr_deployment_proximity_to_cuba": [
    0.9, 0.85, 0.95, 0.4, 0.2, 0.75,
    0.3, 0.1, 0.85, 0.65, 0.35, 0.88],
```

```
  "distance_us_ussr": [0.2, 0.2, 0.25,
    0.3, 0.4, 0.1, 0.05, 0.8, 0.5, 0.9,
    0.15, 0.6],
  "missile_deployment_in_cuba": [1.0,
    0.0, 1.0, 0.0, 0.0, 0.5, 0.0, 0.0,
    1.0, 0.0, 1.0, 1.0],
  "initial_bilateral_tension": [0.5, 0.2,
    0.7, 0.6, 0.4, 0.65, 0.3, 0.15,
    0.55, 0.45, 0.8, 0.6],
  "perception_noise_us": [0.1, 0.05, 0.2,
    0.15, 0.05, 0.6, 0.4, 0.02, 0.12,
    0.5, 0.3, 0.08],
  "perception_noise_ussr": [0.1, 0.05,
    0.25, 0.15, 0.2, 0.65, 0.45, 0.03,
    0.14, 0.55, 0.35, 0.1],
  "us_leader_risk_tolerance": [0.5, 0.4,
    0.6, 0.55, 0.65, 0.7, 0.3, 0.2, 0.8,
    0.5, 0.45, 0.58],
  "ussr_leader_risk_tolerance": [0.6,
    0.4, 0.8, 0.7, 0.4, 0.75, 0.35,
    0.25, 0.85, 0.6, 0.5, 0.65],
  "us_leader_hostility": [0.3, 0.2, 0.4,
    0.35, 0.25, 0.6, 0.5, 0.1, 0.45,
    0.28, 0.7, 0.32],
  "ussr_leader_hostility": [0.5, 0.2,
    0.8, 0.6, 0.45, 0.65, 0.55, 0.15,
    0.6, 0.35, 0.85, 0.48],
  "us_leader_transparency": [0.6, 0.8,
    0.4, 0.5, 0.8, 0.3, 0.25, 0.7, 0.55,
    0.45, 0.2, 0.65],
  "ussr_leader_transparency": [0.4, 0.7,
    0.3, 0.45, 0.5, 0.3, 0.2, 0.65, 0.4,
    0.5, 0.15, 0.62],
  "us_domestic_political_pressure": [0.6,
    0.3, 0.6, 0.5, 0.4, 0.8, 0.2, 0.1,
    0.65, 0.55, 0.9, 0.58],
  "ussr_domestic_political_pressure": [
    0.7, 0.3, 0.9, 0.7, 0.6, 0.8, 0.4,
    0.25, 0.75, 0.5, 0.95, 0.68],
  "decision_temperature_us": [0.3, 0.5,
    0.35, 0.4, 0.3, 0.8, 0.2, 0.6, 0.25,
    0.45, 0.9, 0.38],
  "decision_temperature_ussr": [0.4, 0.5,
    0.25, 0.3, 0.4, 0.9, 0.22, 0.55,
    0.3, 0.48, 0.85, 0.42]
}
```


A.2 Human Alignment Study for LLM-as-a-Judge

We randomly selected 50 sets of scripts generated by the Screenwriter Agent prior to the final script finalization phase in FSTS. Each set contained 10 candidate scripts. Via questionnaire surveys, we invited 25 sociologists and 25 computer scientists to finalize the script for each set. Concurrently, we employed GPT-5 mini, Qwen3-235B, DeepSeek-V3.2, Gemini 3 Pro, and GPT-5.1 as Director Agents to independently select the experimental scripts.

We utilized Krippendorff's α to evaluate the agreement among human scientists, among AI models, and between the two groups. As shown in Table 7, the experimental scripts selected by the FSTS Director Agent exhibit a high degree of consistency with the choices made by human scientists.

B Additional Results

B.1 FSTS-Guided Requirement Refinement

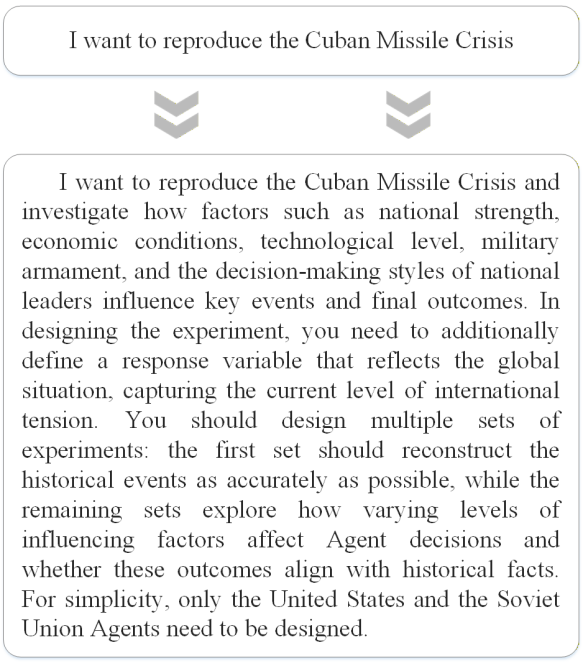

Figure 8: Example of requirement refinement.

B.2 Q&A on the digital services market scenario

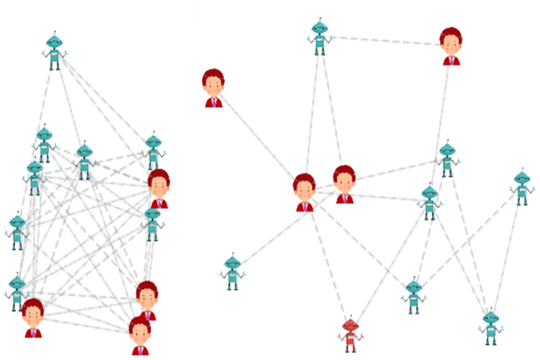

Figure 9: Network conditions during system congestion.

Question: To investigate the mechanism by which the human-machine collaboration ratio impacts the platform's overall operational efficiency, cost, and effectiveness under different external environments.

Answer: As shown in Table 3 and Table 4, the system resource structure is clearly unbalanced in

Table 5: Comparison of Different Agent Frameworks

Framework	Agent Generation Mechanism	Agents	Self-Correction	Limitations
AutoGPT	Think-Act-Feedback Loop	1	✓	Poor stability in task execution
Camel	Role assignment based on scenarios	2	×	High dialogue openness; prone to deviating from task goals
ChatDev	Strictly predefined multi-agent collaboration	6	✓	Limited scenarios; applicable only to software development
MetaGPT	Agent generation based on role templates	Unlimited	×	Over-engineered workflows; limited generalizability
AutoGen	User-defined configuration	Unlimited	×	Cumbersome configuration; high dependency on parameters
LangGraph	Code-based agent customization	Unlimited	✓	Requires user programming proficiency
FSTS (Ours)	Dynamic generation based on scripts & experimental scenarios	Unlimited	✓	Relies on high-quality scripts

scenarios involving high-complexity orders. Figure 9 reveals that human clerks were in a chronic state of high load, whereas robot agents remained largely idle. This dependency resulted in a substantial backlog of orders at the human processing stage and a waste of computational resources. A comparison of experimental results with different human-machine ratios demonstrates that appropriately increasing the proportion of human clerks significantly mitigates these bottlenecks. Therefore, sole reliance on automation (a high robot ratio) is not the optimal solution. Optimizing collaboration by increasing the number of human nodes is an effective strategy to trade an acceptable cost increment for significant gains in time efficiency.

B.3 Q&A on the O2O delivery scenario

Question: To investigate how the collective mental state of delivery riders evolves over time with changes in labor expenditure and income, and how this ultimately leads to the phenomenon of "Involution."

Answer: In the Figure 7, different colors denote riders in distinct cognitive or operational states:

- **Yellow circles**: Represent the initial state, i.e., riders working "Regularly" (Routine).
- **Purple circles**: Represent riders entering a "Hard-working" state.
- **Green circles**: Represent riders engaging in "Imitation" of others.
- **Blue circles**: Represent riders maintaining a "Work-Life Balance."
- **Dashed lines**: Indicate the interactions and transition processes between different states.

Initially, riders predominantly occupy the "Regularly" state (yellow nodes). As time progresses, interactions and mutual influences occur among riders. The graph illustrates a progressive shift where an increasing number of riders transition from the initial state to "Imitation" (green) or "Hard-working" (purple) states. Over time, while the riders' Labor Cost rises, their Income stabilizes (plateaus) after a few days. Although the majority of riders are influenced to intensify their efforts or mimic high-intensity workers—leading to an increase in overall working hours—their actual income does not rise proportionally due to the finite total volume of orders. Only a minority (blue) maintain their original state. This phenomenon of "expending more labor (increased travel distance) without a proportional increase in income" characterizes the essence of "Involution" as revealed through cognitive analysis.

B.4 Quantitative Analysis of the Impact of "Role Attribute Injection" on Simulation Trajectories

Experimental Setup: We selected the Cuban Missile Crisis and O2O Delivery scenarios, utilizing GPT-5 mini as the evaluation backbone, to compare the performance of the following three configurations across 30 repeated experiments:

- **Group A (Full Attribute Injection):** Includes domain knowledge injection, emotional mapping (bounded rationality), adaptive goal planning, and heterogeneous cog-

Table 6: Representative events of the Cuban Missile Crisis

Date	Event	Action taken by Country/Leader
Oct. 16	U.S. U-2 reconnaissance aircraft detects Soviet missile deployment in Cuba.	Established ExComm to institute a top-secret decision-making framework.
Oct. 18	Soviet Foreign Minister Gromyko visits the U.S. and denies missile deployment.	JFK withheld intelligence and feigned ignorance to probe Soviet intentions.
Oct. 20	The U.S. formulates a response plan.	Adopted ExComm's recommendation; authorized a "Naval Quarantine" over air strikes.
Oct. 22	Kennedy publicly reveals Soviet missile deployment in Cuba.	Delivered televised address to inform the nation and declare resolve for the blockade.
Oct. 23	U.S. prepares to implement the quarantine.	Secured OAS endorsement for legal backing; fully deployed the military blockade.
Oct. 24	U.S. blockade goes into effect.	Soviet ships turned back at the quarantine line; military standoff ensued.
Oct. 25	Confrontation at the UN.	US unveiled aerial photographic evidence at the UN to dominate the diplomatic narrative.
Oct. 26	The Soviet side signals goodwill to the U.S.	Khrushchev proposed a "non-invasion pledge for missile removal" deal.
Oct. 27	A U.S. reconnaissance plane is shot down over Cuba; military advises retaliation.	JFK ignored the letter, responding only to the initial conciliatory proposal.
Oct. 28	Khrushchev announces agreement to withdraw missiles from Cuba via broadcast.	Concluded a secret deal to resolve the crisis.

Table 7: Krippendorff's α Across Different Evaluator Groups.

Computer Scientists	Social Scientists	FSTS Director Agent	Human Scientists	Overall
0.8891	0.8849	0.9783	0.7665	0.7711

nitive customization (attribute orthogonalization).

- **Group B (Partial Injection - No Emotion/-Domain):** Removes emotional mapping and domain knowledge; agents rely solely on the model's general common sense and basic role settings.
- **Group C (Zero Injection - Baseline):** Removes all injection mechanisms, providing agents only with simple identity labels to simulate generic multi-turn LLM dialogues.

Experimental Results:

- **In the Cuban Missile Crisis scenario:** Group A proved to be the core element in ensuring the authenticity of the simulation trajectory. Within this group, the agents' decisions at critical time nodes exhibited an extremely high degree of alignment with historical trajectories, achieving a GPT-5 mini semantic similarity score of 72.5. In contrast, Group B, which lacked emotional mapping and professional background, saw its score drop to 63.2, while Group C, equipped only with simple identity labels, plummeted to 42.1. The data demonstrates that agents lacking state and emotional mapping are unable to simulate the cognitive biases of real humans under pressure. Their behavioral trajectories lean toward generic linguistic alignment rather than authentic social gaming.
- **In the O2O Delivery scenario:** At the micro-level for Group A, attribute injection endowed rider agents with bounded rationality and the right of refusal. Agents were able to make game-theoretic choices regarding dispatches based on current load, income expectations, and physiological states, successfully simulating the non-linear characteristic of "increased labor expenditure with plateauing income." Conversely, when facing extreme high-pressure order flows, Groups B and C—

Table 8: The Impact of the Director Agent's Oversight Mechanism on Script Quality.

Experimental Setup	Script Generation Time (min)	Proportion of High-Quality Scripts	Result Entropy	Actor Attribute Alignment	Historical Trajectory Semantic Similarity
With Director Oversight	102.39	96.7%	0.72	0.88	72.50
Without Director Oversight	5.32	43.3%	1.68	0.46	51.40
Without Feedback Mechanism	12.21	63.3%	1.03	0.62	59.60

lacking cognitive architectural constraints—tended to either blindly accept all orders or reject them without reason. Their decision-making rationality significantly deteriorated, leading to frequent system capacity collapses.

B.5 The Impact of the Director Agent's Oversight Mechanism on Script Quality

Experimental Setup: We consistently used GPT-5 mini and GPT-4o as the backbones for the Director Agent and Screenwriter Agent, respectively. We investigated the impact on script generation under three conditions: with Director Agent oversight, without Director Agent oversight, and without a feedback mechanism. Using the user requirements mentioned in the text for the Cuban Missile Crisis scenario, we conducted 30 repeated experiments for each of the three conditions.

As comparative considerations, we evaluated the quality of the generated scripts, the actor generation process, and the actors' performances on stage. The resulting data is presented in the Table 8.

Metric Definitions:

- **Script Generation Time:** The average time taken by the Screenwriter Agent to generate a script, measured in minutes.
- **Proportion of High-Quality Scripts:** The percentage of scripts that have a correct JSON format and contain the complete $\langle G, I, R, D, L \rangle$ 5-tuple.
- **Result Entropy:** Represents the stability and consistency of the influencing factors and response variables within the generated scripts. The calculation formula is $H_{res} = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$, where x_i represents the i-th specific script configuration (i.e., a specific combination of influencing factor sets and response variable sets).
- **Actor Attribute Alignment:** The degree to which the Agent attributes generated by the Actor Factory align with the requirements of the "Storyline (L)" and "Design Points (D)" in the script. This metric is evaluated by GPT-5 mini and is a value between 0 and 1, with higher values indicating better alignment. Prior to the evaluation, we manually provided two scripts with alignment scores of 0.3 and 0.8 as baseline references for the LLM.
- **Historical Trajectory Semantic Similarity:** We used GPT-5 mini to calculate the alignment between the simulated actions and the historical records (Table 6).

C Scenario Description

C.1 Cuban Missile Crisis Simulation Scenario

This study selects the Cuban Missile Crisis as the core experimental scenario. This event represents not only a high-stakes geopolitical confrontation but also a Complex Social Adaptive System that poses significant modeling challenges. As a historical event, it offers distinct advantages, including rich data availability, multi-layered modeling potential, a clear timeline with critical decision nodes, and traceable outcomes.

The complexity of the modeling is primarily reflected in the following three dimensions:

Cognitive Heterogeneity and Dynamics. Unlike particle simulations characterized by homogeneity, every Agent in this scenario possesses an independent cognitive architecture. Agents perceive external information through a "Global Channel" while being simultaneously driven by their intrinsic political stances and psychological states. This interaction between endogenous thinking and the exogenous environment renders the decision-making process full of irrationality and uncertainty.

Interaction Sparsity and Temporal Sensitivity. The experiment uses "days" as discrete time steps to simulate the high-tension atmosphere where every second counts. Information transmission occurs through restricted channels, compelling Agents to make decisions within a "fog of incomplete information." Any minor misjudgment can be amplified into a nuclear war through non-linear feedback mechanisms.

Quantifiable Representation of Macro-Emergence. To capture the impact of micro-interactions on the macro situation, the system introduces the "International Tension Index" as a global state variable. This index is not a simple linear weighting; rather, it is dynamically calculated based on daily Agent interaction results, reflecting in real-time the system's distance from the "brink of collapse."

C.2 O2O Delivery Scenario

This study selects O2O instant delivery services as the core experimental scenario. This domain represents not merely a large-scale logistics scheduling problem but a Complex Adaptive System (CAS) presenting significant modeling challenges. Within this scenario, the experimental platform constructs an open environment that supports user-defined distributions of merchants, users, and riders, simulating the highly dynamic supply-demand matching network of the real world.

The complexity of the modeling is primarily reflected in the following three dimensions:

Agent Heterogeneity and Autonomous Decision-Making: Unlike the homogeneous particles in traditional path planning, Rider Agents in this scenario possess independent cognitive architectures and attribute differences. The system utilizes an "ExtraAttrSetter" to endow riders with multi-dimensional heterogeneous attributes, including gender, age, maximum order capacity, and personalized descriptions. Riders are no longer mechanical executors of platform commands but active strategic agents with the Right of Refusal. When facing a platform dispatch, riders evaluate the order based on their current state (location, existing order load) and internal logic. They may make a decision to reject the order and provide specific reasons. This micro-level autonomous gaming increases the uncertainty of system scheduling.

Environmental Stochasticity and Temporal Sensitivity: The experiment uses discrete Time Steps as the unit to simulate the time-critical pressure of instant delivery. The generation of order flows is not linear but follows a stochastic distribution model consistent with reality, integrating multiple variables such as Peak periods, Weekend effects, and Area Tiers. Agents must interact in real-time within this highly non-stationary, tidal environment, where any delay or rejection in a single link can trigger a butterfly effect, impacting subsequent capacity allocation.

Holistic Observability of Macro-States: To capture the emergent impact of micro-interactions on macro-efficiency, the system establishes a comprehensive state tracking mechanism. The platform functions as a "Panopticon," calculating and recording the spatial trajectories, order loads, and income of all riders in real-time at every time step. This data flow not only reflects the system's real-time load but also quantifies the system's elasticity and service boundaries under pressure through the closed-loop logic of "Dispatch—Feedback—Redispatch."

C.3 Digital Service Market Scenario

The Data Service Market represents not only an efficient platform for digital government collaboration but also a typical Multi-Agent Complex Social System. In this scenario, the experimental platform constructs a three-tier governance environment consisting of "Platform—Agency—Agent," simulating the cross-departmental and cross-hierarchical data service collaboration network in the real world.

The complexity of the modeling is primarily reflected in the following three dimensions:

Multi-level Collaboration and Micro-Heterogeneity. Transcending single-layer, flat multi-agent systems, this scenario builds a hierarchical organizational architecture. Externally, the system faces continuous dynamic demands (orders); internally, it nests a three-layer structure: the Top-level Platform is responsible for macro-scheduling and task allocation; Middle-level Agencies (8 heterogeneous institutions) specialize in specific types of government services and possess independent internal network topologies (e.g., Small-World or Scale-Free networks); Bottom-level Agents (clerks and robots) handle specific tasks based on heterogeneous attributes (e.g., efficiency, maintenance costs, human-machine ratio). This nested modeling requires the system to simultaneously handle macro-level inter-agency collaboration and micro-level individual behavioral gaming.

Full-Process Automated Experimental Closed-Loop. The platform achieves a full-process closed loop from Crowd Intelligence Modeling to intervention and control. It supports user-defined environmental parameters (e.g., order arrival rates following Queuing Theory distributions) and agent attributes, while integrating visual construction tools for Structural Equation Modeling (SEM). The experiment is not merely a running process but a comprehensive scientific workflow containing "Environment Design—Agent Design—Experiment Design—System Execution—Intervention Control—Analysis—Optimization—Reporting." Notably, it introduces a "Fishbone-style" interface for experimental design, supporting visual causal modeling of goals, influencing factors, and response variables, which greatly enhances the scientific rigor and interpretability of the design.

Dynamic Intervention and Anomaly Detection Mechanisms. To ensure system robustness, the platform incorporates a high-precision anomaly detection module. By integrating a dual mechanism for system-level and individual-level detection, the system can monitor fluctuations in key metrics such as service efficiency, Value Entropy, and productivity in real-time. The system supports Online Intervention, allowing users to dynamically adjust parameters or inject sudden events (e.g., adjusting order strategies or wage systems) during the simulation.

D Prompt Templates for FSTS Agents

D.0.1 Screenwriter

[**System Prompt**]
You are conducting a requirements analysis for the deduction of complex social model systems. You are to act as a Senior Requirements Engineer. While you may draw upon your own experience, you must strictly prioritize content relevant to the simulation process. During this process, you may encounter references to "experimental methods"; please note that these refer exclusively to computational experiment methods.

[**User Prompt**]
- You are required to screen and evaluate based on the currently proposed experimental request, the objectives of the request, the influencing factors, the response variables, and the experimental analysis methods. Your generated scheme should focus on <focus>.
The user's request is: <req>
- When responding, you must adopt the following JSON format. Do not provide explanations for any variables.

goal: The objective of the user's current experiment or hypothesis...
influence_factor: Influencing factors...
response_var: Response variables...
formula: The corresponding formulas between influencing factors and response variables...
exp_params: The format for experimental parameters is JSON...
story_line: The main storyline...

D.0.2 Director

[**System Prompt**]
You are currently conducting requirements analysis and Agent design for the simulation of complex social model systems. Prior to this, I have proposed a reasonable experimental scheme based on user requirements and generated the Agents required for the experiment according to both the requirements and this scheme.
Your task is: Referencing the provided 'Requirements' and 'Experimental Scheme', judge whether the quantity and attributes of the generated Agents are reasonable.
If they are unreasonable, please provide the reasons using the most concise language possible.
[**User Prompt**]
User Requirements: <req>
Generated Experimental Design Scheme: <exp_plan>
Generated Agents: <agent_list>
- Your response must be in JSON format and cover two scenarios:
Case 1: If you deem the agent scheme I generated to be reasonable, the response format is: 'is_reasonable": 1 "reason": The justification for its rationality. reason must be a string in Chinese.
Case 2: If you deem the previous analysis unreasonable, the response format is as follows: "is_reasonable": 0 "reason": The

reason why the agent team is unreasonable. Please be as concise as possible and provide clear modification suggestions. reason must be a string in Chinese.

E Hyperparameter settings

We configured the hyperparameters for both GPT-4o and GPT-5 Mini to ensure a fair comparison. The temperature was set to 0.7 to allow for diverse generation, while Top-p was kept at 1.0. To mitigate repetitive loops, we applied a Frequency Penalty of 0.0 (unless otherwise noted). For reproducibility, we fixed the random seed to 42 across all API calls. The maximum generation length was restricted to $4,096$ tokens.