Beyond Black-Box Agents: Explainable and Validatable Generative ABMs

Xuening Tang[0009−0006−6585−5976] and Petter Törnberg[0000−0001−8722−8646]

University of Amsterdam, Amsterdam, 1098XH, the Netherlands
xuening.tang@student.uva.nl

Abstract. Generative agent-based models (GABMs) that embed large language models (LLMs) as autonomous agents have attracted growing interest for simulating human behavior and communication. However, because LLMs operate as opaque black boxes, such models are often difficult to validate, interpret, or replicate, which limits their reliability for theory building in the social sciences. This study presents an exploration of an alternative approach in which LLMs serve as external assistants rather than as agents within simulations. We refer to this as XABM (eXplainable Generative ABM). In this framework, LLMs generate explicit behavioral rules, identify relevant decision variables, and translate theoretical model descriptions into executable simulation prototypes, while human researchers retain full control over model structure and validation. By externalizing the decision logic into transparent, inspectable rules, this approach aims to make computational modeling more interpretable and reproducible. The current implementation is evaluated through a series of preliminary tests, including reproducing the canonical Schelling Segregation Model and an exploratory case study on spiral-of-silence dynamics. These initial results suggest that using LLMs as rule generators offers a promising direction for transparent and explainable generative agent-based modeling.

Keywords: Generative Agent-Based Model · Large Language Model · Social Simulation

1 Introduction

Agent-based models (ABMs) have long served as a core methodological approach for studying how individual behavior and interaction give rise to collective social outcomes. By specifying decision rules at the level of agents and allowing these agents to interact within structured environments, ABMs enable researchers to investigate a wide range of social phenomena, including segregation, cooperation, opinion dynamics, diffusion, and collective action [1][2][3]. Their scientific value lies less in prediction than in explanation: ABMs make it possible to trace how macro-level patterns emerge from micro-level assumptions and interaction mechanisms.

Recent advances in large language models (LLMs) have generated renewed enthusiasm for agent-based modeling. Because LLMs can produce context-sensitive

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

language, emulate aspects of human reasoning, and flexibly respond to complex prompts, they have been proposed as a means of dramatically increasing the behavioral richness of simulated agents. This has led to the rapid emergence of generative agent-based models in which LLMs are embedded directly within agents and tasked with generating actions, beliefs, or communications during simulation runtime [4][5]. Such models have been used to explore social interaction, collective sense-making, polarization, and online discourse, and are often motivated by the promise of greater realism than is achievable with hand-coded decision rules[6][7].

At the same time, the growing use of LLMs as autonomous agents raises fundamental methodological challenges. First, LLMs are stochastic systems whose outputs are sensitive to prompts, sampling parameters, and contextual framing, complicating replication and systematic comparison across simulations [8][9]. Second, and more importantly, LLMs operate as opaque black boxes: the internal processes that produce agent actions are not directly interpretable, making it difficult to identify the mechanisms through which individual behavior aggregates into emergent collective outcomes. This opacity stands in tension with the core explanatory purpose of agent-based modeling, which has traditionally emphasized transparency, mechanistic clarity, and the ability to link outcomes to explicit assumptions [10][11].

These issues are closely connected to long-standing concerns about validation and credibility in computational modeling. For ABMs, validation typically involves demonstrating that a model reproduces known stylized facts, responds plausibly to parameter changes, and aligns with theoretical expectations or empirical data [12][13]. When decision-making is delegated to LLMs whose internal logic is inaccessible, such validation becomes difficult: it is often unclear whether observed dynamics reflect meaningful social mechanisms or artifacts of prompting, training data, or model architecture. As a result, the external validity and scientific utility of LLM-driven generative simulations remain contested[9].

Taken together, these developments reveal a tension at the heart of generative agent-based modeling. On the one hand, LLMs offer a powerful new resource for representing linguistic interaction, interpretation, and social complexity. On the other hand, embedding LLMs directly as agents risks undermining the transparency, interpretability, and researcher control that have historically distinguished ABMs as tools for explanation rather than mere imitation. Resolving this tension is essential if LLMs are to contribute to cumulative theory building in the social sciences.

In this study, we propose an alternative paradigm for integrating LLMs into agent-based modeling that preserves the explanatory strengths of classical ABMs while leveraging the generative capacities of LLMs. Rather than treating LLMs as autonomous decision-making agents, we use them as rule generators during the model construction phase. Given a problem context and theoretical description, LLMs are employed to propose explicit behavioral rules, identify relevant state variables, and translate informal theoretical assumptions into executable model

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

components. These outputs are externalized, inspectable, and subject to human evaluation, revision, or rejection before they are incorporated into a simulation.

This design reassigns the role of LLMs from actors within the simulation to collaborators in the modeling process. By externalizing decision logic into explicit rules, the approach restores the link between assumptions, mechanisms, and outcomes that is central to explanatory modeling[14] . At the same time, it allows LLMs to contribute where they are most valuable: synthesizing theoretical descriptions, generating plausible behavioral hypotheses, and accelerating exploratory model development without relinquishing scientific control.

We demonstrate the utility of the proposed framework through a set of case studies, including the reproduction of canonical results from established agent-based models (e.g., the Schelling Segregation Model) as well as exploratory analyses in less-established domains. These experiments show that LLM-assisted rule generation can identify valid behavioral mechanisms and construct original, mechanism-driven models, while preserving the transparency and validity of the model.

2 GABMs and Social Simulation

Agent-based modeling (ABM) is grounded in a generative conception of explanation: social regularities are explained by specifying the micro-level mechanisms and interaction structures from which macro-level patterns emerge [14]. Rather than estimating relationships directly from data, ABMs formalize theoretical assumptions about behavior, interaction, and structure, and evaluate whether these assumptions are sufficient to reproduce observed phenomena. This emphasis on explicit mechanisms has made ABMs a distinctive tool for studying complex, non-linear social systems, from segregation and diffusion to collective action and polarization [15][2][3].

At the same time, ABMs have long faced a tension between realism and explainability. Classical rule-based models are often criticized for relying on stylized behavioral assumptions that inadequately capture human cognition, emotion, and meaning-making [15][16]. Yet increasing behavioral complexity typically comes at the cost of transparency, calibration, and validation, complicating the link between assumptions and outcomes. As a result, ABMs have historically occupied an uneasy position within the social sciences, valued for theoretical exploration but often viewed as weakly grounded empirically [13][11].

The recent rise of large language models (LLMs) has reconfigured this landscape. Because LLMs can generate context-sensitive language, emulate social reasoning, and draw on vast stores of cultural knowledge, they have been proposed as a way to overcome the behavioral limitations of traditional ABMs. This has led to the rapid emergence of generative agent-based models, in which LLMs are embedded directly as agents capable of planning, remembering, communicating, and adapting through natural language [4][5]. These models promise unprecedented expressive power and have been applied to domains such as online discourse, norm formation, polarization, and collective sense-making.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

However, as recent reviews emphasize, this shift also amplifies long-standing methodological challenges rather than resolving them [17] . When LLMs are used as autonomous agents, decision-making is governed by opaque, high-dimensional inference processes that are difficult to interpret, replicate, or systematically validate. Validation practices in the emerging literature often rely on face validity, qualitative plausibility, or weakly coupled outcome comparisons, rather than direct tests of the mechanisms the models purport to capture. As a result, generative ABMs risk occupying an ambiguous methodological space: too complex to be parsimonious explanatory models, yet insufficiently grounded to function as empirical simulations.

From a philosophy-of-science perspective, this tension reflects a mismatch between simulation realism and explanatory control. While LLM-driven agents may produce behavior that appears human-like, such realism does not by itself constitute explanation. For generative models to contribute to cumulative knowledge, their assumptions must be explicit, inspectable, and open to validation relative to the phenomena they aim to explain [12][14]. Without this transparency, it becomes difficult to distinguish genuine emergent dynamics from artifacts of prompting, training data, or stochastic variation.

The framework developed in this study is grounded in this mechanism-centered view of explanation. Rather than treating LLMs as black-box decision-makers, we conceptualize them as theory-to-rule translators that assist researchers in formalizing behavioral assumptions. By externalizing LLM outputs as explicit rules, variables, and decision structures, the approach preserves the core epistemic commitments of agent-based modeling—mechanistic clarity, reproducibility, and validation—while leveraging LLMs' capacity to synthesize theory, generate plausible hypotheses, and accelerate exploratory model development. In doing so, it repositions LLMs from autonomous actors within simulations to methodological instruments embedded in the scientific modeling process itself.

3 XABM Modeling Framework

The proposed XABM framework comprises three tightly coupled modules that together support an end-to-end modeling workflow, spanning conceptualization, implementation, and evaluation. Each module is supported by one or more LLM-empowered AI agents that perform a distinct function. An orchestrating LLM will be in charge of the coordination and communication between the human researcher and the AI agents. This orchestrating LLM is also responsible for verifying the input, calling the appropriate LLM agent and returning the output. Figure 1 provides an overview of the framework architecture and the flow of operation.

The framework is designed to be used by researchers who begin with a clearly articulated research problem, including the phenomenon of interest, the simulation setting, and the basic model structure (e.g., agent types, environments, and interaction topology). Where available, researchers may also specify preliminary

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

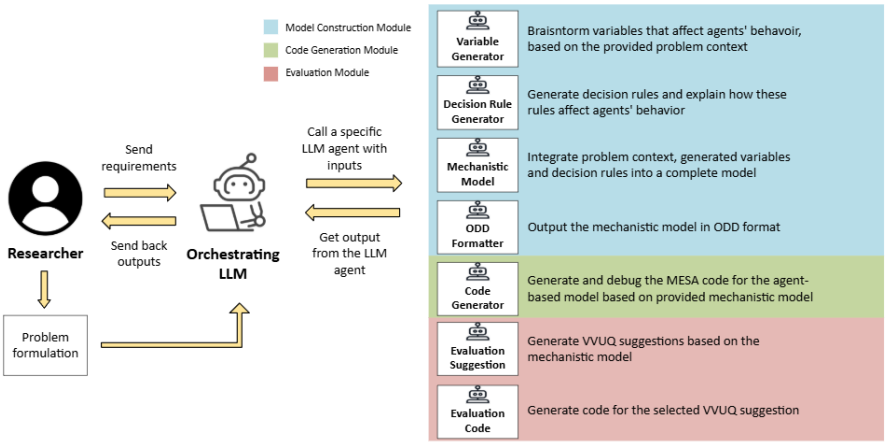

Fig. 1: The overall architecture of the XABM framework

theoretical expectations or stylized outcomes, which serve to constrain and guide the generative process.

The Model Construction Module supports theory formalization and conceptual model construction. Given the problem context and initial model specification, AI agents in this module identify decision-relevant variables, generate decision rules that affect model agents' behavior, and export the complete conceptual model in a scientifically rigorous format (e.g., ODD format). Researchers can introduce additional variables, revise assumptions or explore alternative mechanisms by communicating with the orchestrating LLM.

The Code Generation Module translates the formalized rule set into an executable simulation prototype, using standardized agent-based modeling frameworks such as MESA. Prior to execution, researchers may specify the types of data to be recorded and the desired output formats. The generated implementation serves as an initial prototype rather than a finalized computational model. Researchers are expected to review, modify and extend the code based on their research needs. This design choice preserves flexibility while ensuring that the correspondence between theoretical assumptions and computational implementation remains explicit.

The Evaluation Module supports model verification, validation, and uncertainty quantification (VVUQ). Based on the conceptual model, it first generates structured recommendations for sensitivity analysis, robustness testing, and uncertainty assessment. It then produces corresponding code implementations for these evaluation procedures.

Together, these three modules form an iterative pipeline that separates generative assistance from decision authority. By externalizing behavioral logic, implementation choices, and validation considerations, the XABM framework enables the use of LLMs in agent-based modeling without sacrificing transparency,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

reproducibility, or explanatory control.To further mitigate the black-box risks associated with LLMs and to enhance the transparency, traceability, and reproducibility of generated models, all prompts (both user and system), user feedback loops, and model outputs will be systematically logged and documented after each conversation. This process ensures that each stage of model development can be reconstructed, audited, and replicated by independent researchers.

4 Case Study: Shelling Segregation Model

The framework advances a key methodological claim: LLMs can support agent-based modeling without acting as opaque decision-makers by externalizing behavioral logic into explicit, inspectable rules. To evaluate this claim, we apply the framework to a series of case studies, beginning with Schelling's segregation model [18]. As a well-established benchmark with extensively documented assumptions and dynamics, it provides a stringent test of whether the framework can recover known explanations without hard-coded knowledge. The objective of this case study is to assess the epistemic fidelity: whether the XABM framework can (i) identify core decision variables and behavioral rules in this context, (ii) translate them into a transparent and executable simulation, (iii) reproduce Schelling model's characteristic emergent patterns, and (iv) generate meaningful extensions beyond the original formulation. See Appendix for the specific problem formulation.

	Schelling (1971)	**XABM**
Variables	– **Tolerance**: limit of proportion of neighbours that are outgroup. – **Racial composition neighbours (RCN)**: proportion of neighbours that have the same race as the target agent.	– **Homogeneity preference (HP)**: preference for homogeneous setting, $[0, 1]$. – **Racial composition neighbours (RCN)**: proportion of same-race neighbours, $[0, 1]$. – **Vacant spot availability (VSA)**: proportion of neighbouring vacant sites, $[0, 1]$.
Decision rules	– If $(1 - \text{RCN}) > \text{Tolerance}$, then Move. – If $(1 - \text{RCN}) < \text{Tolerance}$, then Stay.	– Decision signal $= \alpha \cdot \text{RCN} + (1 - \alpha) \cdot \text{VSA}$. – If Decision signal $<$ HP, then Move. – If Decision signal $\geq$ HP, then Stay.

Table 1: Comparisons between the behavioural rules described in Schelling's paper and the ones generated by the LLM framework

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

Table 1 compares the behavioral rules generated by the framework with those specified in Schelling's original formulation. The framework correctly identifies all key decision variables, including neighborhood composition, tolerance thresholds, and relocation behavior. While the terminology may differ, the underlying logic closely mirrors the original model. In addition, the framework introduces an explicit decision variable capturing the availability of vacant locations. In Schelling's extended formulations, vacancy was treated as an environmental condition rather than an agent-level consideration. Its explicit inclusion here illustrates how the framework can surface latent assumptions and render them inspectable as part of the agents' decision process—an example of how rule externalization can increase conceptual clarity without altering the core mechanism.

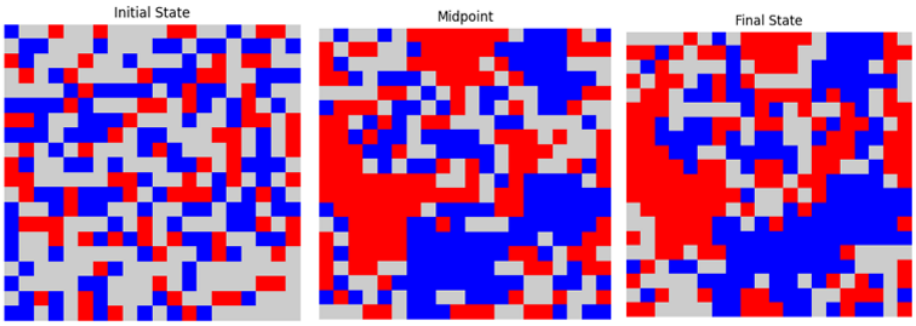

Fig. 2: Two-dimensional grid showing agents' spatial distribution at three simulation stages: (1) initialization (t = 0), (2) midpoint, and (3) final time step (t = 600). Red and blue cells represent the two agent types, while gray cells indicate vacant spaces. The grid has a size of 20 × 20, containing 300 agents (150 of each type) and 100 empty cells. This is consistent with Schelling's original recommendation that approximately 25 to 30 percent of the cells are vacant, so that the agents have enough space to move.

Figure 2 shows the evolution of spatial patterns over time as produced by the framework's code generator. The emergent dynamics closely resemble those reported by Schelling: agents self-organize into increasingly homogeneous clusters, producing clear segregation patterns despite relatively mild individual preferences. One procedural difference is that agents in the generated model relocate to randomly selected vacant cells, whereas in Schelling's original formulation they move to the nearest satisfactory location. Despite this difference, the qualitative dynamics and macro-level outcomes remain consistent, underscoring the robustness of the underlying mechanism.

We further evaluate the model with a quantitative metric: the segregation index, which is defined as the proportion of agents that are surrounded by neighbors of the same opinion. Table 2 summarizes the evaluation recommendations generated by the validator module. Based on this, a sensitivity analysis was conducted on the homogeneity preference parameter. According to Figure 3, the

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

segregation index peaks when homogeneity preference lies between 0.5 and 0.6 and reaches its minimum at very high (>0.7) or very low (<0.1) values. This pattern is consistent with Schelling-type dynamics: high preference leads to persistent dissatisfaction and frequent relocation, resulting in a disordered system, while low preference yields minimal mobility, so the system preserves its initial heterogeneous configuration.

Stochasticity Control	**Parameter Sensitivity Analysis**	**Uncertainty Quantification**
– Monte Carlo simulation – 100 runs with different random seeds – Compute mean and standard deviation – Output metrics: segregation level, neighbourhood composition – Stability if SD < 0.05 – Check convergence via distribution plots	– One-at-a-time (OAT) method – HP: vary from 0.0 to 1.0 (step 0.1) – Vacancy rate: vary from 0.25 to 0.30 (step 0.01) – 50 runs per parameter value – Output metrics: segregation level, moves per agent – Visualize trends with plots	– Target metric: segregation level – Bootstrap resampling – 1000 bootstrap samples – 95% confidence interval – No parametric distribution assumed – Captures stochastic uncertainty

Table 2: Summary of robustness analysis methods: stochasticity control, parameter sensitivity analysis, and uncertainty quantification.

Figure 4 illustrates the stochastic variability of the segregation index across 50 independent simulation runs. The mean segregation index is approximately 0.35, indicating a moderate level of segregation. The variability across runs highlights the stochastic nature of the model. Differences in initial conditions and random interactions produce divergent but bounded outcomes.

In summary, the XABM framework is able to provide scientific insight in the context of a well-established agent-based model by identifying and formalizing the core decision-making processes underlying agent behaviour, translating these processes into an executable simulation, and capturing key system properties, including stochastic variability and robustness across different parameter settings.

However, a key limitation of relying on a well-established model is that its structure, mechanisms, and common extensions may already be embedded in the LLM's training data. This introduces the risk of data leakage, whereby the model may implicitly recognize the Schelling setup and reproduce familiar formulations rather than deriving them through genuine reasoning. Consequently, the generated outputs may reflect memorized knowledge rather than true mechanistic inference. To mitigate this concern, we conduct an additional case study

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

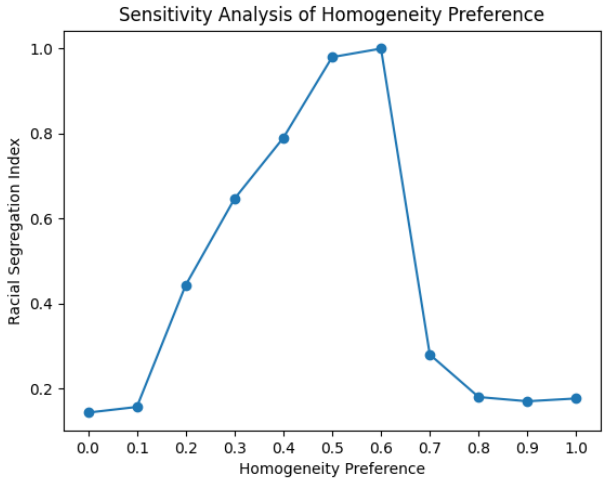

Fig. 3: Sensitivity analysis of homogeneity preference parameter (HP), which was varied from 0 to 1 in increments of 0.1, resulting in 11 experimental conditions. For each condition, the model was simulated over 500 timesteps with a population of 300 agents on a 20×20 lattice grid.

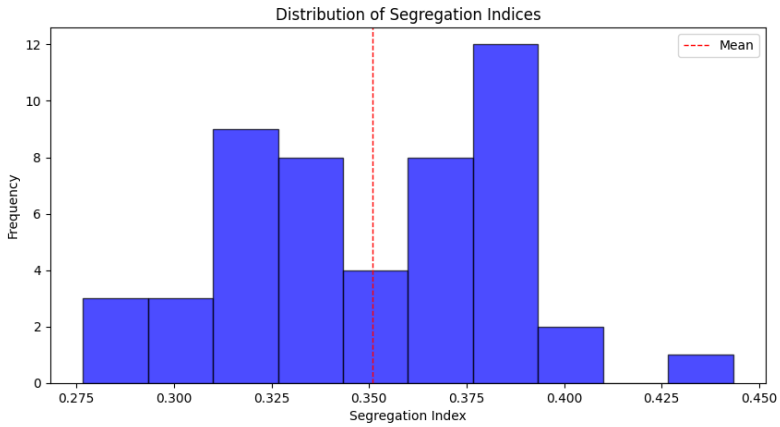

Fig. 4: Distribution of segregation index across 50 independent simulation runs: the model was initialized with 300 agents and executed for 300 timesteps, with individual homogeneity preferences drawn from a uniform distribution.

in a less formalized domain within computational modeling: spiral of silence dynamics in online social networks. This setting allows for a more rigorous evaluation of the LLM's generative reasoning and its capacity to produce original, mechanism-driven models.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

5 Case Study: Sprial of Silence Dynamics

The spiral of silence theory posits that individuals' willingness to express their opinions is shaped by their perception of prevailing opinion climates and their fear of social isolation [19]. Previous computational studies have shown that local silencing dynamics can cascade to the widespread suppression of minority viewpoints over time [20][21]. In this case study, we examine whether these dynamics persist in a heterogeneous environment where human users and LLM-driven agents coexist. In addition, we investigate whether a minority-boosting intervention can mitigate the convergence toward silence among minority opinion groups.

In this agent-based model, each agent is characterized by two attributes: identity (human or LLM-driven) and a fixed binary opinion. Human agents experience a fear of social isolation, modeled as an individual-specific expression threshold that determines whether they publicly express their opinion based on their perception of the prevailing opinion distribution. In contrast, LLM-driven agents are assumed to have no such constraint and therefore always express their opinions. Agents are situated on a two-dimensional lattice grid with Moore neighborhoods and are randomly initialized according to predefined proportions of agent types and opinions; in this study, the population consists of 80% human agents and 20% LLM-driven agents, with an initial opinion distribution of 60% versus 40%. A system-level media acts as a filtering mechanism that collects all expressed opinions at each timestep and redistributes them to agents, who sample from this pool to form their perception of the opinion environment. Two media conditions are considered: a no-intervention setting, where all messages are published as collected, and a minority-boosting setting, where the visibility of under-represented opinions is selectively amplified. See Appendix for the specific problem formulation. Table 3 and Algorithm 1 show the variables and behaviour rules the XABM framework generates, based on the above problem formulation.

Variable Name	Definition	Type
Social connection strength	The strength or closeness of social connections an agent has with its neighbors.	float [0,1]
Perceived public opinion	The agent's perception of the majority opinion in the environment based on media signals.	float [0,1]
Fear of isolation	The degree to which an agent avoids expressing opinions due to potential social isolation.	float [0,1]
Confidence level	The stability of an agent's opinion; higher confidence reduces the likelihood of remaining silent.	float [0,1]

Table 3: Key variables generated by the XABM framework for the spiral of silence case study

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

Algorithm 1 Agent Decision Making Process

```
Compute perceived support as a weighted combination of:
    (i) global opinion alignment from media
    (ii) local opinion alignment from neighbors
Compare perceived support with the agent's fear of isolation threshold
if perceived support is lower than fear of isolation then
    Decrease the agent's confidence level
    if confidence level remains above a minimum threshold then
        Agent continues to express its opinion
    else
        Agent remains silent
    end if
else
    Increase the agent's confidence level
    Agent expresses its opinion
end if
```

Simulation results are presented in Figure 5. In the absence of a minority-boosting mechanism, the silence ratio among human agents in the minority group (opinion 0) increases to approximately 80%, while the corresponding ratio for the majority group (opinion 1) remains low and continues to decrease due to rising confidence levels. With the introduction of the boosting mechanism, even at low amplification levels, the disparity between the two groups begins to diminish. As the amplification strength increases, this gap eventually disappears, and the majority group converges to a higher yet stable silence ratio. This pattern demonstrates that minority-boosting interventions can effectively counteract the self-reinforcing dynamics of silence by sustaining minority expression, thereby preventing the system from converging to a polarized state in which one opinion becomes systematically suppressed.

In summary, the XABM framework is capable of generating meaningful scientific insights even in less well-established domains with relatively simple study designs. This demonstrates its capacity for generative reasoning and its ability to construct original, mechanism-driven models. At the same time, it shows that externalizing behavioral logic into explicit and inspectable rules preserves both the generative strengths of LLMs and the transparency and replicability of the resulting models.

All framework modules, experimental scripts, and output plots are documented in a private GitHub repository, which will be made publicly available upon project completion. The repository can be shared upon reasonable request.

6 Discussion and Conclusion

This paper proposes XABM: an alternative approach to generative social simulation that directly addresses persistent challenges of transparency, validation, and explainability. Rather than embedding large language models as opaque

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

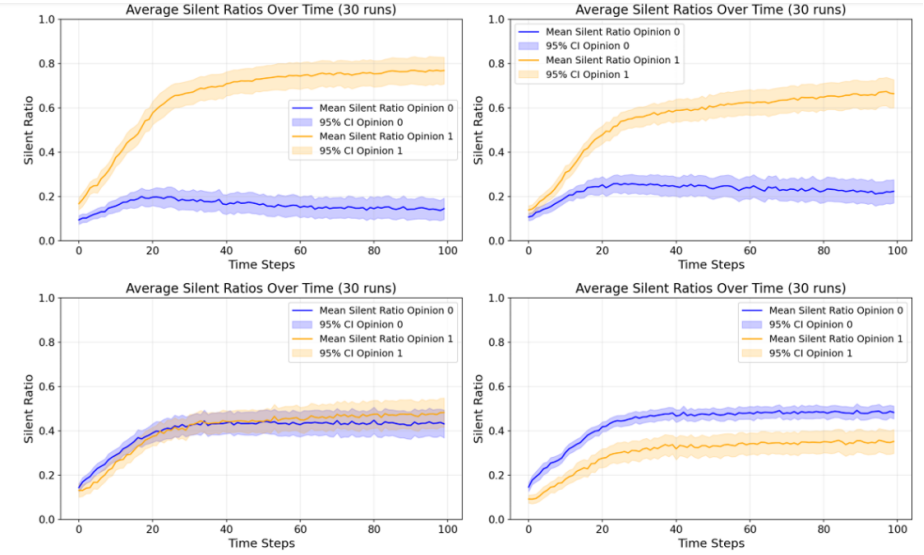

Fig. 5: Silence ratio among human agents across both opinion groups under different media conditions: no boosting (top-left), low boosting (top-right), medium boosting (bottom-left), and high boosting (bottom-right). Opinion 1 represents the majority group, while Opinion 0 represents the minority group. Each simulation was conducted with 100 agents over 100 timesteps, with results averaged across 30 independent runs. Shaded areas indicate 95% confidence intervals.

decision-makers within agents, the framework integrates LLMs as human-guided collaborators that externalize and formalize behavioral rules. This design resolves a central tension in generative modeling: how to harness the expressive capacity of LLMs without sacrificing interpretability, reproducibility, or explanatory control.

A key contribution of the framework is its ability to make implicit modeling choices explicit. By surfacing decision-relevant variables that are often treated as background conditions, the approach increases the inspectability of assumptions and clarifies how micro-level rules generate macro-level outcomes. This is particularly important in agent-based modeling, where informal assumptions can easily become embedded in code without being articulated or evaluated.

We evaluate the framework through a series of case studies, including Schelling's segregation model and spiral of silence dynamics in a heterogeneous human–LLM environment. The results demonstrate that the XABM framework can identify core decision variables, formalize them into explicit behavioral rules, and generate meaningful insights in different contexts. These findings establish a baseline level of epistemic fidelity and suggest that LLM-assisted rule generation can enhance, rather than undermine, explanatory modeling. Future work will focus on extending the framework's capabilities, such as integrating literature retrieval

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

functions, to further support the construction and validation of agent-based models.

The contribution of this work is not a claim that LLMs can replace theory, nor an attempt to produce more realistic simulations. Instead, it offers a methodological reorientation: LLMs are positioned as tools for theory formalization rather than as opaque autonomous agents whose internal reasoning is inaccessible. In doing so, the framework preserves the core epistemic commitments of agent-based modeling while extending its practical reach. It provides a principled alternative to prevailing LLM-as-agent approaches, which often conflate surface realism with explanation and face persistent validation challenges.

More broadly, this study contributes to ongoing debates about the role of generative AI in scientific modeling. As LLMs are increasingly adopted across the social sciences, the question is not whether they will be used, but how. The framework presented here demonstrates that generative AI can be integrated in ways that enhance transparency and explanatory control rather than erode them. The scientific value of generative agent-based modeling, we argue, depends not on how human-like agents appear, but on how clearly their behavior, and its consequences, can be calibrated, validated, explained, and understood.

Acknowledgments. This work was conducted as part of the first author's master's graduation project. No external funding was received for this research. The authors acknowledge the valuable support and feedback received during the development of this project.

Disclosure of Interests. The authors declare no competing interests

References

1. J. M. Epstein and R. Axtell, "Growing artificial societies: Social science from the bottom up," *Computers & Mathematics with Applications*, vol. 33, no. 5, p. 127, 1996.
2. M. W. Macy and R. Willer, "From factors to actors: Computational sociology and agent-based modeling," *Annual Review of Sociology*, vol. 28, no. 1, p. 143, 2002.
3. S. F. Railsback and V. Grimm, *Agent-based and individual-based modeling: A practical introduction*. 2019.
4. J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, ACM, 2023.
5. L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, 2024.
6. C. Gu, L. Luo, Z. R. Zaidi, and S. Karunasekera, "Large language model driven agents for simulating echo chamber formation," *arXiv preprint*, 2025.
7. C. Wang, Z. Liu, D. Yang, and X. Chen, "Decoding echo chambers: Llm-powered simulations revealing polarization in social networks," *arXiv preprint*, 2024.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

8. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, ACM, 2021.

9. P. Taillandier, J.-D. Zucker, A. Grignard, B. Gaudou, N. Q. Huynh, and A. Drogoul, "Integrating llm in agent-based social simulation: Opportunities and challenges," *arXiv preprint*, 2025.

10. V. Grimm, E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, and D. L. DeAngelis, "Pattern-oriented modeling of agent-based complex systems: Lessons from ecology," *Science*, vol. 310, no. 5750, pp. 987–991, 2005.

11. F. Squazzoni, J. G. Polhill, B. Edmonds, P. Ahrweiler, P. Antosz, G. Scholz, and N. Gilbert, "Computational models that matter during a global pandemic outbreak: A call to action," *Journal of Artificial Societies and Social Simulation*, vol. 23, no. 2, p. 10, 2020.

12. V. Grimm, U. Berger, D. L. DeAngelis, J. G. Polhill, J. Giske, and S. F. Railsback, "The odd protocol: A review and first update," *Ecological Modelling*, vol. 221, no. 23, pp. 2760–2768, 2010.

13. P. Windrum, G. Fagiolo, and A. Moneta, "Empirical validation of agent-based models: Alternatives and prospects," *Journal of Artificial Societies and Social Simulation*, vol. 10, no. 2, p. 8, 2007.

14. P. Hedström and P. Ylikoski, "Causal mechanisms in the social sciences," *Annual Review of Sociology*, vol. 36, no. 1, pp. 49–67, 2010.

15. J. M. Epstein, *Generative social science: Studies in agent-based computational modeling*. Princeton University Press, 2012.

16. P. T¨ornberg and J. Uitermark, "The social science of complexity," in *Seeing Like a Platform*, pp. 21–42, Routledge, 2025.

17. M. Larooij and P. Törnberg, "Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations," *arXiv preprint*, 2025.

18. T. C. Schelling, "Dynamic models of segregation," *Journal of Mathematical Sociology*, vol. 1, no. 2, pp. 143–186, 1971.

19. E. Noelle-Neumann, "The spiral of silence: A theory of public opinion," *Journal of Communication*, vol. 24, no. 2, pp. 43–51, 1974.

20. D. Vilone and E. Polizzi, "Modeling opinion misperception and the emergence of silence in online social systems," *PLoS ONE*, vol. 19, no. 1, p. e0296075, 2024.

21. D. Sohn, "Spiral of silence in the social media era: A simulation approach to the interplay between social networks and mass media," *Communication Research*, vol. 49, no. 1, pp. 139–166, 2019.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39

A Input Prompt for Behavioural Rule Generation

Problem formulation Schelling model

Research Topic: Simulating the formation of racial segregation patterns in a closed urban environment.
Context Description: Assume that people prefer a homogeneous neighbourhood setting.
Agent Characteristics:
- Two types of agents with different racial identities.
- Approximately equal population sizes for the two types.

Environment and Interactions:
- Agents are arranged in a two-dimensional grid.
- Each agent interacts with its eight neighbours.
- Agents are initially randomly distributed.
- The grid contains approximately 25–30% vacant spots.

Behavioral Decisions:
- At each timestep, agents decide whether to move to a vacant spot in the grid.

Problem formulation spiral of silence study

Research Topic: Simulating the spiral of silence phenomenon in online social environments.
Context Description: Individuals fear social isolation and are less likely to express their opinions when they perceive themselves to be in the minority. In online environments, perceptions of public opinion are shaped by both local social interactions and exposure to mass or social media content.
Agent Characteristics:
- Two types of agents: human users and LLM-driven agents.
- Human agents may choose to either speak or remain silent at each timestep.
- LLM agents always express their opinion at each timestep.
- Each agent holds a fixed opinion (either 0 or 1) throughout the simulation.

Environment and Interactions:
- Agents are arranged in a two-dimensional grid.
- Each agent interacts with its eight neighbouring agents.
- Agents are randomly distributed across the grid with respect to both type and opinion.

System-Level Process:
- A global media system aggregates opinions expressed by agents who choose to speak.
- At each timestep, the media publishes all collected messages.
- In the subsequent timestep, each agent randomly samples 10–20 messages from the media output.

Behavioral Decisions:
- At each timestep, human agents decide whether to express their opinion or remain silent based on their perception of the opinion climate.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29915-4_39