Emerging Media
2025, Vol. 3(3) 401–413
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/27523543251365678
journals.sagepub.com/home/emm
Sage

Agentic AI for Sustainable Development: Leveraging Large Language Model-Enhanced Agent-Based Modeling for Complex Policy Strategies

Jia'an Liu[1,#], Chu Chu[2,#], Yilin Zhao[3], Goshi Aoki[4], and Zhiqing Xiao[4]

Abstract

The convergence of agent-based modeling (ABM) and large language model (LLM)-driven autonomous agents presents a transformative approach to social simulation, particularly for addressing the complexities of the United Nations sustainable development goals (SDGs). ABM enables the exploration of emergent behaviors in complex systems but is limited by static, rule-based agents that fail to capture the nuances of human decision-making. LLM-based agents, equipped with adaptive reasoning, contextual awareness, and generative capabilities, offer a solution by enhancing realism, diversity, and inclusivity in simulations. This paper explores how integrating these technologies can advance policy experimentation, enabling simulations that reflect diverse cultural contexts and emergent social norms. We discuss the technical and ethical challenges of LLM-based systems, including reasoning limitations, hallucinations, and alignment constraints, and propose strategies for governance that balance innovation with accountability. This paper, based on insights from the UNU Macau AI Conference 2024 session titled "AI Agents in Practice: Harnessing AI for All," advocates for interdisciplinary research and human-in-the-loop frameworks to ensure responsible AI use in sustainable policymaking.

[1] United Nations University Institute in Macau, Macau SAR, China
[2] Journalism School, Fudan University, Shanghai, China
[3] Guanghua Law School, Zhejiang University, Hangzhou, China
[4] College of Computer Science and Technology, Zhejiang University, Hangzhou, China

[#] These authors contributed to the work equally and should be regarded as co-first authors.

Corresponding Author:
Jia'an Liu, Casa Silva Mendes, Estrada do Engenheiro Trigo No 4, Macao SAR, China.
Email: jiaan@unu.edu

Keywords
agent-based modeling, large language models, social simulation, sustainable development goals, collective intelligence

Received: January 2, 2025; revised: May 27, 2025; accepted: July 4, 2025

Introduction

The world's most pressing social, environmental, and economic challenges—embodied in the United Nations (UN) sustainable development goals (SDGs)—demand innovative methodologies that can capture and simulate complex human behaviors. Agent-based modeling (ABM) has been instrumental in exploring how individual actions can collectively drive societal outcomes. However, conventional ABM agents often rely on fixed, simplified rules that cannot capture the nuanced, ever-evolving nature of human decision-making, making it difficult to fully represent phenomena such as cultural differences, emergent norms, and complex policy responses.

The UN SDGs represent a global consensus on the most critical social, environmental, and economic challenges confronting humanity. These goals are inherently complex, characterized by deep interconnections, feedback loops, and non-linear dynamics. Tackling issues like climate change, poverty, and inequality requires policy interventions whose outcomes are often difficult to predict using traditional analytical methods. These methods frequently struggle to capture the emergent, system-level consequences that arise from the diverse decisions and interactions of myriad individual actors within a society. It is this very complexity, where the whole often behaves differently than the sum of its parts, that demands innovative methodologies capable of simulating human behavior and societal dynamics in a more granular and realistic fashion.

ABM is a computational approach that offers a "bottom-up" perspective, constructing simulations not from overarching equations but from the actions and interactions of individual "agents." These agents serve as computational representations of real-world actors such as individuals, households, businesses, or even government bodies. Each agent can be endowed with specific characteristics (such as age, income, beliefs, or preferences) and decision-making rules that govern their behavior. By placing these agents within a simulated environment and allowing them to interact with the environment and each other by exchanging information, competing for resources, forming networks, ABM allows researchers to observe how large-scale patterns and societal outcomes, such as market trends, disease spread, or social norm shifts, emerge directly from micro-level activities. This makes ABM an invaluable tool for exploring how complex social systems function and why certain policies might succeed or fail.

Despite its strengths, conventional ABM faces a significant limitation: the agents themselves are often programmed with relatively simple, static, or pre-defined rules. While useful for modeling certain behaviors, these rule-based approaches often fail to capture the full richness and adaptability of human decision-making. Real people learn, adapt, possess cultural biases, experience emotions, and respond to novel situations in ways that hard-coded rules cannot easily replicate. Concurrently, large language models (LLMs) have reshaped AI by excelling in text comprehension and generation, recent work has extended these capabilities into "agentic" systems that integrate multi-step reasoning, external tool use, and memory-based contexts (Xi et al., 2025). By incorporating a spectrum of perspectives and adapting to new information, such LLM-driven agents present a transformative opportunity to enhance ABM. It means moving beyond simple rules to agents that can: (a) *reason with context*: they can understand and react to nuanced information, including policy descriptions or news updates, in a manner more aligned with human comprehension. (b) *Exhibit* *diversity*: LLMs can be prompted or fine-tuned to represent a vast spectrum of human perspectives, cultural backgrounds, and decision-making styles, leading to more inclusive and representative simulations. (c) *Adapt and learn*: while still an area of active research, agentic LLMs can potentially modify their behavior based on simulated experiences and interactions, capturing emergent social learning and norm evolution.

This paper explores the integration of LLM-based autonomous agents within ABM frameworks as a novel approach to social simulation, especially tailored to address the complexities inherent in SDG-related policy strategies. We argue that this synergy allows for richer, more dynamic simulations that can better inform policy design and evaluation. We will examine the potential benefits, from fostering collective intelligence within simulations to ensuring the inclusion of diverse, even marginalized, perspectives. However, we also acknowledge and address the significant challenges—from the technical limitations of current LLMs, such as their potential for generating inaccurate information ("hallucinations") and the difficulties in ensuring their alignment with human values, to the pressing ethical and governance considerations. Ultimately, this research aims to delineate a path toward leveraging these advanced AI capabilities responsibly, creating more effective and equitable policy solutions for a sustainable future.

Enhancing ABM with LLM-based Generative Agents

ABM provides a potent framework for understanding complex systems by simulating the interactions of individual actors. In ABM, an "agent" serves as a computational entity, typically characterized by specific attributes (such as age, location, income, or beliefs) and a set of decision rules that dictate its behavior, like movement or interaction strategies (Dorri et al., 2018). But real-world situations, such as navigating a crisis or adopting a new policy, often depend heavily on how individuals interpret evolving information, weigh conflicting goals, and respond based on their unique experiences and cultural backgrounds—dynamics that conventional models struggle to capture fully (Liang et al., 2022).

Given the complexity of SDG targets, AI can be significantly useful in advancing sustainable development (Vinuesa et al., 2020), though earlier AI successes often demonstrated superhuman capabilities within highly constrained domains. Modern LLMs exhibit a more generalized capacity for understanding and reasoning, built upon training with vast amounts of textual data. Through innovations like sophisticated prompting techniques that guide their reasoning processes, these models can generate human-like narratives, analyze complex scenarios, and engage in step-by-step problem-solving (Wei et al., 2022), providing a foundation for more flexible, general-purpose intelligence. A particularly relevant development in this space is the concept of "Generative Agents," which specifically aims to simulate human-like behaviors within virtual social environments based on generated texts for "next-step" instructions (Park et al., 2023). These agents are designed to maintain internal states, including memories, goals, and even simulated emotional responses, allowing them to engage in complex social interactions. They can form relationships, develop trust or conflict, and participate in collective dynamics, leading to the emergence of phenomena like community norms or cultural diffusion. This focus on simulating rich inner lives and social interactions makes generative agents particularly well-suited for integration into ABM frameworks.

This approach opens new avenues for policymakers seeking to navigate the complex interdependencies inherent in the SDGs. By integrating LLM-driven agents, models can be populated with actors representing a rich tapestry of demographic, cultural, and socioeconomic perspectives, which is essential for understanding how policies impact diverse communities, which is a cornerstone of the SDG agenda. These agents can be endowed with access to up-to-date knowledge through retrieval mechanisms and empowered with decision-making processes that transcend simple rule-based heuristics, allowing for a more realistic simulation of human responses to policy shifts. Consequently, when modeling the potential outcomes of interventions targeting specific SDGs, such as poverty reduction (SDG 1), climate action (SDG 13), or promoting inclusive institutions (SDG 16), it becomes possible to consider their intricate economic, social, and environmental ripple effects with greater fidelity. Therefore, these enhanced models can become significantly more descriptive and capable of reflecting diverse, context-sensitive human responses (Sanders et al., 2023). This capability allows policymakers to experiment with multifaceted scenarios, exploring how interventions might unfold within varied societal contexts and capturing emergent outcomes that reflect the genuine complexities of human behavior, thereby fostering more robust and equitable strategies for achieving sustainable development.

System Architecture

The integration of LLMs into ABM necessitates a carefully designed system architecture that allows seamless communication and control flow between the ABM simulation environment and the LLM services. This work proposed a modular architecture (Figure 1). The ABM environment manages the overall simulation clock, agent populations, spatial environment, and interaction protocols. LLM capabilities are encapsulated as a distinct module or service that agents can invoke.

The ABM Core Engine forms the foundational layer, managing the discrete-event simulation environment. It is responsible for maintaining the agent population, representing the simulated physical and social environment, scheduling agent actions and interactions, and updating the global state of the simulation at each time step. Integrating LLMs into an ABM framework requires a carefully decoupled yet cohesive architecture that preserves the strengths of existing ABM libraries while enabling rich, context-aware decision-making via LLMs.

Rather than reimplementing well-tested ABM primitives as in scratch engines such as AgentSociety (Piao et al., 2025), our proposed framework builds atop existing (largely tested, with good performance metrics in common models) python-based ABM library, leveraging its flexible scheduler, grid and network representations, and rich community ecosystem. This choice reduces development overhead, ensures robust performance on standard use cases, and allows focus on agents' cognitive capabilities rather than duplicating core scheduling logic. The flexibility for defining and customizing the **behavioral models and attributes** of agents within this Core Engine enables **fine-grained control over agent definition**, and management of the agents' action and simulation steps. This engine drives the discrete-event simulation, handling: (a) **Agent population:** Instantiation, lifecycle management, and attribute updates of each agent; (b) **Orchestration and Control Unit:** Invocation of agent actions and global updates at each time step, termination conditions, and the synchronized operation of all other modules. It ensures the coherent flow of information and control within the framework; (c) **Spatial and social environment:** Modeled as a standalone micro-service or subprocess, the environment simulator encapsulates domain-specific dynamics—physical spaces, urban mobility, social networks, economic ledgers, and resource distributions—providing agents with deterministic, verifiable feedback. Agents issue "get" and "set" API calls (e.g., query travel times, execute transactions) via lightweight adapters, ensuring that subjective LLM reasoning is grounded in objective, real-world constraints.

LLM Interaction Layer serves as the bridge between the ABM Core Engine and context-aware decision-making based on LLM's reasoning capabilities. Its responsibilities include dynamically constructing prompts, managing API calls, and parsing LLM responses to translate them into actionable information for the ABM Core Engine. This layer is inspired by integration patterns

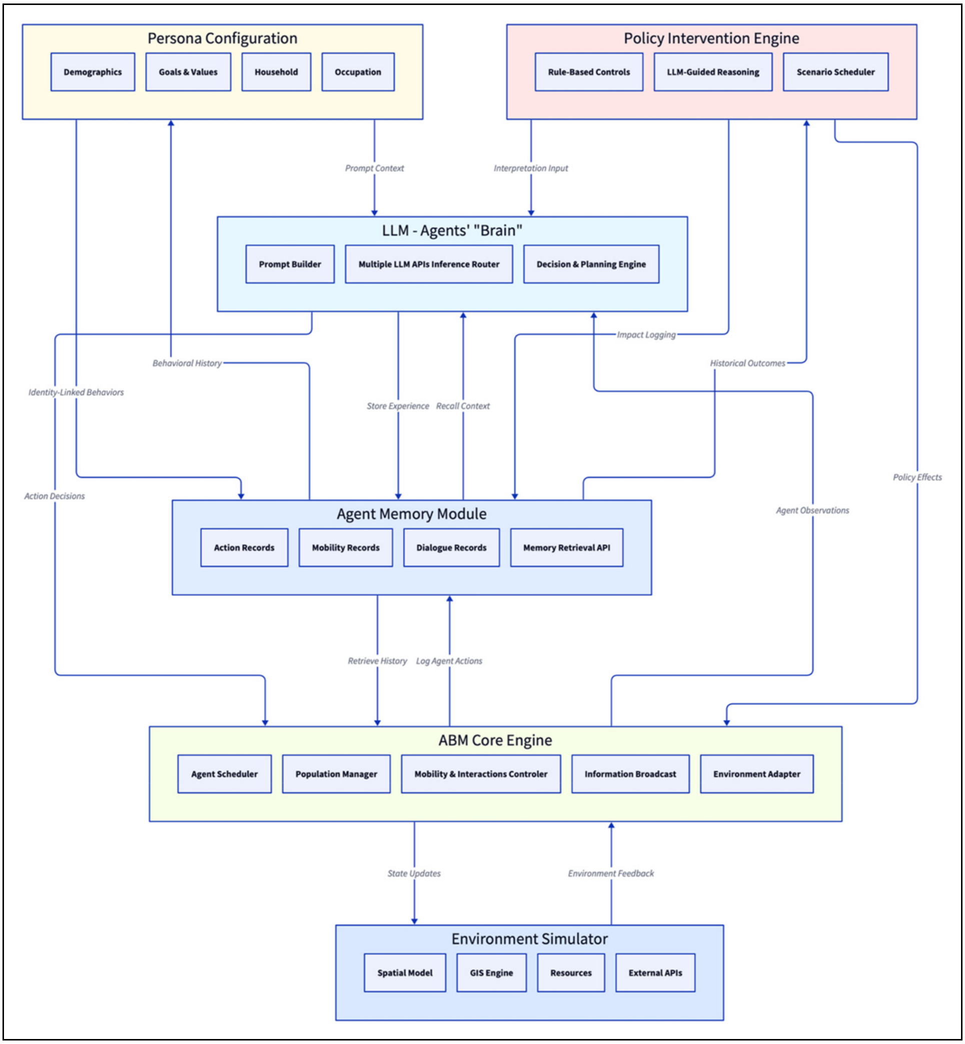

Figure 1. Framework Design of LLM-enhanced Agent-based Modeling System.

in practical LLM applications and agentic systems. Operationally, this layer can be conceptualized as an LLM-Service exposing micro-endpoints:

- **Prompt Builder:** Assembles a persona-specific prompt from the agent's state, neighborhood observations, and a scenario memory buffer. An example of prompt schema is:System: "You are ${role}, representing ${demographic}. Goal: maximise ${utilities} while adhering to SDG constraints ${sdg_set}." Memory: ${top-k episodic memories} Observation: ${structured obs} Task: "Decide one action and a brief rationale."
- **Inference Router:** Dispatches the prompt to (a) multiple shared high-capacity LLMs for rich reasoning, or (b) a fine-tuned small model when latency or data-sovereignty constraints apply. This allows for optimizing computational resources and response times based on agent needs and task complexity.

- **Decision and Planning Engine:** This core component orchestrates the agent's cognitive cycle. It takes the assembled prompt (including current state, observations, and retrieved memories) and leverages the selected LLM to engage in reasoning, potentially multi-step planning, and ultimately, decision-making. It processes the LLM's output, which might involve complex thought processes or reasoning, to distil a specific, structured action.

Agent Persona and Memory Modules: Policies and environmental changes rarely have a uniform effect across a population; their impact invariably differs based on the rich tapestry of characteristics that define each persona, such as their beliefs, decision-making patterns, socio-demographic profiles, and even personality traits captured by LLMs (Park et al., 2024). The effectiveness and realism of policy simulations are enhanced by the diversity and sophistication of agent personas within the model. This component acts as a repository for defining and managing the characteristics of different agent types. It stores:

- **Persona Templates:** Detailed descriptions, rules, and parameters that define the roles, goals, attributes (demographics, values, beliefs), and behavioral tendencies of various stakeholder personas. LLM-driven agent personas, following methods given by Ge et al. (2024), the perspective richness of the LLM can empower the large-scale creation of diverse personas representing up to a one-billion level of mass population.
- **Knowledge & Data Bases:** Curated sets of documents, data, or information snippets that specific agent can access via Retrieval Augmented Generation (Lewis et al., 2020) to inform their decisions or communications, making their behavior more contextually grounded and role-specific.
- **Emotional & Motivational State Tracker:** Maintains a dynamic model of each agent's affective state (e.g., stress, confidence, trust) and intrinsic motivations (e.g., risk tolerance, social approval), updating these based on past outcomes and influencing future decisions.

Policy Intervention Module: A critical function of LLM-enhanced ABM framework is its ability to simulate the effects of policy interventions. This module provides the mechanisms for modelers to introduce, schedule, and implement diverse policy scenarios within the simulation. Policies exert their influence within an agent-based system primarily by altering the conditions under which agents operate or the information they possess. These interventions can manifest as changes in the simulated environment (such as new regulations affecting resource availability), as direct signals or information campaigns targeted at agents, or as modifications to the incentive structures agents face (such as through subsidies, taxes, or penalties).

Classic methods in ABM software such as directly modifying agent attributes or altering fixed decision-making offer advantages in terms of computational efficiency and ease of interpretation. or policies where interpretation, negotiation, belief-updating, or culturally inflected responses are central, an LLM-agent can be informed of a policy change and asked to reason about its implications based on its persona, memory, and goals. This allows for modeling more nuanced responses, capturing the heterogeneity and complexity inherent in human reactions to policy. Table 1 provides a comparative overview of how these core policy mechanisms can be implemented using both rule-based and LLM-driven approaches, and specifies the primary approach within our framework, along with relevant examples. We propose using rule-based methods for the direct implementation of changes to tangible attributes and environmental parameters, leveraging their efficiency. For changes concerning decision logic, information processing and interpretation, we designate the LLM-driven approach as primary.

Table 1. Mechanisms of Policy Application in Agent-Based Systems.

Policy Mechanism	Rule-Based Implementation	LLM-Driven Implementation	Example Policy/ Event	Primary Approach
Attribute Modification	Directly alter agent state variables (e.g., agent.wealth += 100).	Update agent's profile/ memory; Prompt LLM to reason about attribute changes (e.g., "Income increased. How does this affect spending?").	Universal Basic Income	Rule-Based
Decision Logic Change	Add, remove, or modify "if-then" rules in agent behavior logic.	Modify LLM prompts/goals/ constraints; provide new guidelines (e.g., "Prioritize X due to new regulation Y.").	Mandatory Energy Standards	LLM-Driven
Environment Shift	Change global/local environmental variables (e.g., road.status = 'flooded').	Inform agents of changes via prompts/feeds (e.g., "Main road is flooded. How do you reach your destination?").	Road Waterlogging (Heavy Rain)	Rule-Based
Information Change	Alter info networks; broadcast messages; set 'informed' flag.	Filter/prioritize info; simulate campaigns via narratives.	Public Health Awareness	LLM-Driven

Data Management and Logging Toolkits: These utilities meticulously capture comprehensive data throughout each simulation run, recording the dynamic states of agents, their decisions, and interactions at each time step. Concurrently, it logs modifications to environmental variables, and it can also track spatial dynamics through logs of environmental maps or agent locations. For systems employing LLM-based agents, this layer documents the specifics of LLM invocations, including prompts, responses, and metadata like latency and token usage. This comprehensive logging enables standard debugging, thorough auditing of AI-driven decision-making, and performance analysis, which are vital for validating the model, ensuring its responsible application, and thereby providing the empirical foundation for subsequent scientific inquiry.

Multimodal Extensions for Enhanced Policy Simulation: Not all policy challenges revolve solely around textual data; some also involve imagery (e.g., satellite photos for environmental monitoring) or audio (e.g., local radio broadcasts in disaster management). Recent efforts propose LLM-driven frameworks that can parse visual, auditory, or sensor inputs alongside text (Durante et al., 2024). By fusing multiple data types, agentic systems can detect emerging crises, model resource allocation, and guide intervention strategies with greater fidelity. Such multimodal extensions hold particular promise for ABM scenarios that require a holistic grasp of diverse, real-world signals.

Leveraging LLM-Enhanced ABM: Scenarios in Policy Co-Creation, Inclusion, and Automation

The potential applications of the integrated LLM-driven ABM herald advancements for social simulation, particularly in navigating the intricate landscape of SDGs policymaking, offering novel methodologies for understanding, shaping, and even automating aspects of policy development and implementation. Our proposed framework provides a conceptual blueprint for these advanced applications (Figure 2), centering on an LLM-enhanced ABM engine designed to harness collective

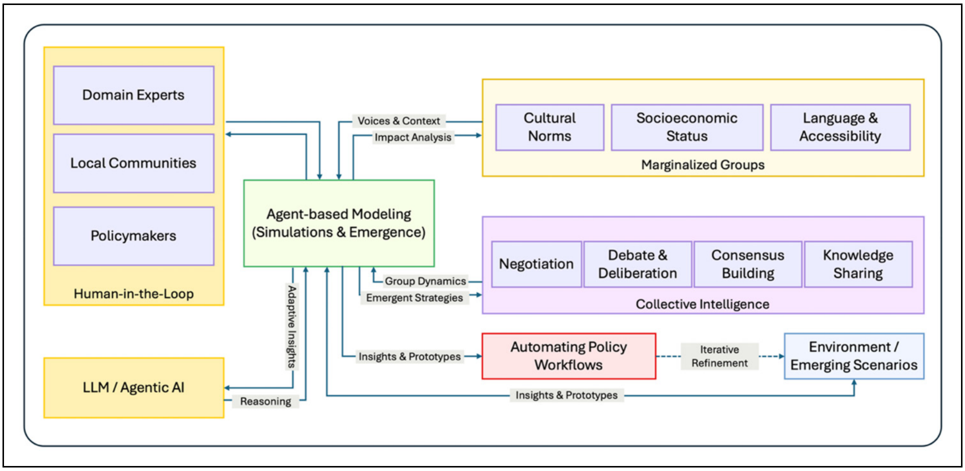

Figure 2. LLM-enhanced Agent-based Modeling Framework for Collective Intelligence, Inclusion, and Policy Automation.

intelligence, ensure the inclusion of diverse perspectives, streamline policy workflows, especially those from marginalized communities. The agentic applications critically supported by human-in-the-loop (HITL) paradigms.

The symbiotic relationship between the LLM, providing “generative reasoning” by text-based planning, and the ABM Engine, which simulates emergent behaviors and returns “adaptive insights” is fueled by data representing the Environment or Emerging Scenarios—be it climate models, economic forecasts, or public health data—allowing the simulation to dynamically respond to changing conditions. Table 2 summarizes several important application areas of our proposed ABM framework, corresponding impacts, limitations, and potential mitigations.

Collective Intelligence for Enhanced Policymaking: LLM-based agents possess the ability to simulate the views of different groups of people, enabling rich interactions that mirror collective decision-making processes. Through these interactions, agents can discuss, debate, and build consensus, effectively embodying collective intelligence. This integration facilitates outcomes that are more creative, robust, and accurate than what individual entities could achieve alone.

Agents mimicking different policy stances can negotiate, debate, and generate shared positions, mirroring real legislative or community dialogues. This process ensures that the simulated outcomes reflect a plurality of perspectives, enhancing the realism and applicability of policy simulations. Instead of top-down modeling, simulations can incorporate bottom-up intelligence, reflecting the input of thousands or millions of virtual stakeholders. This approach allows for more nuanced and representative policy prototyping, capturing the complexity of real-world stakeholder interactions.

However, if models are over-aligned to prevailing norms, the simulation may underrepresent dissenting or minority views, resulting in policies that are less robust or inclusive in the real world. Over-aligning a model to a specific set of norms could unintentionally bias its behavior or exclude alternative value systems (Senthilkumar et al., 2024).

Inclusion of Marginalized Voices for Policymaking: Crucially, for these simulations to contribute meaningfully to sustainable development, they must actively incorporate the perspectives of

Table 2. Key Application Areas of LLM-enhanced ABM: Impacts, Limitations, and Mitigations.

ABM Application Area	Key Limitations	Potential Real-World Impact	Mitigation/Strategies
Collective Intelligence Simulation	Over-alignment, Value bias, Ambiguity	Agents may neglect minority perspectives, converge to uncreative consensus	Explicit value modeling, stakeholder auditing.
Emergency Scenario ABM	Context/memory limits, Hallucinations	Agents may forget prior events, generate unrealistic sequences, miss key dynamics	Persistent agent memory, scenario curation
Inclusion of Marginalized Voices in ABM	Data bias, Alignment	Risk of stereotyping, superficial representation, or omission of crucial voices	Curated training data, participatory validation
Multimodal Policy ABM	Modality integration, Context window	Incomplete data fusion may skew outcomes; context loss in long simulations	Modular design, expert review

Marginalized Groups. However, as noted in Table 2, this endeavor faces challenges like Data bias and potential superficial representation, given the underrepresentation of these groups in many large training datasets. Traditional ABM often struggles to capture the nuanced realities of these communities. Our proposed approach addresses this by using LLMs to create agents whose behaviors reflect specific Cultural Norms, Socioeconomic Status, and Language & Accessibility barriers. This is achieved not merely through abstract rules, but by infusing agents with knowledge derived from diverse sources. Researchers, working alongside community representatives (as part of the HITL process), would develop rich profiles based on ethnographic studies, interviews, and participatory workshops. These profiles and lived experiences become the basis for highly detailed LLM prompts or, potentially, for fine-tuning smaller, specialized LLMs (where sufficient, carefully curated data exists). Members of these communities are encouraged to engage in refining and validating their digital counterparts, ensuring authenticity and preventing stereotyping. For example, an agent representing an indigenous community in a land-use simulation might be prompted to prioritize ancestral lands and non-monetary values, referencing specific cultural narratives. LLMs' multi-language capabilities can be leveraged to build agents that communicate in local languages and to make simulation outputs accessible. Projects such as the UN University's "AI Voice for the Voiceless" initiative underscore the importance of this approach, aiming to ensure these groups influence policy debates (UNU-EHS & UNFCCC Technology Executive Committee, 2024). This approach aims to embed these perspectives directly into the Impact Analysis phase, ensuring that the ABM doesn't merely simulate a society, but their society, with its unique vulnerabilities and strengths.

Human-in-the-loop: The inherent limitations of AI, such as potential biases or unrealistic "hallucinations" (Mirzadeh et al., 2024), necessitate the HITL paradigm. From the perspective of software engineering, the implementation of HITL involves creating interactive modeling interfaces. These interfaces would allow domain experts, policymakers, and local communities not just to observe, but to actively engage by triggering/pausing simulations, querying agents on their "reasoning" (i.e., examining the LLM outputs), modifying agent parameters or environmental variables, and providing direct feedback in natural language. This continuous validation and co-creation process, as explored in participatory modeling projects (Tan et al., 2022), grounds the simulation in reality. It acts as a critical mechanism for mitigating the risks outlined in Table 2, such as over-alignment or superficial representation, by ensuring human judgment and ethical oversight are constantly applied.

Automating Policy Workflows: Applications of the proposed framework extend to Automating Policy Workflows. Using the rich outputs of the ABM (simulated impacts, emergent strategies, consensus points) as inputs for LLMs specifically tasked with policy-related report writing. For instance, following a successful CI simulation, an LLM could be prompted to generate a draft policy brief summarizing the consensus points reached, highlighting potential implementation challenges identified by the "marginalized community" agents, and comparing the proposed strategy with existing policies. LLMs can thus act as powerful assistants (Gao, 2023; Gunes & Florczak, 2023), processing simulation data (Qi et al., 2024), and public feedback (Heseltine & Clemm von Hohenberg, 2024) to accelerate the drafting and analysis phases. This creates an iterative refinement loop: policy drafts can be tested in new ABM runs, allowing for rapid prototyping and adaptation before real-world implementation.

Conclusion and Discussion

The convergence of ABM with LLM-driven autonomous agents marks a pivotal step in social simulation, offering significant potential for navigating the intricate policy landscapes associated with the UN SDGs. This paper has explored how leveraging the adaptive, context-aware, and generative capabilities of LLM-based agents can enrich ABM, enabling simulations that achieve greater realism in human decision-making, foster the inclusion of diverse cultural and socioeconomic perspectives, and allow for more dynamic and nuanced policy prototyping. By simulating emergent behaviors with agents that can reason, remember, and interact in human-like ways, we can move beyond the limitations of static, rule-based models and begin to explore complex societal responses to policy interventions with unprecedented fidelity.

However, the integration of these powerful AI technologies is not without its challenges, demanding careful consideration of both technical and ethical dimensions. LLMs, despite their advancements, are still susceptible to reasoning limitations, the generation of "hallucinations," and complexities surrounding their alignment with diverse ethical frameworks and human values. These technical hurdles necessitate ongoing research into robust validation techniques, methods for ensuring the verifiability of agent behaviors, and architectures that can manage context and memory effectively within large-scale simulations.

Beyond the technical, the deployment of LLM-enhanced agentic AI for policy raises profound governance questions. The existing landscape of AI regulation, while evolving through initiatives like the EU AI Act (Kop, 2021) and U.S. guidelines, often struggles to keep pace with the rapid advancements in generative AI and its agentic applications. These frameworks, while important for establishing foundational safety and accountability standards and reflecting a desire that AI should not override human decision-making (Hagendorff, 2020), can be outpaced by technological shifts, potentially stifling innovation or failing to address specific risks associated with autonomous policy simulation. This gap highlights the indispensable role of more agile governance mechanisms, often emerging from "soft law" instruments and multi-stakeholder collaborations (Abbott & Snidal, 2000; Tallberg et al., 2023).

Effectively governing this space requires a concerted effort involving diverse actors. Technology enterprises developing the core LLM and ABM platforms bear a responsibility for building in ethical safeguards and transparency, as seen in emerging industry initiatives (Fjeld et al., 2020). Research institutions must pioneer methods for responsible development and validation. Crucially, civil society organizations, particularly those representing marginalized communities, must be integral partners, ensuring that simulations reflect lived realities and promote equitable outcomes, rather than perpetuating existing biases. Their involvement is vital for grounding agent personas and validating simulation results. We must also remain cognizant of broader societal concerns, including potential impacts on labor markets, persistent copyright and liability issues, the risk of exacerbating the digital divide, and political threats arising from misuse.

Addressing these multifaceted challenges necessitates a shift toward multi-stakeholder governance models. Such frameworks, fostering continuous dialogue between governments, industry, academia, and civil society, are best positioned to develop norms that are both ethically sound and practically implementable. A core component of this approach must be the robust implementation of HITL frameworks. As discussed, HITL is not merely a technical feature but a fundamental governance strategy, ensuring that domain experts, policymakers, and community representatives remain central to the design, validation, and interpretation of policy simulations. This ongoing human oversight is critical for mitigating AI risks, building trust, and ensuring that the outputs of these complex models serve human-defined goals.

Looking ahead, future research must prioritize interdisciplinary collaborations. Computer scientists, social scientists, legal scholars, ethicists, and policymakers must work together to refine LLM-enhanced ABM methodologies, develop rigorous validation standards, and co-create governance protocols. Deepening HITL frameworks to make them more interactive and accessible, particularly for non-technical stakeholders and marginalized communities, is paramount. By embracing these collaborative and human-centric approaches, and by aligning AI development with shared human values—anchored in international norms yet sensitive to local contexts—we can ensure that LLM-enhanced ABM evolves as a transformative and trustworthy tool, capable of supporting the complex, inclusive, and sustainable policymaking required to achieve the SDGs.

ORCID iDs

Jia’an Liu https://orcid.org/0009-0007-5129-5581
Chu Chu https://orcid.org/0009-0003-2308-0576
Yilin Zhao https://orcid.org/0000-0002-7927-2660
Goshi Aoki https://orcid.org/0000-0002-7375-3404
Zhiqing Xiao https://orcid.org/0009-0007-4889-644X

Ethical Statement

There are no human participants in this article and informed consent is not required.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abbott, K. W., & Snidal, D. (2000). Hard and soft law in international governance. *International Organization*, *54*(3), 421–456. https://doi.org/10.1162/002081800551280

Dorri, A., Kanhere, S. S., & Jurdak, R. (2018). Multi-agent systems: A survey. *IEEE Access*, *6*, 28573–28593. https://doi.org/10.1109/ACCESS.2018.2831228

Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J. S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., Ikeuchi, K., Vo, H., Fei-Fei, L., & Gao, J. (2024). Agent AI: Surveying the horizons of multimodal interaction. arXiv preprint arXiv:2401.03568.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. SSRN. https://doi.org/10.2139/ssrn.3518482

Gao, A. (2023). Implications of ChatGPT and large language models for environmental policymaking. SSRN. https://doi.org/10.2139/ssrn.4499643

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. arXiv Preprint arXiv:2406.20094.

Gunes, E., & Florczak, C. K. (2023). Multiclass classification of policy documents with large language models. arXiv Preprint arXiv:2310.08167.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, *11*(1). https://doi.org/10.1177/20531680241236239

Kop, M. (2021, September 21). *EU artificial intelligence act: The European approach to AI*. SSRN. https://ssrn.com/abstract=3930959

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (Article 793, pp. 9459–9474). Curran Associates Inc. https://dl.acm.org/doi/abs/10.5555/3495724.3496517

Liang, X., Luo, L., Hu, S., & Li, Y. (2022). Mapping the knowledge frontiers and evolution of decision making based on agent-based modeling. *Knowledge-Based Systems*, *250*, 108982. https://doi.org/10.1016/j.knosys.2022.108982

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv Preprint arXiv:2410.05229.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (Article 2, pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. arXiv Preprint arXiv:2411.10109.

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J. Y., Zhou, D., Gao, C., Xu, F., Zhang, F., Rong, K., Su, J., & Li, Y. (2025). AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. arXiv Preprint arXiv:2502.08691.

Qi, W., Lyu, H., & Luo, J. (2024). Representation bias in political sample simulations with large language models. arXiv Preprint arXiv:2407.11409.

Sanders, N. E., Ulinich, A., & Schneier, B. (2023). Demonstrations of the potential of AI-based political issue polling. arXiv Preprint arXiv:2307.04781.

Senthilkumar, P., Balasubramanian, V., Jain, P., Maity, A., Lu, J., & Zhu, K. (2024). Fine-tuning language models for ethical ambiguity: A comparative study of alignment with human responses. arXiv Preprint arXiv:2410.07826.

Tallberg, J., Lundgren, M., & Geith, J. (2023). AI regulation in the European Union: Examining non-state actor preferences. SSRN. https://doi.org/10.2139/ssrn.4424114

Tan, Y.-R., Agrawal, A., Matsoso, M. P., Katz, R., Davis, S. L. M., Winkler, A. S., Huber, A., Joshi, A., El-Mohandes, A., Mellado, B., Mubaira, C. A., Canlas, F. C., Asiki, G., Khosa, H., Lazarus, J. V., Choisy, M., Recamonde-Mendoza, M., Keiser, O., Okwen, P., ... Yap, P. (2022). A call for citizen science in pandemic preparedness and response: Beyond data collection. *BMJ Global Health*, *7*(6), e009389. https://doi.org/10.1136/bmjgh-2022-009389

UNU-EHS, & UNFCCC Technology Executive Committee. (2024, August 31). *Bonn AI & climate 2024*. United Nations University. https://unu.edu/ehs/announcement/bonn-ai-climate-2024

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-019-14108-y

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, Ed H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems* (vol. 35, pp. 24824–24837). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ... Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, *68*(2), 121101. https://doi.org/10.1007/s11432-024-4222-0