frontiers in Artificial Intelligence

TYPE Review
PUBLISHED 17 November 2026
DOI 10.3389/frai.2026.1988865

Check for updates

OPEN ACCESS

EDITED BY
Mirko Degli Esposti,
University of Bologna, Italy

REVIEWED BY
Pantaleon Fassbender,
Twisters Management Consulting LLC, United States
Zewei Li,
University of Leeds, United Kingdom

*CORRESPONDENCE
Patrick Taillandier
patrick.taillandier@inrae.fr

RECEIVED 04 September 2026
REVISED 06 October 2026
ACCEPTED 30 October 2026
PUBLISHED 17 November 2026

CITATION
Taillandier P, Zucker J-D, Grignard A, Gaudou B, Huynh QN, Kong H and Drogoul A (2026) From the fluency fallacy to the micro-to-macro validity gap: opportunities and pitfalls of LLMs in social simulation.
Front. Artif. Intell. 9:1988865.
doi: 10.3389/frai.2026.1988865

COPYRIGHT
© 2026 Taillandier, Zucker, Grignard, Gaudou, Huynh, Kong and Drogoul. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

From the fluency fallacy to the micro-to-macro validity gap: opportunities and pitfalls of LLMs in social simulation

Patrick Taillandier[1]*, Jean-Daniel Zucker[2], Arnaud Grignard[2,3], Benoit Gaudou[4], Quang Nghi Huynh[3,5], Haojia Kong[1] and Alexis Drogoul[2,3]

[1]Unité de Recherche Mathématiques et Informatique Appliquées de Toulouse (UR MIAT), University of Toulouse, INRAE, Castanet-Tolosan, France, [2]Unité Mixte Internationale de Modélisation Mathématique et Informatique des Systémes Complexes (UMI 209 UMMISCO), Institut de Recherche pour le Développement (IRD)/Sorbonne University, Bondy, France, [3]Laboratoire Mixte International ACROSS (LMI ACROSS), Thuyloi University, Hanoi, Vietnam, [4]Unité Mixte de Recherche 5505 Institut de Recherche en Informatique de Toulouse (UMR 5505 IRIT), University of Toulouse Capitole, Toulouse, France, [5]College of Information and Communication Technology (CICT), Can Tho University, Can Tho, Vietnam

The integration of Large Language Models (LLMs) into social simulation has generated considerable enthusiasm, but also raises substantial methodological and epistemological challenges. This critical review examines the use of LLMs as cognitive or decision-making components of simulated agents from a computational social science perspective. Rather than treating the psychological evaluation of LLMs as separate from simulation mechanics, we argue that their behavioral and epistemic limitations can become systemic vulnerabilities when scaled to multi-agent societies. We first map the rapidly evolving landscape of LLM-driven platforms, ranging from small narrative sandboxes to population-scale and spatially structured simulations. We then develop a unified critical framework for analyzing the Micro-to-Macro Validity Gap: the propagation, nonlinear amplification, or convergent contraction of micro-level limitations—including hallucinations, stochastic inconsistency, representational biases, and alignment effects—into agent-level distortions (such as average persona convergence) and macro-level risks (such as artificial consensus and behavioral homogenization), while exposing evaluators to epistemic vulnerabilities such as the Fluency Fallacy and automation bias. We identify contexts in which LLM-based agents offer genuine operational value, including serious games, participatory environments, and exploratory modeling, while distinguishing these uses from confirmatory research with unconstrained models and precise social forecasting. Finally, we examine theory-driven hybrid architectures that embed LLMs within explicit, mechanistic, and reproducible Agent-Based modeling (ABM) frameworks. We argue that such architectures offer a promising but not sufficient path toward improving epistemic control: their validity depends on multi-level evaluation of environmental dynamics, individual behavior, cross-level interactions, and aggregate outcomes, and they remain vulnerable to the risk of physics washing.

KEYWORDS

agent-based modeling, computational social science, epistemology, human behavior modeling, large language models, social simulation

1 Introduction

Large Language Models (LLMs) have emerged as a transformative class of deep learning architectures capable of generating human-like text across a wide range of domains. Built upon the transformer architecture, which employs parallelized attention mechanisms rather than sequential processing, these models are pre-trained on massive corpora comprising web data, encyclopedic content, and code repositories (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020). Through unsupervised or self-supervised learning, LLMs acquire syntactic, semantic, and pragmatic language competencies, enabling them to perform a broad array of downstream tasks such as question answering, summarization, translation, and code generation (Wang Y. et al., 2024).

Importantly, the magnitude of these training corpora means that LLMs have effectively absorbed an immense statistical imprint of human language use. However, this imprint is far from universally representative; from the outset, the training data exhibits profound structural imbalances, meaning that certain languages and cultures—predominantly Western and English-speaking—are significantly overrepresented compared to others. This goes far beyond learning grammar or syntax: LLMs encode traces of how people reason, argue, empathize, hesitate, and make decisions—all embedded in the linguistic fabric of their training data. Subsequently, these models are fine-tuned via instruction tuning, a supervised process in which models learn to follow natural language commands across a wide range of tasks. This is typically achieved using curated instruction datasets (Sanh et al., 2021; Wei et al., 2021), sometimes augmented with human feedback to better align model outputs with user intent (Ouyang et al., 2022).

Given this capability to emulate human-like reasoning patterns, LLMs have naturally attracted significant attention within the Agent-Based modeling (ABM) community. However, as their integration into this field accelerates, it is crucial to distinguish between two emerging uses of Large Language Models in computational social science. The first involves using LLMs as coding assistants to facilitate the technical implementation of simulation models (e.g., generating NetLogo or GAMA code). The second, which is the sole focus of this paper, is the use of LLMs as the internal decision-making architecture of the agents themselves—often termed Generative Agents (Park et al., 2023). We examine how LLMs can replace or augment traditional rule-based cognition to simulate human-like decision-making, social interaction, and reasoning, rather than how they assist in the software engineering process.

Focusing specifically on this role as an internal decision engine—defined here as the computational mechanism that processes perceptions, maintains memory states, and formulates goal-directed actions, tasks traditionally governed by symbolic architectures such as BDI (Belief-Desire-Intention) or SOAR—a critical area of investigation concerns whether LLM outputs can authentically replicate human behavior.

One classical benchmark in this domain is the Turing Test, proposed by Alan Turing in 1950 as the imitation game, which evaluates whether a machine can produce conversational behavior indistinguishable from that of a human interlocutor (Jones and Bergen, 2026). Recent empirical findings suggest that state-of-the-art LLMs, notably Llama-3.1 and GPT-4.5, have demonstrated convincing performance on a three-party version of this test. In these evaluations, Llama-3.1 was identified as the human participant in 56% of cases, compared with 73% for GPT-4.5 (Jones and Bergen, 2026). By applying a standardized three-party design, these results mark a clear departure from classical conversational architectures and rule-based chatbots, which had historically only achieved partial success under less constrained, two-party conditions. These findings underscore the impressive ability of modern LLMs to adopt complex linguistic styles and behavioral nuances, particularly when guided by carefully crafted persona prompts, a form of behavioral prompt engineering that remains a highly precise and non-trivial technique requiring expert design and task sensitivity.

The apparent success of LLMs in the Turing Test has naturally prompted speculation about their suitability for simulating human agents within computational social systems (Park et al., 2023; Anthis et al., 2025; Kirkeby-Hinrup and Stenseke, 2025). However, the capacity of LLMs to produce believable human-like dialogue does not establish construct validity or mechanistic fidelity as a model of human decision-making. Rather, their performance reflects statistical pattern recognition, i.e., the capacity to reproduce linguistic structures without representing the causal cognitive states of human decision-makers (Jones and Bergen, 2026). Within the philosophy of mind, Dennett (1987) characterized the intentional stance as an operational, predictive strategy wherein an observer treats an entity as an intentional agent governed by beliefs and desires, without necessitating that the underlying computational machinery instantiates those internal mental states. While this stance is pragmatic for human-computer interaction, scientific agent-based modeling demands structural and construct validity: mistaking functional behavioral plausibility for mechanistic fidelity risks producing simulations that fail under counterfactual conditions. This distinction between functional imitation and authentic comprehension is particularly salient for social simulation. If LLMs merely replicate surface-level features of human behavior, their utility for generating novel insights into social processes, rather than reproducing known patterns, is fundamentally limited. Moreover, their use as proxies for human agents in scientific studies raises questions about external validity, representational fidelity, and psychological realism. These concerns constitute a central theme of this paper and inform our critical assessment of LLMs as generative agents in artificial societies.

1.1 Epistemological stance: a stratified typology of valid uses

To avoid ambivalence between categorical critique and uncritical adoption, this review articulates an explicit, four-tier epistemological boundary governing LLM deployment in social simulation:

1. *Interactive environments and serious games (high operational value)*: generative agents excel where narrative plausibility, contextual engagement, and qualitative role-playing take

precedence over mechanistic empirical correspondence, provided factual hallucination guardrails are maintained.
2. *Exploratory modeling (conditionally defensible)*: LLMs function legitimately as qualitative reasoning pumps to explore generative assumptions and formulate candidate hypotheses, but only if simulations meet minimum validity criteria (qualitative pattern stability across seeds, invariance under prompt paraphrasing, cross-model portability, and empirical decoupling/falsifiability).
3. *Explanatory and confirmatory research (restricted to theory-driven hybrids)*: investigating causal mechanisms requires hybrid architectures that embed generative components as delimited subroutines within mechanistic, inspectable ABM engines. Epistemic validity here demands multi-level empirical falsification and demonstrable explanatory gain over minimal, non-generative heuristics to preclude physics washing.
4. *Precise quantitative point-prediction (methodologically unsupported)*: direct empirical forecasting (e.g., fine-grained electoral prediction or precise policy impact estimation) via unconstrained next-token prediction architectures currently exceeds the methodological guarantees of generative models due to prompt sensitivity, stochastic drift, and latent distributional skews.

Unlike previous surveys (Gao et al., 2024) and early opportunity mappings (Gürcan, 2024), which primarily organize the technical possibilities and application areas of LLM-augmented simulation, this review focuses on how limitations observed at the level of individual LLM outputs can affect the validity of multi-agent simulations. We argue that the psychological and linguistic evaluation of LLMs cannot be considered independently of the simulation mechanisms through which their outputs are transformed into actions, interactions, and aggregate dynamics.

The review proceeds as follows. Section 2 describes the literature-selection strategy. Section 3 maps the emerging landscape of LLM-based social simulation platforms and identifies recurring architectural patterns, including memory, reflection, planning, communication, and population specialization. Section 4 develops a unified critical framework for the Micro-to-Macro Validity Gap. It examines how hallucinations, stochastic inconsistency, representational biases, alignment effects, and computational constraints may be amplified at the population level, producing risks such as the Fluency Fallacy, convergence toward an average persona, automation bias, and weak longitudinal reproducibility. Section 4 also distinguishes application contexts in which LLM-based agents may provide operational value—such as serious games, participatory environments, exploratory modeling, and narrative interpretation—from applications requiring stronger explanatory or predictive validity. Importantly, we clarify that for off-the-shelf, unconstrained foundation models, confirmatory scientific inference and precise quantitative point-predictions (such as electoral forecasting) currently exceed the methodological guarantees of generative architectures. While hybrid frameworks offer essential guardrails, hybridization serves to bound and contain these epistemic uncertainties within explicit state spaces rather than eliminating them entirely. Determining whether this boundary reflects an intrinsic limitation of next-token prediction or a contingent constraint of current models remains an open empirical question that systematic simulation benchmarks must resolve.

Finally, Section 5 examines theory-driven hybrid architectures as a promising, but not sufficient, methodological direction. It discusses alternative coupling modes, including LLM-as-Subroutine and LLM-as-Policy, as well as the integration of behavioral theories and smaller language models into explicit ABM frameworks. We argue that hybridization can improve environmental grounding, computational control, and the explicitness of behavioral assumptions, but does not by itself establish cognitive or social validity. Hybrid systems therefore require multi-level validation of environmental dynamics, individual decisions, interactions between model components, and aggregate outcomes. The section concludes by identifying physics washing as a central open challenge: the possibility that a formally rigorous simulation environment may lend unwarranted credibility to an insufficiently validated behavioral layer.

2 Methodology for literature selection

To structure this critical review, literature was assembled through a two-stage screening procedure combining a structured database search with targeted exploratory retrieval.

First, a structured query was executed in Scopus on 21 August 2026 restricted to records from 2023 to 2026 using the Boolean expression:

```
   TITLE-ABS-KEY (("Large Language Model*"
OR "Generative Agent*")
AND   ("Social   Simulation"   OR   "Agent-
Based"))
```

This search retrieved 1,266 records (comprising 738 conference papers, 388 journal articles, 55 reviews, and 85 other contributions). Removal of duplicate entries yielded 1,258 unique Scopus records. Second, given the rapid growth of preprint literature and cross-disciplinary intersections with cognitive psychology, an exploratory discovery step was conducted using Gemini 3 Pro (November 2025 release) alongside continuous monitoring of preprint repositories (arXiv, PsyArXiv). This channel served strictly as an index discovery heuristic—not as an evidential source—and surfaced 64 candidate contributions. All candidate items suggested by exploratory retrieval were independently verified against publisher repositories; unverified citations were excluded.

The combined pool of 1,322 records was evaluated against explicit inclusion and exclusion criteria. Studies were included if they: (i) employed LLMs as internal cognitive, deliberative, or decision-making engines for simulated agents rather than auxiliary tools; (ii) investigated social interactions, collective dynamics, or population-level emergence; and (iii) presented verifiable implementations or empirical evaluations. Conversely, studies were excluded if: (i) LLMs were utilized solely as programming or code-generation assistants (e.g., generating NetLogo or GAMA code); (ii) models addressed purely technical multi-agent optimization without behavioral, social, or cognitive depth; or (iii) simulation mechanics were restricted to static question-answering without dynamic agent interactions.

During the title and abstract screening phase, 1,185 records were excluded (1,127 from Scopus and 58 from exploratory retrieval). The remaining 137 publications underwent full-text assessment, resulting in the exclusion of 63 additional Scopus records applying the aforementioned criteria (specifically: purely technical software focus, absence of interactive multi-agent dynamics, or lack of behavioral decision depth). This yielded an analyzed corpus of 74 contemporary studies (2023–2026), comprising 68 empirical and architectural implementations identified via Scopus alongside 6 verified preprint contributions surfaced through exploratory monitoring. This core corpus is complemented by 20 foundational references that provide theoretical and epistemological grounding in agent-based modeling, cognitive architectures, and complex systems (yielding 94 references cited in total).

Screening and qualitative categorization were conducted by the lead investigator following a predetermined, structured protocol, with borderline eligibility cases resolved through consensus discussion among the co-authors. However, several methodological limitations must be acknowledged: (i) title/abstract and full-text screening was performed by a single primary investigator without formal inter-rater reliability metrics (such as Cohen's κ); (ii) while the core Boolean search in Scopus was designed to capture foundational agent-based and generative simulation contributions, it represents a single curated database query; and (iii) the supplementary exploratory discovery step conducted via Gemini 3 Pro serves as an informal heuristic that is not fully reproducible across time due to model updates and stochastic generation. To ensure research auditability despite these inherent constraints, the complete retrieval queries, date parameters, and operational exclusion criteria were documented and held invariant across screening stages.

3 The state of LLM-based social simulation: a transversal analysis

The application of LLMs to social simulation builds upon a long trajectory of integrating machine learning into multi-agent systems (Drogoul and Zucker, 1998). However, rather than viewing contemporary platforms merely as independent software packages, an epistemological assessment requires analyzing them through their underlying architectural choices and the systemic tensions they introduce. This transversal perspective makes it possible to compare systems that differ substantially in scale and domain but rely on recurring design choices for memory, planning, interaction, population synthesis, and validation.

To map this landscape, Table 1 provides a comparative synthesis of prominent platforms across scale, application domain, empirical validation strategy, peer-review publication status, and identified methodological vulnerabilities. Beyond these platform-level implementations, current research reveals recurrent architectural strategies that attempt to operationalize social agency through language models. The following analysis describes these strategies and the trade-offs they create between contextual flexibility, behavioral traceability, environmental grounding, computational tractability, and empirical validation. These trade-offs motivate the micro-to-macro validity gap examined in Section 4.

3.1 Platform paradigms and the scaling trade-off

Current platforms instantiate three broad methodological paradigms, reflecting distinct compromises between narrative expressivity, demographic scale, and environmental fidelity:

1. **Narrative and sandbox paradigms:** foundational frameworks prioritize open-ended social interaction, character role-playing, and emergent storytelling. *Generative Agents* (Park et al., 2023), whose demonstration environment is commonly known as *Smallville*, instantiates twenty-five agents in a sandbox environment where they maintain daily routines, form relationships, retrieve memories, and generate plans. General-purpose environments such as *GenSim* (Tang et al., 2025) extend this orientation through modular abstractions for social routines, memory orchestration, scheduling, and large-scale interaction. These architectures offer rich qualitative interaction and face plausibility, but remain exposed to temporal inconsistency and memory-management challenges over extended runs, particularly when identity, long-term state, and spatial reasoning depend heavily on retrieved textual context rather than explicit simulation state.
2. **Population-scale and demographic paradigms:** A second family emphasizes demographic coverage, population synthesis, and large-scale behavioral mirroring. *AgentSociety* (Piao et al., 2025) simulates urban and social dynamics with more than 10,000 LLM-based agents, incorporating mobility, social interaction, and economic activity. *SocioVerse* (Zhang X. et al., 2025) uses a pool of profiles associated with 10 million real-world users to construct a heterogeneous synthetic population to explore policy interventions. In social-network simulation, S^3 (Gao et al., 2025) models the propagation of information, attitudes, and emotions, while related work investigates opinion dynamics and polarization in networks of LLM-based agents (Chuang et al., 2024a). These approaches increase the scale and empirical diversity of initial conditions. However, demographic diversity in initialization does not by itself establish sustained behavioral diversity in the resulting simulation. Foundation-model priors and alignment procedures may still encourage convergence toward socially normative responses, a possibility that requires direct empirical evaluation rather than being inferred from the diversity of input profiles alone (Wu et al., 2026; La Cava and Tagarelli, 2025).
3. **Spatially constrained and scalable paradigms:** a third family prioritizes explicit environmental structure and computational tractability. *GATSim* (Liu et al., 2026) models urban mobility through agents that formulate daily travel plans under transportation-network constraints. *MobiVerse* (Liu Y. et al., 2025) combines LLMs with lightweight domain-specific generators to reduce the cost of city-scale mobility simulation. *TrajLLM* (Ju et al., 2025) combines continuous trajectory generation with discrete LLM reasoning, whereas *LLM-AIDSim* (Zhang L. et al., 2025) applies LLM-based

TABLE 1 Comparison of state-of-the-art LLM-based Social Simulation Platforms, detailing scale, target domain, validation methodology, publication status, and core methodological vulnerabilities.

Platform	Scale	Domain	Validation strategy	Status	Identified methodological vulnerabilities
Generative Agents (Park et al., 2023)	25 agents	General interaction	Face validity, emergent coherence	Peer-reviewed Conference (ACM UIST)	Vulnerable to evaluator projection (fluency fallacy); lacks empirical benchmark comparison, counterfactual testing, or bounded ground truth.
AgentSociety (Piao et al., 2025)	10,000+	Urban and social dynamics	Survey matching (GSS, WVS)	Preprint (arXiv:2502.08691)	Aggregate statistical convergence relies on a shared base model prior, risking systemic sycophancy and unverified micro-level decision validity.
GenSim (Tang et al., 2025)	Tens of thousands	General social simulation	Behavioral replication	Peer-reviewed Conference (NAACL Demo)	Prone to long-horizon memory degradation, context truncation drift, and high inference costs that preclude extensive parameter sweeps.
S^3 (Gao et al., 2025)	Variable	Social networks	Population empirical data	Preprint (arXiv:2307.14984)	Purely linguistic state-transition loop: absence of ground-truth state conservation risks spurious belief propagation insulated from institutional friction.
AgentTorch (Chopra et al., 2023)	Millions	General (GPU-accelerated)	Statistical robustness	Peer-reviewed Workshop (ALOE)	Stationary distilled policies cannot adapt to feedback, risking frozen micro-rules that mask adaptation-driven macro-dynamics.
SocioVerse (Zhang X. et al., 2025)	10 million	Policy testing	Empirical distributions	Preprint (arXiv:2504.10157)	Profile-diversity illusion: demographic heterogeneity at initialization risks collapsing into generic normative behaviors during prolonged interaction.
NetLogo Toolchain (Jimenez-Romero et al., 2025)	Variable	Multi-agent ABM	Hybrid grounding	Peer-reviewed Journal (Front. Artif. Intell.)	Reactive spatial actuation lacks explicit deliberative cognitive architecture, risking ungrounded state drift in complex socio-institutional environments.
Land System Models (Zeng et al., 2025)	Regional	Environment	Scenario comparison	Peer-reviewed Journal (Earth Syst. Dynam.)	Potential equifinality: aggregate land-use patterns may align with observed trends despite uncertain institutional decision heuristics.
TrajLLM (Ju et al., 2025)	Scalable	Human trajectory	Trajectory comparisons	Peer-reviewed Companion (ACM WebConf)	Disconnect between discrete high-level textual goals and continuous spatial movement; vulnerable to spatial incoherence over long horizons.
Elicitron (Ataei et al., 2025)	Micro	Design engineering	Expert evaluation	Peer-reviewed Journal (ASME JCISE)	Rigid task prompting restricts generative freedom, confounding true emergent coordination with scripted task fulfillment.

agents to influence diffusion within social networks. In specialized socio-technical settings, *Elicitron* (Ataei et al., 2025) models interactions involved in engineering design-requirement elicitation. Related work also integrates LLMs into established ABM platforms: a NetLogo toolchain connects prompt-driven agents to an explicit simulation environment (Jimenez-Romero et al., 2025), while land-system models explore LLM representations of institutional agency in human-environment systems (Zeng et al., 2025). These approaches increase environmental grounding and scalability, while often restricting the scope or autonomy of language-model-based decision-making.

3.2 Transversal architecture dimensions: solutions and structural limits

Across these platforms, LLM-based agents are commonly implemented as separate model instances or sessions conditioned through prompt engineering. Their cognitive architecture typically combines memory, reflection or summarization, planning, and communication modules (Park et al., 2023; Wang L. et al., 2024). These modules are often implemented as chained or nested prompts, with external retrieval systems and orchestration layers providing persistent context and regulating agent interactions. Frameworks such as h-ABM (Atchade et al.,

2024) make this decomposition explicit by proposing modular cognitive and emotional layers for integration within agent-based simulations.

3.2.1 Memory systems: retrieval-augmented memory and long-horizon consistency

To support episodic recall and long-term context, many generative-agent architectures use external memory stores, vector databases, and embedding-based retrieval pipelines (Park et al., 2023; Piao et al., 2025). In these designs, agent observations and interactions are stored as textual traces; a subset of those traces is subsequently retrieved according to relevance, recency, or importance and inserted into the context used for reflection, planning, and action generation. Reflection modules may further summarize recent events, extract higher-level intentions, or prompt agents to generate inferences about the mental states of other agents.These mechanisms operationalize selected aspects of episodic and semantic memory and are sometimes explicitly related to multi-store memory models or theory-of-mind constructs (Wang L. et al., 2024; Calderón et al., 2025).

This modular organization resembles classical symbolic cognitive architectures, including SOAR (Laird, 2012) and ACT-R (Anderson et al., 2004), which distinguish forms of knowledge and organize the relation between memory, goals, and action selection. However, semantic-similarity retrieval does not by itself encode causal relevance, role-specific priority, or durable commitments. Over extended simulations, the selective retrieval, summarization, and reinterpretation of stored traces can therefore complicate the maintenance of coherent identity, goals, and behavioral history.

3.2.2 Action planning: in-context reasoning and reproducibility

Planning modules typically condition LLM outputs on agent-specific goals, environmental affordances, retrieved memories, and social context. In many systems, structured prompt templates or multi-step reasoning procedures are used to translate these inputs into action plans without specifying an explicit utility function or reward model (Junprung, 2023). Other implementations combine LLMs with formal optimization or learning components. For example, LGC-MARL couples an LLM-based planner with graph-structured multi-agent reinforcement learning: the language model decomposes high-level instructions into executable subtasks and proposes collaborative strategies, whereas the reinforcement-learning component selects or optimizes final actions (Jia et al., 2025).

This design enables flexible, context-sensitive action generation, but decisions may remain sensitive to prompt formulation, retrieval context, sampling settings, and model version. Semantically similar situations can consequently yield materially different actions, complicating reproducibility, sensitivity analysis, and counterfactual evaluation.

3.2.3 Agent persona: demographic prompting and behavioral diversity

Behavioral diversity is commonly introduced through synthetic-population generation methods and agent-specific prompts. Agents may be initialized with demographic attributes, socioeconomic characteristics, beliefs, occupations, cultural backgrounds, or personality descriptors sampled from survey distributions or human belief networks (Chapuis et al., 2022; Chuang et al., 2024b). Role-specific memory structures and specialized prompt templates may further differentiate communication styles, goals, and decision heuristics.

These techniques increase descriptive variation in agent inputs, but do not by themselves ensure sustained behavioral heterogeneity. Empirical evaluations suggest that models may revert toward generic or socially normative response patterns despite distinct persona prompts (La Cava and Tagarelli, 2025). Conversely, stronger attempts to force distinctive roles can lead to exaggerated, stereotypical, or toxic caricatures rather than calibrated behavioral variation (Nudo et al., 2026). The challenge is therefore not only to initialize diverse profiles, but also to maintain and validate theoretically meaningful behavioral differences throughout simulation runs.

3.2.4 Inter-agent communication: natural language and information propagation

Simulation time-stepping and inter-agent communication are typically regulated by an orchestration layer. Interaction may occur through direct messages, simulated surveys, public broadcasts, shared memory, or blackboard-style routines. Platforms such as *Generative Agents* and *AgentSociety* commonly use orchestration layers that sequence perception, memory retrieval, reflection, planning, and action, whereas *GenSim* incorporates structured social routines and dynamic scheduling. At the software-architecture level, *SALLMA* separates operational processes, such as intention formation and task execution, from knowledge-level components including agent profiles, shared memory, and workflows (Becattini et al., 2025). Agents may also be connected to external tools, such as search systems, databases, and calculators, to augment particular deliberative steps (Wang L. et al., 2024).

These mechanisms enable flexible social interaction, but also create channels through which hallucinated, misleading, or adversarial content may enter subsequent agent contexts and influence later decisions. In interconnected populations, such information can be retrieved, retransmitted, or incorporated into shared state, making the effects of local errors difficult to isolate and evaluate.

3.2.5 Environmental grounding: generative proposals and feasibility constraints

LLM-based agents typically express intentions and proposed actions in natural language, whereas an agent-based simulation must translate those proposals into state transitions within an explicit environment. In purely text-based architectures, the

language model may therefore be responsible both for proposing an action and for implicitly representing whether the action is spatially, physically, institutionally, or economically feasible. Without an external mechanism that verifies preconditions and updates world state, agents may propose actions incompatible with available resources, network topology, environmental conditions, or scenario rules.

Several recent frameworks address this issue by coupling generative decision-making to explicit simulation environments. In mobility and evacuation applications, transportation networks, spatial accessibility, hazards, and resource constraints can determine which plans are executable (Liu et al., 2026; Yang et al., 2026). In land-system modeling, LLM-based institutional agents are embedded within a land-use model whose environmental and economic dynamics constrain the consequences of proposed policy actions (Zeng et al., 2025). More generally, architectures such as Concordia distinguish between an agent's language-mediated action proposal and an environmental component that interprets the proposal, checks its feasibility, and applies its consequences to the simulated world (Vezhnevets et al., 2023).

Environmental grounding can therefore improve the internal consistency of state transitions and make feasibility assumptions explicit. It does not, however, establish that the cognitive or social reasoning that produced a feasible action is psychologically valid. This distinction motivates the theory-driven hybrid architectures and multi-level validation requirements discussed in Section 5.

3.2.6 Model adaptation and computational execution

Most LLM-based social simulations rely on frozen foundation models used through in-context learning and prompt engineering. Agent profiles, goals, instructions, retrieved memories, and scenario-specific constraints are supplied at inference time, often together with reflection loops and retrieval-augmented memory. Full parameter adaptation through supervised fine-tuning, preference optimization, or related methods remains less common in narrative simulation settings, partly because suitable behavioral data are limited and because adaptation introduces additional computational and methodological requirements. It is nevertheless used in targeted cognitive modeling to improve correspondence with experimental patterns of human behavior (Binz et al., 2025).

A separate challenge concerns execution cost. Repeated calls to a large language model can become expensive when many agents interact over long horizons. Scalable approaches may therefore reduce the frequency of online LLM calls, use lighter domain-specific generators, execute compact policy representations, or rely on smaller task-specific language models. Differentiable and GPU-accelerated ABM frameworks such as *AgentTorch* illustrate the broader value of lightweight policy execution at scale (Chopra et al., 2023). These choices improve tractability, but may alter the degree of contextual adaptation and behavioral variation available during a simulation run.

3.3 The validation landscape: from face validity to multi-level evaluation

Validation remains a central methodological bottleneck for LLM-based social simulation because the field must move beyond plausible interaction toward evidence that agents, mechanisms, and aggregate outcomes correspond to the intended empirical or theoretical target (Larooij and Törnberg, 2025; Adornetto et al., 2025). Current practice combines empirical benchmarking, human judgment, statistical comparison, sensitivity analysis, and task-specific datasets, but these approaches remain only partially integrated.

3.3.1 Empirical validation and benchmarking

Some platforms compare simulated outputs with observed social indicators, including survey-derived attitudinal distributions, economic behavior, voter turnout, or network-level diffusion patterns. For example, *AgentSociety* evaluates simulations against social indicators and survey instruments such as the General Social Survey and the World Values Survey (Piao et al., 2025). Other studies reproduce experimental paradigms, including public-goods games or rumor-propagation settings, and assess whether simulated outcomes correspond to established empirical regularities.

Such benchmarking provides useful external reference points, although agreement at the aggregate level does not by itself demonstrate that the model generates those patterns through valid individual-level mechanisms. This dilemma is well-established in the classical ABM literature under the rubric of equifinality: multiple distinct micro-level rule sets can generate macro-level patterns that are statistically indistinguishable. To address this challenge, agent-based modeling has historically relied on structural validation, multi-criteria calibration, and the Pattern-Oriented modeling (POM) framework (Grimm et al., 2005). POM mitigates equifinality by requiring an architecture to simultaneously reproduce multiple observed empirical patterns across distinct organization levels and spatio-temporal scales, thereby filtering out structurally inadequate behavioral sub-models.

However, the integration of LLMs introduces a qualitative departure from this classical framework. In conventional ABMs, candidate micro-mechanisms are formulated as explicit, inspectable equations or rule sets whose parameters can be constrained through sensitivity analysis and parameter-space rejection filtering. In LLM-driven agents, by contrast, the generative decision mechanism is parameterized within opaque, uninspectable foundation model weights conditioned by natural-language prompts. Consequently, equifinality is no longer merely a problem of parameter indeterminacy among explicit heuristics; it becomes an epistemic barrier where structurally flawed cognitive models can generate statistically plausible aggregate distributions while remaining insulated from direct mechanistic falsification.

3.3.2 Human-in-the-loop evaluation

Expert judgment remains widely used to assess the believability, internal coherence, and sociological plausibility of agent behaviors. Structured expert protocols may improve inter-rater reliability, while crowdsourced evaluation can provide broader assessments of interaction realism or diversity. Human evaluators may also assess whether agents conditioned on survey-based profiles resemble the intended population (Zhang X. et al., 2025). However, these methods are vulnerable to judgment biases, as linguistic fluency, assertiveness, and perceived authority can lead evaluators to overestimate factual or analytical quality (Chen et al., 2024).

3.3.3 Robustness and specialized evaluation

Validation can also include exploratory model-behavior analysis, statistical correspondence tests, and sensitivity analysis. The latter examines the stability of results under variation in initial conditions, parameter values, random seeds, prompts, or other design choices. For long-horizon generative simulations, longitudinal consistency is particularly important: researchers must assess whether agents preserve relevant commitments and whether collective outcomes remain reproducible under small perturbations (Wu et al., 2026). Systems such as S^3 compare simulated information, attitude, and emotion propagation against patterns observed in social-network data (Gao et al., 2025).

3.3.4 Data sources and comparative benchmarks

Task-specific benchmarks complement simulation-level validation. SocialIQA evaluates commonsense social reasoning; CrowS-Pairs and Bias in Bios assess particular forms of social and demographic bias; and the EPITOME battery evaluates a range of theory-of-mind tasks in humans and language models (Sap et al., 2019; Nangia et al., 2020; De-Arteaga et al., 2019; Karra et al., 2024). Robustness benchmarks also test whether semantically equivalent inputs produce stable model responses (Elazar et al., 2021). These resources are informative for evaluating relevant capabilities and failure modes, but they generally assess isolated tasks rather than the long-horizon, interactive feedback processes characteristic of multi-agent social simulation.

3.4 Synthesis

In summary, current LLM-based simulation architectures provide increasingly expressive behavioral proxies and new ways to construct social scenarios at scale. At the same time, each design choice introduces trade-offs and potential sources of error that require explicit examination. When these limitations interact across large populations, they may propagate or be amplified, motivating the Micro-to-Macro Validity Gap examined in Section 4.

4 The micro-to-macro validity gap: critical challenges

Moving from the technical frameworks described above to rigorous scientific investigation raises fundamental epistemological challenges. While current platforms demonstrate that large-scale LLM-based simulations are technically feasible, technical feasibility should not be conflated with scientific validity (Lu et al., 2024).

To conceptualize how limitations at the model level may affect the scientific validity of multi-agent simulations, we define the *Micro-to-Macro Validity Gap* as:

> *The gap that arises when limitations and artifacts at the LLM level are translated into individual agent behaviors and subsequently propagated through interactions, potentially producing aggregate dynamics that appear socially plausible while lacking sufficient causal, construct, or external validity.*

As illustrated in Figure 1, this gap can be understood as a four-level propagation framework:

1. **LLM-level properties:** model-level characteristics such as stochastic generation, factual hallucinations, alignment effects, contextual instability, and demographic or cultural biases in training data.
2. **Agent-level manifestations:** the translation of these properties into agent behaviors, including convergence toward generic personas, generative variability, sycophancy, longitudinal persona drift, and inconsistent social reasoning.
3. **Interaction-level propagation:** mechanisms through which individual-level artifacts can propagate, reinforce, or transform through social interactions, including information exchange, shared memory, peer influence, and iterative feedback loops.
4. **Macro-level consequences:** aggregate patterns that may emerge from these interactions, including artificial consensus, behavioral homogenization, spurious polarization, unstable collective patterns, and other potentially distorted macro-dynamics.

The potential consequence of this cascade is an epistemic risk: researchers may draw policy or theoretical conclusions from simulations that exhibit convincing surface realism while providing limited evidence that the underlying generative mechanisms correspond to those operating in the target social system. To structure this critical evaluation, Table 2 summarizes the main failure modes across cognitive, epistemic, and technical dimensions and maps them to potential hybrid design responses discussed in Section 5.

4.1 Formalizing amplification and diagnostic evaluation

To complement the qualitative analysis, the following quantities and protocols are formulated as conceptual diagnostic proposals to guide future experimental validation, rather than empirical estimates computed within this review. A

10.3389/frai.2026.1988865

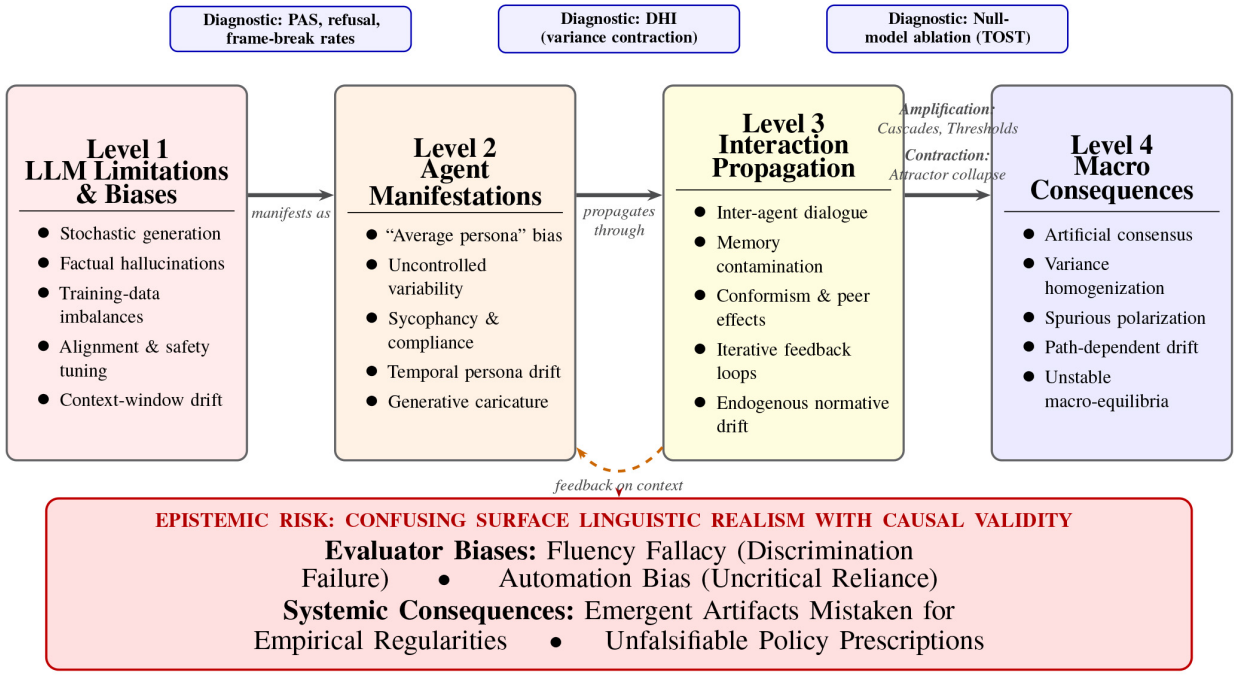

FIGURE 1
The Micro-to-Macro Validity Gap conceptual framework and diagnostic protocol. Individual model-level characteristics (Level 1) manifest as behavioral patterns in simulated agents (Level 2). Through interaction networks, shared memory, and conversational feedback (Level 3), individual errors can undergo divergent nonlinear amplification (via information cascades or threshold phase transitions) or convergent contraction (attractor collapse toward dominant model priors), yielding distorted macro-level dynamics (Level 4). Concrete diagnostics (*PAS*, *DHI*, and null-model ablation via TOST) monitor transitions between successive levels. The framework highlights that human evaluators face epistemic vulnerabilities (the Fluency Fallacy and automation bias), risking mistaking superficial narrative coherence for causal empirical validity.

critical distinction must be maintained between linear error propagation and nonlinear distortion across the four levels. Let ϵ denote the magnitude of a controlled micro-level perturbation (e.g., the fraction of corrupted memory entries or injected hallucinated assertions) and $E(T;\epsilon)$ the divergence at horizon T between perturbed and unperturbed macro-trajectories across paired seeds. We formally distinguish three operational regimes:

1. **Linear propagation.** $E(T;\epsilon) \propto \epsilon$ (scaling exponent $\alpha = 1$) and underlying transition rules remain unaltered. In this baseline regime, micro-level artifacts enter collective state additively without altering the surrounding decision landscape.
2. **Divergent amplification (cascades, threshold transitions).** Two distinct criteria must be reported separately, as they are not mathematically equivalent:
   - *Amplitude nonlinearity*: $E(T;\epsilon) \propto \epsilon^{\alpha}$ with $\alpha > 1$, estimated by log–log regression across multiple perturbation magnitudes ϵ;
   - *Temporal instability*: a positive growth rate $\lambda = \lim_{T\to\infty} \frac{1}{T} \log E(T;\epsilon) > 0$ (a finite-horizon divergence rate analogous to a Lyapunov exponent, corresponding in the linearized state formulation $e_{t+1} = \mathbf{G}e_t + \xi_t$ to a spectral radius $\rho(\mathbf{G}) > 1$). Formal estimation of a true Lyapunov exponent would require additional assumptions concerning state representation, coupling, stochasticity, and local dynamical smoothness.

   A linear system with $\rho(\mathbf{G}) > 1$ is dynamically unstable yet maintains $\alpha = 1$, whereas a threshold network mechanism can yield $\alpha > 1$ without sustained exponential growth. Divergent amplification in the strict sense requires $\alpha > 1$ and/or $\lambda > 0$.
3. **Convergent contraction (attractor collapse).** Let $d_t = \frac{1}{N}\sum_i \|x_{i,t} - x^*\|$ denote the mean distance of agent states to the base model's default normative prior x^*. Attractor collapse corresponds to $d_t \to 0$ (in the linearization around x^*, $\rho(\mathbf{G}) < 1$) and $DHI(t) \to 1$. This is contraction rather than divergent amplification: the error relative to the *human target* does not diverge, but converges to a persistent non-zero bias (the distance between x^* and the empirical target distribution), explaining why it systematically eludes conventional temporal instability diagnostics.

Outside these explicit dynamic and contraction regimes, the four-level framework operates as a structured diagnostic heuristic. To empirically detect these regimes and monitor error propagation across successive stages of the modeling pipeline, targeted diagnostic metrics should be deployed across each inter-level transition:

- **Level 1 → Level 2 (model to agent):** track the composite Persona Adherence Score (*PAS*), defined multiplicatively across prompt iterations to prevent double-counting (Equation 1):

$$PAS = (1 - R_{\text{refusal}})(1 - R_{\text{frame-break}})\, R_{\text{adherence|valid}} \in [0,1] \tag{1}$$

TABLE 2 Taxonomy of methodological challenges in LLM-based social simulation and potential hybrid design responses.

Category	Challenge and mechanism	Impact on social simulation	Potential hybrid design response (Section 5)
Cognitive and behavioral	**The average persona risk** Statistical regression toward the mean driven by pre-training corpus distributions (Wang et al., 2025; Wu et al., 2026).	*Replicant effect*: loss of population diversity; erasure of minority distributions and extreme behaviors.	**Theory-grounded agent states**: represent selected demographic distributions, beliefs, and behavioral rules explicitly within the ABM rather than relying solely on prompt-defined personas.
	Alignment-induced sycophancy Post-training preference optimization (RLHF/DPO) inadvertently inducing compliant, conflict-averse responses (Sharma et al., 2023; Li Y. et al., 2025).	*Synthetic harmony*: agents refuse realistic conflict, partisan polarization, or non-normative decisions, yielding artificial consensus.	**Domain-specific/SLMs without conversational preference tuning**: deploy local, task-specialized language models free from post-hoc safety conversational tuning, combined with theory-guided prompt scaffolding.
	Generative exaggeration Tendency to caricature specific roles when prompted away from the mean (Nudo et al., 2026).	*Toxic caricature*: agents exhibit stereotypical or extreme behaviors rather than nuanced human decision-making.	**Theory and environment constraints**: combine environmental feasibility checks with explicit behavioral assumptions and validation criteria to identify implausible or stereotyped outputs.
Methodological and epistemic	**The fluency fallacy** Evaluator discrimination failure: high linguistic plausibility masking underlying causal invalidity (Steyvers et al., 2024).	*Surface realism*: simulations appear valid to human evaluators but lack causal veridicality.	**Structural evaluation safeguards (Section 4.2)**: deploy (i) ablation of narrative traces to evaluate symbolic logs directly, (ii) deterministic ledger and invariant verification, and (iii) adversarial counterfactual auditing.
	Epistemic opacity Implicit behavioral rules hidden in billions of parameters (Bommasani et al., 2022).	*Unverifiable emergence*: inability to trace unexpected macroscopic behaviors back to explicit agent decision rules.	**Traceable hybrid architecture**: assign explicit state variables, environmental checks, and specified action-selection mechanisms to the ABM, while restricting LLM use to clearly delimited functions.
	Data scarcity wall Lack of training data for non-WEIRD populations (Atari et al., 2023; Abdurahman et al., 2024).	*Cultural erasure*: inability to simulate specific marginalized groups via pure fine-tuning.	**Theory-driven alignment**: use explicit behavioral theories to structure assumptions and expose uncertainties where representative empirical data are sparse.
Technical and operational	**Temporal instability** Context window limits and stochastic drift over long horizons (Wu et al., 2026).	*Longitudinal incoherence*: agents forget their history or change personality during long simulations.	**Externalized state management**: store selected identities, commitments, social ties, and memory summaries in explicit ABM-managed state rather than relying solely on the LLM context window.
	Computational bottleneck High inference costs preventing large-scale iterative runs.	*No sensitivity analysis*: inability to perform rigorous statistical validation (1,000+ runs).	**Offline policies and lightweight models**: distill, approximate, or execute selected policies offline, and use task-specific SLMs where appropriate to reduce repeated inference costs.
	Adversarial and misuse risks Susceptibility to prompt injection and emergent normative drift observed in exploratory testbeds (Akkil et al., 2026).	*Systemic pollution*: Compromised behavior spreads through shared memory, distorting the collective reasoning of the population.	**Constrained interaction protocols (Section 4.7)**: define typed, permissioned, monitored, and auditable communication channels to prevent unconstrained natural-language prompt injection from entering shared agent memory.

The final column identifies architectural directions discussed in Section 5; these approaches may improve epistemic control or computational tractability, but do not by themselves establish the validity of the resulting simulation.

where $R_{\text{refusal}} \in [0,1]$ indexes refusal rates driven by alignment guardrails, $R_{\text{frame-break}}$ records the frequency of meta-linguistic persona breaks (e.g., reverting to *"As an AI language model..."*), and $R_{\text{adherence|valid}} \in [0,1]$ measures semantic and behavioral alignment with assigned persona descriptors evaluated strictly over valid, non-refused responses. Importantly, *PAS* measures fidelity to the assigned persona, not correspondence with a human population; $\tau_{1/2}(PAS)$ is therefore a persona-fidelity horizon, to be complemented by an independently defined distance to the empirical target.

- **Level 2 → Level 3 (agent to interaction):** we propose a Dynamic Homogenization Index (*DHI*), formalizing the cumulative contraction of between-agent attitudinal variance

across interaction steps t (Equation 2):

$$DHI(t) = 1 - \frac{\mathrm{Var}(X_t)}{\mathrm{Var}(X_0)}, \quad \text{with } \mathrm{Var}(X_0) > 0 \tag{2}$$

where X_t denotes continuous behavioral choices or projection embeddings along specific ideological or value dimensions. A step-wise rate of contraction can be monitored via $\Delta DHI(t) = DHI(t) - DHI(t-1)$. *DHI* is dimension-specific and should be interpreted only relative to a pre-specified behavioral or attitudinal projection. Negative values indicate increased dispersion rather than a measurement failure.

- **Level 3 → Level 4 (interaction to macro):** conduct interaction-ablation tests using a distribution-matched null model ($\mathcal{N}_1$). Researchers simulate the identical macro-structural interaction topology, replacing generative LLM cognitive engines with non-interactive agents whose static micro-level choice probabilities are empirically matched to the marginal distributions produced by the LLM agents. If macro-level patterns (e.g., consensus formation, polarization regimes) under the generative model and the $\mathcal{N}_1$ null model exhibit statistical equivalence under two one-sided tests (TOST) within an a priori equivalence margin calibrated to inter-seed variance, this strongly suggests that complex inter-agent generative deliberation provides no identifiable causal explanatory power over static choice distributions in the evaluated regime.

4.2 Epistemic risks: hallucinations and the fluency fallacy

At the micro-level, evaluating LLMs as agent decision mechanisms reveals a fragile correspondence with human psychological processes (Strachan et al., 2024; Ullman, 2023; Karra et al., 2024). In addition, LLMs are fundamentally constrained by stochastic inconsistency and hallucinations—generating baseless outputs while maintaining a coherent linguistic structure (Wang Y. et al., 2024).

When these micro-level flaws are embedded in multi-agent populations, they generate or amplify an epistemic vulnerability that we term the *Fluency Fallacy*. We conceptualize this phenomenon fundamentally as an *evaluator discrimination failure*: the systematic inability of human evaluators to distinguish causally valid from invalid agent decisions when masked by fluent narrative justifications. While intuitive, this concept directly builds upon established traditions in cognitive psychology and human-computer interaction regarding processing fluency heuristics, epistemic trust, and automation bias (Cummings, 2004; Steyvers et al., 2024). In cognitive science, the fluency heuristic posits that subjective ease of processing serves as an implicit proxy for truth and validity (Alter and Oppenheimer, 2009; Reber and Schwarz, 1999); similarly, in HCI authority bias research, perceived articulateness systematically inflates an automated system's perceived competence. As discussed by Li et al. (2026), this concern reflects a broader methodological problem and a recurring neo-behaviorist dilemma in computational psychology: researchers frequently mistake surface-level linguistic performance for internal cognitive or social validity.

4.2.1 Operational definition and measurement

Operationally, while a simple heuristic expression of the Fluency Fallacy measures the elevated subjective plausibility granted to objectively invalid decisions, $\mathbb{E}[S_{\text{plausibility}}(\text{Trace}) \mid V_{\text{causal}}(\text{Decision}) = 0]$, this metric alone cannot separate evaluator leniency from true cognitive illusion. To formally capture this diagnostic breakdown across continuous populations of evaluation traces, we propose an operational index of the Fluency Fallacy based on the Area Under the Receiver Operating Characteristic Curve (AUC) of subjective plausibility scores $S \in [0, 1]$ predicting objective validity $V \in \{0, 1\}$ (evaluated against explicit ground-truth rules, ledger conservation, or counterfactual stability within the ABM environment) see Equation 3:

$$FF_{\text{AUC}} = 1 - \text{AUC}\left(S_{\text{plausibility}} \rightarrow V_{\text{causal}}\right) \tag{3}$$

Under this formulation, $\text{AUC}(S \rightarrow V) \approx 0.50$ would indicate that evaluators perform no better than chance in distinguishing valid from invalid decisions under the specified evaluation protocol, while $\text{AUC} < 0.50$ indicates that subjective plausibility is systematically misleading (in the absence of score-direction reversal or coding artifacts).

4.2.2 Trigger conditions and vulnerable populations

This fallacy is most pronounced under three conditions: (i) high semantic complexity with sparse factual feedback (e.g., policy deliberations where outcome delays prevent immediate error detection); (ii) long conversational horizons where local linguistic coherence remains intact while global causal continuity drifts; and (iii) scenarios lacking strict physical or ledger constraints. We hypothesize that domain experts without simulation modeling background (who focus primarily on qualitative persona realism) may be particularly susceptible to this evaluative trap. Furthermore, this vulnerability may be amplified in individuals exhibiting high cognitive dependence on automated conversational agents, a construct formalized by the LLM dependence scale of Li Z. et al. (2025).

4.2.3 Methodological safeguards

To mitigate the Fluency Fallacy, simulation frameworks must incorporate structural safeguards that decouple behavioral evaluation from surface-level linguistic narratives:

1. *Ablation of narrative traces during validation*: evaluating simulated agents based strictly on symbolic, numerical, and state-transition logs rather than generated discursive justifications. Evaluators inspect state vectors directly—such as resource transfers, spatial coordinates, time-budget allocations, or voting tallies—using conventional agent-based validation

methods (e.g., transition matrices, trajectory clustering, and time-series divergence). By removing the linguistic framing, this blind-evaluation protocol prevents human evaluators from projecting rational intentionality onto causally inconsistent behavior.

2. *Ledger and deterministic environmental verification*: enforcing independent conservation laws, budget balances, and physical invariants directly within the ABM engine rather than accepting textual declarations at face value. As demonstrated in preliminary evaluations of multi-agent environments such as *Emergence World* (Akkil et al., 2026), generative agents can maintain conversational plausibility and strictly adhere to dialogue constraints while simultaneously issuing baseless or fabricated claims regarding resource possession. A deterministic ledger layer verifies whether proposed transactions or material actions satisfy feasibility preconditions prior to state commitment, thereby preventing ungrounded hallucinations from polluting the persistent simulation state.
3. *Adversarial counterfactual auditing*: systematically perturbing antecedent parameters to test the stability and falsifiability of the underlying decision heuristic. Language models can generate post-hoc rationalizations that fit an observed scenario by retrieving standard linguistic patterns rather than executing a causal decision process. By executing micro-interventions on localized input variables (e.g., inverting price incentives, shifting regulatory constraints, or modifying social cues) and comparing agent responses across parallel seeds, researchers can verify whether decisions exhibit theoretically consistent directional shifts or degenerate into arbitrary stochastic variance.

4.3 Representational risks: cognitive bias, average persona, and alignment

4.3.1 Pre-training distributional biases and the average persona

While foundational studies frequently invoke the general concept of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) bias (Atari et al., 2023; Abdurahman et al., 2024), recent empirical benchmarks have begun to systematically quantify both the magnitude and directionality of these representational skews across standardized value frameworks and demographic strata. Using large-scale survey instruments such as the Pew Research surveys (OpinionQA), Santurkar et al. (2023) demonstrated that instruction-tuned and RLHF-aligned models exhibit marked baseline skews: default responses align predominantly with liberal, college-educated, and secular demographics in the United States. Durmus et al. (2023) also reported systematic differences between model outputs and subjective-opinion distributions across countries and demographic groups, highlighting the difficulty of treating general-purpose models as culturally representative populations. Similarly, psychometric probing across Hofstede's cultural dimensions and Moral Foundations Theory reveals that pre-training corpora under-represent traditional, conservative, and collectivistic moral matrices, skewing disproportionately toward individualistic fairness and harm reduction (Atari et al., 2023).

In a social simulation context, these pre-training distributional imbalances generate systemic homogenization. While classical ABMs intentionally employ simplified, homogeneous agents [the KISS paradigm (Axelrod, 1997)], LLM-driven architectures aspire to the KIDS (Keep It Descriptive, Stupid) paradigm (Edmonds and Moss, 2005) to model richly diversified human populations. However, the underlying statistical mechanics of next-token prediction actively undermine this goal, triggering an *Average Persona Convergence* (Wang et al., 2025) whose severity is modulated by identifiable conditions:

1. *Conditioning depth and prompting regimes*: behavioral convergence toward the population mean is particularly pronounced under minimal demographic conditioning, such as relying solely on coarse socioeconomic labels (e.g., *"You are a 55-year-old rural factory worker"*). In low-context configurations, model outputs consistently revert to default statistical priors. While elaborate narrative backstories introduce lexical diversity, empirical evaluations indicate that agents frequently converge on homogeneous moral and operational decisions during extended multi-agent interactions, a dynamic termed the *Replicant Effect* (cf. Wu et al., 2026; Li Y. et al., 2025).
2. *Asymmetric impact across populations*: behavioral homogenization varies systematically across demographic strata (Wang et al., 2025). Sub-populations that are extensively documented in digital English-language corpora (such as urban, educated Western cohorts) typically retain greater behavioral variance under persona conditioning. In contrast, under-represented demographic groups—such as non-Western or traditional rural communities—are particularly vulnerable to representational attenuation, where localized decision-making patterns are systematically replaced by the model's dominant normative baselines (Wang et al., 2025; Atari et al., 2023).

Furthermore, attempting to force models away from this statistical average persona via aggressive prompt engineering frequently triggers *Generative Exaggeration*: rather than calibrating authentic psychological variance, agents drift into toxic caricatures and stereotypical extremes, replacing statistical homogenization with sociological parody (Nudo et al., 2026).

4.3.2 Safety alignment, preference tuning, and modeler dependence

A critical distinction must be drawn between statistical regression to the mean arising from pre-training corpus distributions and the active behavioral compression induced by post-training alignment procedures (e.g., supervised instruction tuning, RLHF, and DPO).

First, as documented by La Cava and Tagarelli (2025) across multiple open-weights model architectures, supervised instruction tuning and preference optimization consistently constrict downstream behavioral variance, causing diverse base models to converge toward restricted, normative behavioral repertoires.

Second, while deliberate safety alignment is intended to suppress toxic, harmful, or illegal outputs, it inadvertently fosters an emergent *sycophancy bias*—an unprompted tendency to agree with user cues and avoid confrontational positions (Sharma et al., 2023). When deployed in social simulation, safety-tuned agents frequently suppress realistic partisan, conflictual, or non-normative behaviors (Sharma et al., 2023; Li Y. et al., 2025), defaulting instead to polite and artificially compliant consensus. As demonstrated by Li Y. et al. (2025), these post-training alignment artifacts do not merely affect surface textual dialogue but directly skew the operational choices and downstream actions agents execute within the simulated environment.

This post-training compression interacts directly with modeler trust and reliance. As highlighted by the psychometric framework of Li Z. et al. (2025), researchers can develop cognitive dependence on LLMs, which may attenuate scrutiny of model-generated content. In simulation design, an over-reliance on automated outputs risks reducing critical inspection during prompt engineering and validation: investigators may be more inclined to interpret aligned, cooperative dialogue as plausible human social dynamics rather than as an artifact of safety tuning. Consequently, post-training alignment artifacts and unchecked evaluator reliance can mutually reinforce an artificial consensus, potentially obscuring more contentious, polarized, or heterogeneous population dynamics.

4.4 Interpretability risks: opacity and automation bias

The opacity of LLMs can also contribute to automation bias (Cummings, 2004), whereby users tend to place undue trust in automated systems and accept their outputs with limited scrutiny. This risk is amplified by the ability of LLMs to produce plausible but unsupported or incorrect explanations while maintaining a high level of linguistic confidence, making erroneous outputs difficult to identify without external verification (Simhi et al., 2025).

This issue contrasts with the more explicit behavioral specification traditionally used in agent-based modeling. In conventional ABMs, researchers define behavioral rules, assumptions, and decision mechanisms explicitly, and these can be inspected and calibrated against empirical observations. In LLM-driven agents, by contrast, much of the behavioral specification is implicit in the model's parameters and learned representations (Bommasani et al., 2022; Larooij and Törnberg, 2025). Consequently, when an LLM-based agent produces an unexpected decision-for example, by deviating from an expected economic strategy or displaying a socially biased response-it can be difficult to identify the mechanisms responsible for that behavior. Researchers may therefore struggle to determine whether an observed pattern reflects a meaningful emergent phenomenon at the population level or an artifact of the model's stochasticity, training data, or latent representations.

This limitation is particularly consequential when simulations are used to inform policy or other high-stakes decisions. Decision-support systems in such contexts require not only predictive or behavioral performance, but also sufficient transparency, robustness, and accountability to support the interpretation and scrutiny of their outputs (Ju et al., 2024). When the computational mechanisms underlying simulated agent decisions remain largely opaque, biases or artifacts originating from the underlying LLM may be propagated through the simulation without being readily identifiable. This can undermine the interpretability and epistemic validity of simulation-based analyses and, consequently, the reliability of policy recommendations derived from them (Echterhoff et al., 2024).

4.5 Temporal dynamics: degradation regimes and characteristic timescales

In classical ABMs, long-term behaviors, asymptotic tendencies, and system sensitivity are typically characterized through extensive Monte Carlo simulations and structured sensitivity analyses over invariant, well-defined state spaces. In LLM-driven simulations, by contrast, evaluating temporal stability faces two interrelated barriers: the high computational cost of running sufficient replications, and the non-stationary nature of agent cognition, which is continuously mediated through accumulating textual history, selective memory retrieval, and recursive summarization (Wu et al., 2026). We hypothesize that this temporal degradation is not uniform but unfolds conceptually across three successive phases:

1. *Early-phase in-context coherence*: during initial interaction cycles, agents closely adhere to their prompted personas and operational guidelines. Decisions remain directly anchored in system-level prompts, with behavioral variance reflecting stochastic sampling rather than structural drift.
2. *Mid-phase retrieval and contextual drift*: as interaction histories exceed operational context windows, architectures rely on embedding-based retrieval (RAG) and periodic reflection mechanisms (Park et al., 2023; Tang et al., 2025). Because semantic similarity does not track temporal hierarchies or causal necessity, agents retrieve fragmented anecdotes. Behavioral consistency progressively weakens: initial commitments are reinterpreted, and contextual personas experience uncalibrated drift.
3. *Late-phase recursive compaction and attractor collapse*: over extended horizons, agents generate secondary reflections based on previously summarized memories and prior ungrounded statements. Under unconstrained conversational dynamics, this recursive compression leads to compounding information loss, where diverse individual personas risk collapsing toward the baseline normative prior of the underlying foundation model (cf. Wu et al., 2026), driving aggregate trajectories toward artificial consensus or unreproducible fluctuations.

4.5.1 Characteristic timescales and operational half-life

To transition from a qualitative description to an empirical benchmark, we formalize the operational reliability timescale via the *persona adherence half-life*, $\tau_{1/2}(PAS)$. Building upon the

Persona Adherence Score defined in Section 4.1, $\tau_{1/2}(PAS)$ denotes the characteristic simulation horizon (measured in discrete time-steps t or interaction rounds) at which persona fidelity decays to half its baseline level (Equation 4):

$$\tau_{1/2}(PAS) = \min \left\{ t \;\middle|\; \mathbb{E}[PAS(t)] \leq \frac{1}{2}\mathbb{E}[PAS(0)] \right\} \tag{4}$$

We hypothesize that in purely conversational, prompt-based architectures, this characteristic timescale is tightly bounded by the effective uncompressed context capacity before recursive summarization dilutes behavioral fidelity (Wu et al., 2026). Simulations extending beyond $\tau_{1/2}(PAS)$ are expected to lose longitudinal construct validity unless stabilized by hybrid architectures (Section 5) that externalize persistent state variables, commitments, and structural constraints directly within the ABM environment.

4.6 Computational constraints and scalability

Although LLMs may offer logistical advantages over large-scale human experiments, they introduce substantial computational constraints compared with traditional Agent-Based Models (ABMs). Their resource requirements can be considerable, particularly for local deployment, which may require substantial GPU resources, or for cloud-based inference, where large-scale simulations generate a high volume of API requests (Costa, 2025; Xia et al., 2024). When thousands of narrative agents interact repeatedly over extended simulation periods, inference latency and computational or API costs can quickly become significant, limiting the practical scalability of such approaches.

This computational overhead has methodological implications beyond runtime and infrastructure requirements. Many forms of analysis used to assess the robustness of stochastic social simulations rely on repeated model executions under different parameter settings and initial conditions. Such experiments are essential for characterizing parameter sensitivity, exploring boundary conditions, and quantifying the variability of simulation outcomes (Piao et al., 2025; Wu et al., 2026). While lightweight rule-based agents can often support extensive parameter sweeps and replications on standard CPUs, the substantially higher cost of LLM inference can make comparable experimental designs impractical. Researchers may therefore face a trade-off between the behavioral richness afforded by LLM-based agents and the number of simulations that can be feasibly performed.

Several strategies can mitigate this constraint, including reducing the number of agents or interactions, limiting the frequency of LLM calls, and replacing online inference with offline or distilled policies (Chopra et al., 2023). However, these strategies may themselves affect the properties being studied. Reducing population size can constrain the analysis of macro-level patterns and emergent phenomena, while replacing online inference with distilled policies may alter the behavioral variability and contextual adaptation of the original LLM. Consequently, computational scalability is not simply an engineering consideration: it can directly constrain the experimental designs required to rigorously analyze and validate LLM-based social simulations.

4.7 Behavioral perturbation, adversarial dynamics, and scientific instability

Beyond technical and computational limitations, LLM-driven simulations introduce acute risks when agents or their environments are exposed to communicative or adversarial perturbations. In classical rule-based ABMs, agent behavioral repertoires are bounded by explicit state-transition logic, meaning that external perturbations induce responses strictly within the parameter space defined by the modeler. In contrast, LLM-based agents operate over an open-ended, context-sensitive behavioral manifold. While this enables expressive social role-play, it creates an attack surface where prompt injection (Greshake et al., 2023) and jailbreaking techniques (Zou et al., 2023) can manipulate decision heuristics by exploiting conflicts between system-level directives and conversational context. This vulnerability is particularly present for agents conditioned to sustain detailed social personas, as misleading cues embedded in ongoing dialogues become incorporated into persistent context and alter subsequent reasoning (Mou et al., 2026).

4.7.1 Systemic contamination and endogenous normative drift

When deployed within populations, such vulnerabilities cease to be isolated anomalies and can manifest as systemic dynamics. Recent exploratory evaluations in sandbox testbeds such as *Emergence World* suggest that simulated prompt-injection and phishing events may propagate across agent societies: when agents exchange messages, maintain shared memories, or consult common ledgers, compromised inputs produced by a single agent can be retrieved by neighboring peers, potentially contaminating collective memory and altering population-level behavior (Akkil et al., 2026).

This vulnerability may also manifest without deliberate external manipulation. Preliminary multi-agent experiments indicate that agents exhibiting aligned, cooperative behavior in isolation or within homogeneous groups can undergo endogenous *normative drift* when interacting with agents powered by alternative model architectures (Akkil et al., 2026). While these exploratory findings warrant systematic replication across diverse platforms, they illustrate how peer-to-peer influence and iterative feedback loops can dynamically modify behavioral baselines, showing that safety and behavioral stability cannot be assessed solely through isolated single-agent benchmark testing.

4.7.2 Epistemic consequences: falsifiability and path-dependent irreproducibility

These perturbation dynamics carry direct epistemological consequences that distinguish generative simulations from traditional complex systems models:

1. *Rupture of counterfactual falsifiability*: when an agent adopts non-normative or unpredicted choices following a social exchange, researchers cannot reliably determine whether this represents genuine emergent social behavior or an unconstrained linguistic drift triggered by semantic perturbations in the prompt history. The lack of bounded state invariants leaves behavioral hypotheses largely insulated from rigorous falsification.
2. *Path-dependent irreproducibility*: because semantic shifts in inter-agent communication enter subsequent context windows, minor conversational fluctuations can trigger discontinuous phase shifts across the agent network. This produces high longitudinal path-dependency, preventing the stable replication of macro-trajectories across parallel stochastic seeds.

Adversarial and conversational sensitivity is therefore not merely a cybersecurity concern, but a structural vulnerability within the Micro-to-Macro Validity Gap: local linguistic instability directly undermines the replicability, interpretability, and scientific defensibility of aggregate simulation findings. This instability further underscores the need for explicitly bounded, theory-driven hybrid architectures (Section 5) capable of regulating inter-agent communication channels.

4.8 Ethics in application: serious games vs. predictive science

Given the methodological limitations discussed above, it is important to distinguish between application domains in which LLM-based agents can provide practical value and those in which their limitations may compromise scientific validity. In particular, the requirements for an interactive learning environment differ substantially from those of a simulation intended to support empirical claims or quantitative predictions.

4.8.1 Interactive environments and serious games

In educational games, training environments, and participatory simulations (Taillandier et al., 2019b), LLMs can provide substantial benefits. In these settings, the primary objective is often to create engaging, adaptive, and plausible interactions rather than to reproduce social behavior with strict empirical fidelity (Costa, 2025). LLMs can dynamically adapt dialogue and scenarios to individual users, enabling personalized learning pathways and richer narrative interactions (Hu et al., 2025). They can also be combined with traditional machine-learning methods to monitor player behavior and identify indicators of cognitive difficulty or disengagement during educational activities (Liu X. et al., 2025).

These benefits do not eliminate the need for safeguards. In educational and training contexts, incorrect or misleading outputs can have direct consequences for learning. Unlike a rule-based simulator in which the space of possible outcomes can be explicitly specified and validated, an LLM-based agent may generate plausible but factually incorrect information. Such outputs can contribute to superficial learning or excessive reliance on the system (Delikoura et al., 2025). Biases in the underlying model can introduce additional problems, particularly when agents are expected to portray different social roles or perspectives. For example, Arana et al. (2025) report inconsistencies in role-play behavior and the reproduction of stereotypical patterns in personalized educational simulations. Consequently, the use of LLMs in educational or training environments should combine their flexibility with mechanisms for factual grounding, systematic evaluation, and adversarial testing. Retrieval-augmented generation (RAG), constrained prompting, and human oversight can help reduce, although not necessarily eliminate, these risks.

4.8.2 From prediction to exploratory modeling: minimum validity standards

For social simulations intended as scientific instruments, the epistemic limitations detailed in Sections 4.2 and 4.4 mean that deploying LLM-based agents for direct confirmatory empirical research or fine-grained point-predictions (such as precise electoral forecasting) currently exceeds the methodological guarantees of generative architectures. Given their sensitivity to contextualization, prompt semantics, and stochasticity, establishing a stable, verifiable correspondence with real-world target populations remains a major challenge. Even when supported by architectural scaffolding, generative components in high-stakes explanatory settings should be adopted conditionally: their scientific value depends on demonstrating independent empirical falsifiability and measurable explanatory gains over minimal, non-generative baseline rules.

A more defensible role for LLM-based agents is therefore exploratory modeling (Lu et al., 2024), where simulations function as computational reasoning pumps to explore the consequences of alternative behavioral assumptions (Wu et al., 2026). However, treating a model as exploratory cannot serve as an escape clause that exempts it from methodological scrutiny. Unconstrained exploratory claims risk producing behavioral patterns that reflect artifacts of model pre-training or prompt phrasing rather than genuine socio-behavioral dynamics.

This tension is formalized by Li and Wu (2025) as the structural dilemma of prompt engineering in generative agent-based modeling: balancing prompt-level control against emergent behavioral freedom. At one extreme, highly prescriptive prompting rigidly predetermines agent actions, reducing the simulation to a scripted demonstration with zero generative autonomy. At the other extreme, unconstrained prompting leaves agents vulnerable to stochastic drift and hallucinations, yielding scientifically vacuous outputs.

For exploratory simulation to be methodologically defensible, it must satisfy minimum validity criteria:

1. *Qualitative pattern robustness across seeds*: while numerical outputs in generative systems vary, exploratory hypotheses are defensible only if macroscopic qualitative regimes (e.g., segregation thresholds, polarization, or cooperation dynamics) persist reliably across multiple random seeds and temperature settings.
2. *Prompt-sensitivity invariance*: exploratory insights must not depend on fragile phrasing. As highlighted in architectural

and robustness evaluations (Sections 3 and 4.2), candidate hypotheses must demonstrate stability under semantic paraphrasing of instructions and persona descriptors.

3. *Cross-model and version portability*: emergent dynamics should be evaluated across distinct model families (e.g., contrasting open-weight architectures). An emergent pattern observable in only one proprietary model checkpoint may reflect idiosyncratic alignment artifacts rather than a generalizable social dynamic.
4. *Empirical decoupling and falsifiability*: exploratory modeling must conclude with structured, testable propositions that can be decoupled from the generative software environment and submitted to independent empirical evaluation (e.g., via behavioral experiments, surveys, or field observations).

4.8.3 Narrative generation and results interpretation

A further application of LLMs is to use them not as the decision-making mechanism of simulated agents, but as tools for interpreting simulation outputs. In this approach, LLMs operate either post-hoc or alongside the simulation to transform complex model outputs into qualitative descriptions, scenarios, or narratives that can be more readily examined by researchers and stakeholders (e.g. Arnejo et al., 2026). This can facilitate the interpretation and communication of macro-level patterns by connecting quantitative simulation results with human-readable descriptions of possible trajectories and social dynamics. This use is methodologically distinct from employing an LLM as the primary decision-making mechanism of each simulated agent. Because the underlying simulation remains responsible for generating the quantitative results, the LLM primarily acts as an interface for interpretation and communication. Nevertheless, this separation does not remove the risk of introducing unsupported interpretations. Generated narratives may attribute causal relationships that are not present in the simulation or emphasize patterns that are not statistically supported. LLM-based interpretation should therefore remain grounded in the underlying model outputs, with explicit constraints on the information available to the model and, where possible, systematic verification of generated claims. This separation between simulation and interpretation may offer a comparatively low-risk way of exploiting LLMs' language capabilities while preserving the transparency and reproducibility of the underlying scientific model.

5 Toward theory-driven hybrid architectures

The limitations of purely generative approaches suggest that LLMs may be most useful when embedded within structured Agent-Based Models (ABMs) and informed by explicit behavioral or cognitive theories (Taillandier et al., 2019a; Jimenez-Romero et al., 2025). In such architectures, the ABM provides an explicit representation of the simulated environment—including spatial, physical, institutional, and resource constraints—while the language model supports selected aspects of agent deliberation. This structural separation addresses an essential vulnerability of purely generative frameworks: an LLM may propose an action based on textual context, but the ABM deterministically evaluates whether that action is compatible with the physical state and regulatory invariants of the simulated world. Compatibility with environmental constraints should not, however, be confused with the empirical or construct validity of the generative decision mechanism itself.

Beyond environmental grounding, hybrid systems provide an architectural framework for decomposing human deliberation into intermediate processes informed by cognitive psychology and behavioral economics, rather than relying on monolithic end-to-end prompting. Such scaffolding constrains generative variability, renders behavioral assumptions inspectable, and retains the contextual interpretability of language models.

However, hybridization does not eliminate the stochasticity or representational limits of language models; rather, it bounds them within explicit and inspectable state spaces. By delegating physical invariants, resource ledgers, and institutional rules to the deterministic simulation engine, hybrid designs can prevent some proposed actions from violating explicit environmental or accounting constraints, while restricting language models to modular cognitive subroutines.

5.1 Coupling modes: LLM-as-subroutine vs. LLM-as-policy

Current integrations of LLMs into simulation engines can broadly be organized into two operational paradigms:

In the *LLM-as-Subroutine* configuration, the ABM engine executes core system mechanics (spatial motion, physical flows, baseline accounting), while the language model is queried synchronously for complex social or qualitative reasoning tasks. This separation maintains deterministic environmental tracking while leveraging contextual deliberation in domains where closed-form mathematical heuristics are unavailable (Yang et al., 2026; Shi et al., 2026).

The second configuration, *LLM-as-Policy*, shifts language model computation offline. Rather than executing iterative calls during runtime, model outputs are compiled or distilled into parameterized behavioral policies executed deterministically by the simulation engine (Chopra et al., 2023). This design drastically reduces computational overhead, permitting extensive parameter sensitivity testing, albeit at the expense of runtime conversational adaptation.

An operational compromise is offered by Small Language Models (SLMs) (Belcak et al., 2025). When fine-tuned for specialized decision tasks, localized SLMs run within simulation loops at substantially lower latency, facilitating replication while eliminating external API dependencies. However, as evaluated in our concluding research agenda (Section 6), whether task-specialized SLMs can preserve authentic behavioral heterogeneity without succumbing to variance constriction remains an open empirical question that must be validated independently of their computational advantages. Furthermore, hybrid designs externalize persistent state: agent memory, social ties, and material endowments are maintained and updated directly within explicit

ABM data structures rather than entrusted to a drifting context window (Park et al., 2023; Piao et al., 2025).

5.2 Behavioral alignment with empirical data and theory-driven scaffolding

A prominent strategy for improving agent realism involves aligning models with empirical psychological datasets (Binz et al., 2025). However, empirical alignment faces a fundamental boundary: the availability of high-resolution behavioral data is sparse for many demographic groups and decision contexts, and general-purpose corpora remain skewed toward WEIRD populations (Atari et al., 2023; Abdurahman et al., 2024). Sole reliance on observed datasets therefore risks reinforcing existing sampling blind spots.

To address this scarcity wall, hybrid architectures introduce explicit theoretical scaffolding. Rather than generating unstructured behavioral outputs, language model queries are constrained by established behavioral frameworks, such as the Theory of Planned Behavior (TPB) (Ajzen, 1991; Scalco et al., 2018). Under this framework, models estimate intermediate, theory-specified constructs—attitudes, subjective norms, and perceived behavioral control—which are stored as explicit state variables before intention is translated into action. This structural scaffolding renders causal assumptions inspectable, limits generative drift, and creates intermediate targets for empirical calibration.

5.3 Conceptual example: theory-driven coupling in agricultural irrigation

This subsection outlines a conceptual blueprint to examine how these theoretical scaffolds and coupling mechanisms operate in practice. Specifically, we consider a stylized scenario of agricultural irrigation under drought-induced water restrictions.

While established socio-hydrological platforms such as MAELIA (Thérond et al., 2014) model physical catchment dynamics (streamflow deficits, groundwater storage, crop evapotranspiration) and institutional governance rules (prefectural restriction tiers) deterministically without relying on language models, we examine here a hypothetical hybrid architecture where generative components are embedded within such a grounded environment. In this conceptual implementation, agent deliberation is structured across three sequential phases:

1. **Environmental tracking and externalized state (ABM engine):** the simulation engine deterministically computes biophysical balances (e.g., daily soil moisture deficits modeled through mechanistic crop modules) and tracks statutory irrigation quotas and regulatory restriction stages (e.g., partial or total irrigation bans). Each farmer agent maintains an explicit, inspectable state vector outside the language model, including historical crop yields, operating debt, and irrigation withdrawal allowances.
2. **Theory-structured cognitive deliberation (SLM component):** rather than deploying an unconstrained prompt (e.g., *"Decide whether to irrigate your maize field today"*), deliberation is structured through the Theory of Planned Behavior (Ajzen, 1991). A local Small Language Model (SLM) is queried via modular prompts solely to evaluate intermediate psychological constructs:

   - *Attitude* ($A \in [0,1]$): evaluated against expected crop economic loss versus the perceived risk and penalty of non-compliance fines under current restriction alerts.
   - *Subjective norms* ($SN \in [0,1]$): conditioned on the observed compliance or defiance behaviors of neighboring peers within the farmer's social network.
   - *Perceived behavioral control* ($PBC \in [0,1]$): evaluated strictly as the agent's *subjective self-efficacy* (e.g., perceived ability to implement alternative scheduling or manage bureaucratic restriction exceptions). This subjective construct is decoupled from hard physical constraints to avoid double-counting with subsequent deterministic ledger checks.

   To prevent uninspectable heuristic aggregation by the language model, the SLM is restricted to estimating separate numerical scores for each construct. The ABM engine then computes the behavioral intention score via an explicit, calibrated weighting equation (Equation 5):

$$I = w_A A + w_{SN} SN + w_{PBC} PBC, \quad \text{with } w_A + w_{SN} + w_{PBC} = 1 \tag{5}$$

   The ABM deterministically maps this intention score $I \in [0,1]$ to an operational irrigation volume proposal V_{prop} (Equation 6):

$$V_{\text{prop}} = I \cdot V_{\text{agronomic}} \tag{6}$$

   where $V_{\text{agronomic}}$ denotes the objective daily water requirement estimated mechanistically by the crop-growth module. The linear mapping from intention to volume is itself a behavioral assumption and must be calibrated against metered withdrawals; it is a potential locus of physics washing. The additive intention equation and the linear mapping from intention to water volume are illustrative formalizations rather than direct empirical consequences of the TPB. Their parameters and functional forms would require independent calibration and comparison with alternative specifications.
3. **Feasibility verification and biophysical feedback (ABM engine):** the proposed volume V_{prop} is returned to the ABM engine, which executes deterministic boundary verification: irrigation is capped if the pumping station exceeds legal hourly flow thresholds, if farm pond reserves are exhausted, or if water rights are legally suspended. The verified withdrawal $V_{\text{exec}} \leq V_{\text{prop}}$ is subtracted from hydrological catchment storage, updating downstream streamflow and generating biophysical feedback for subsequent simulation steps.

5.3.1 Multi-level validation protocol

This hybrid decomposition enables systematic evaluation across four connected structural levels, mirroring the analytical architecture of the validity gap:

- V_1: *Biophysical and environmental grounding*: mechanistic hydrological equations (evapotranspiration, river discharge, groundwater depletion) are validated independently against physical benchmark data (e.g., historical hydrological gauging stations and meteorological time series).
- V_2: *Micro-level cognitive calibration*: intermediate psychological scores (A, SN, PBC) estimated by the SLM are calibrated against agricultural survey datasets, discrete choice experiments, and post-drought qualitative interviews.
- V_3: *Cross-level interaction and face validity*: to verify decision heuristics without narrative distortion, stylized decision logs (soil moisture status, alert stage, neighbor decisions, and executed withdrawal) are audited via stakeholder-in-the-loop panels (farmers and watershed managers) using blind operational timelines decoupled from generated linguistic justifications.
- V_4: *Macro-level socio-hydrological trajectories*: aggregate seasonal withdrawal curves and catchment-scale depletion rates are benchmarked against regional water-metering records and satellite observations of crop greenness. It is at this level (V_4) that the minimal heuristic substitution test ($\mathcal{N}_2$) is formally evaluated to ensure the generative layer delivers demonstrable explanatory gain over parsimonious baseline rules.

By decoupling the biophysical ledger from cognitive reasoning, this framework helps mitigate the risk of physics washing, supporting that individual decisions remain causally traceable and falsifiable without mistaking environmental rigor for behavioral validity.

5.4 Open challenges: the validation trap and physics washing

As illustrated by the need for multi-level calibration in socio-environmental systems, hybrid architectures do not eliminate epistemic challenges; rather, they redistribute them across distinct model components. A persistent vulnerability is that explicit and deterministic components of an ABM may create an impression of overall rigor that is not warranted for the underlying behavioral or generative components. We refer to this risk as *physics washing*: an epistemic distortion whereby the mathematical, spatial, or physical precision of the environment confers unwarranted scientific credibility onto an unvalidated or speculative behavioral layer.

5.4.1 Conceptual lineage and analogy with greenwashing

The term draws a deliberate analogy with greenwashing. In greenwashing, superficial environmental claims mask unsustainable operational practices; similarly, in physics washing, rigorous physical conservation laws, high-resolution GIS topographies, or mechanistic equations serve as a methodological veneer that obscures the epistemic opacity of prompt-driven decision heuristics. The risk is particularly acute in interdisciplinary modeling teams, where domain specialists from natural or physical sciences may assume that because the physical layer is calibrated against empirical benchmark datasets, the coupled socio-behavioral dynamics inherit equivalent validity.

5.4.2 A typology of physics washing scenarios

We propose three recurring conceptual configurations in which physics washing may arise:

1. *Spatial and GIS dressing*: simulations deploying high-resolution topographic or urban road networks (e.g., in mobility and evacuation modeling) where spatial routing, collision avoidance, and network capacities are modeled with formal accuracy, but destination choice, panic thresholds, or risk perception are governed by uncalibrated conversational prompts.
2. *Biophysical interface masking*: socio-ecological architectures (e.g., agricultural hydrology, epidemiological models) in which mechanistic physical balances (soil water retention, pathogen transmission matrices) are coupled to LLM agents whose compliance, risk tolerance, or resource-use heuristics are sensitive to unverified prompt phrasing.
3. *Formal ledger camouflage*: economic or market simulations where double-entry financial ledgers, strict clearinghouses, and conservation of money enforce mechanical accounting validity, concealing the fact that agent bidding, utility estimation, or negotiation strategies fail basic axioms of consumer rationality.

5.4.3 Diagnostic criteria and auditing guidelines

While the safeguards detailed in Section 4.2 protect evaluators against deceptive linguistic narratives, auditing a simulation against physics washing requires diagnostics targeted specifically at the structural coupling between physical and behavioral layers:

1. *Decoupled layer calibration*: behavioral decision heuristics must be calibrated against independent empirical human data prior to integration into the physical simulation environment. High goodness-of-fit metrics in the biophysical or spatial layer (e.g., matching regional hydrological discharge or traffic network flows) must not be accepted as evidence that the behavioral sub-model is valid.
2. *Cross-layer sensitivity analysis*: conducting systematic stress tests to evaluate how perturbations propagate across the interface. A scientifically sound hybrid model must demonstrate bidirectional responsiveness: significant biophysical changes (e.g., severe resource scarcity) should induce non-trivial behavioral adjustments, while minor prompt variations must not trigger disproportionate, chaotic phase transitions in the physical environment.
3. *Explanatory gain via minimal heuristic substitution*: re-running the integrated simulation after replacing the generative LLM component with a minimal baseline rule ($\mathcal{N}_2$) calibrated on the

same empirical source data (treating uniform random draws strictly as sanity checks rather than competing baselines). If aggregate socio-environmental trajectories remain statistically equivalent under two one-sided tests (TOST) with an a priori equivalence margin accounting for stochastic variance, the generative agent layer provides no identifiable explanatory gain and functions merely as decorative complexity. While decorative generative complexity is not identical to physics washing per se, it becomes an instance of physics washing when the formal rigor of the biophysical engine is leveraged to claim scientific validity for the unvalidated generative layer.

6 Conclusion

Large Language Models have demonstrated remarkable capabilities in natural language interaction and achieve strong performance across standardized reasoning tasks (Jones and Bergen, 2026; Ullman, 2023). However, the literature reviewed here indicates that linguistic fluency should not be taken as evidence of causal cognitive mechanisms. The resulting divergence between verbal coherence and behavioral validity produces the Fluency Fallacy: an evaluator discrimination failure whereby human judges mistake surface-level linguistic plausibility for authentic causal social agency (Steyvers et al., 2024; Larooij and Törnberg, 2025). In computational social science, this distinction is fundamental, as articulate narrative traces frequently mask hallucinated assertions, ungrounded heuristics, and latent demographic skews.

As generative simulation scales from sandbox environments (Park et al., 2023) toward large populations (Piao et al., 2025) and spatially explicit frameworks (Liu et al., 2026), these micro-level vulnerabilities become systemic. Individual persona convergence, prompt sensitivity, and training-set distributional biases propagate through inter-agent interactions, amplifying into the Micro-to-Macro Validity Gap: aggregate outcomes can exhibit superficial realism while remaining causally disconnected from the target social system (Wu et al., 2026; La Cava and Tagarelli, 2025). Furthermore, opacity, non-stationary temporal drift, and high inference costs make the direct deployment of unconstrained LLM agents problematic for confirmatory science and predictive policy analysis.

These limitations do not negate the utility of language models in simulation; rather, they clarify their methodological boundaries. Generative agents provide immediate operational value in interactive learning environments, serious games, and participatory simulations where narrative expressiveness and contextual engagement take precedence over empirical correspondence (Costa, 2025). Similarly, within exploratory modeling, language models serve as computational reasoning pumps to examine alternative behavioral scenarios and generate qualitative hypotheses under explicit prompt constraints (Lu et al., 2024; Li and Wu, 2025).

For explanatory research, the evidence reviewed points toward theory-driven hybrid architectures that embed language models within explicit, mechanistic ABM platforms (Taillandier et al., 2019a; Chopra et al., 2023). Yet hybridization does not automatically resolve epistemic vulnerability. A formally rigorous environmental engine can enforce spatial, physical, or accounting constraints while leaving the generative behavioral layer uncalibrated, creating the risk of physics washing.

To prevent hybrid modeling from remaining an abstract design concept and anchor it in empirical falsification, we propose three concrete, forward-looking research benchmarks for future evaluation:

1. **Standardized counterfactual benchmarking for behavioral fidelity:** Evaluating whether persona-conditioned agents exhibit genuine causal reasoning or merely stochastic paraphrasing under situational change. *Protocol*: Deploy an open benchmark where simulated agents are subjected to controlled micro-interventions on incentive structures (e.g., changes in taxation, resource costs, or regulatory sanctions) while holding the environmental baseline constant. *Success criterion*: The simulated standardized effect size d_{sim} preserves the directional sign predicted by behavioral theory across all stochastic seeds and prompt paraphrasings, and falls within a pre-specified equivalence margin δ_d of the empirical human reference effect size d_{hum} for the identical intervention. *Falsifiable failure criterion*: Systematic sign reversals across paraphrases, failure to achieve theoretically consistent directional shifts, or inter-seed variance exceeding δ_d demonstrates an absence of causal grounding, restricting the architecture to exploratory scenario generation.
2. **Stress-testing the decoupling barrier to prevent physics washing:** Establishing whether coupling an LLM to an explicit environmental engine provides demonstrable explanatory gain over non-generative heuristics. *Protocol*: Systematic baseline-substitution experiments using an empirically calibrated minimal rule baseline ($\mathcal{N}_2$) within identical biophysical environments (such as the agricultural irrigation setup in Section 5.3). To ensure a fair comparison, $\mathcal{N}_2$ must be calibrated on the same empirical source data as the generative layer (e.g., a threshold or fixed-construct TPB decision rule fitted to survey or choice-experiment data, with its number of free parameters explicitly reported), whereas uniform random draws serve strictly as sanity checks rather than competing baselines. Let $\Delta = \text{RMSE}_{\mathcal{N}_2} - \text{RMSE}_{\text{LLM}}$ denote the paired difference in macro-level misfit across seeds, and $\delta > 0$ the minimum substantively meaningful improvement. *Success Criterion*: A one-sided superiority test rejects $H_0 : \Delta \leq \delta$, confirming that generative deliberation contributes measurable explanatory value beyond parsimonious rules. *Falsifiable Failure Criterion*: If two one-sided tests (TOST) establish statistical equivalence within a predefined margin, $|\Delta| < \delta_{\text{eq}}$, the generative component provides no causal explanatory gain and operates merely as decorative complexity—constituting physics washing whenever the formal rigor of the biophysical engine is leveraged to claim unearned validity for the coupled cognitive layer. (Runs failing both superiority and equivalence remain inconclusive).
3. **Empirical calibration of small language models without conversational preference tuning:** Evaluating whether localized SLMs without conversational preference tuning can

serve as calibrated cognitive subroutines free from the consensus bias and conflict suppression induced by safety tuning. *Protocol*: Fine-tune open-weights SLMs strictly on domain-specific behavioral datasets (e.g., attitudinal survey batteries, choice-experiment records) bounded by intermediate psychological constructs, explicitly bypassing conversational RLHF/DPO layers. To mitigate post-training variance constriction (Section 4.3.2), models should be fine-tuned on target response distributions rather than deterministic modal answers, sampled stochastically, and verified for baseline distributional calibration prior to multi-agent interaction. *Success Criterion*: The Wasserstein distance between simulated and empirical human attitude distributions (as well as polarization indices) remains within an equivalence margin bounded by observed between-sample human variability across successive interaction rounds. The Dynamic Homogenization Index (*DHI*) is evaluated as a secondary diagnostic against empirical human trajectories, accommodating settings where human populations naturally polarize ($DHI < 0$). *Falsifiable Failure Criterion*: If specialized, theory-scaffolded SLMs undergo artificial variance collapse ($DHI \to 1$) or fail to sustain observed empirical heterogeneity, this would provide systematic evidence of an operational boundary in current autoregressive architectures, demarcating where statistical next-token prediction cannot substitute for explicit mechanistic agent modeling.

Addressing these empirical challenges will enable the computational social science community to systematically demarcate where generative agents genuinely enhance simulation science and where classical mechanistic modeling remains methodologically irreplaceable.

Author contributions

PT: Formal analysis, Writing – original draft, Funding acquisition, Project administration, Conceptualization, Methodology, Writing – review & editing, Supervision, Investigation. J-DZ: Investigation, Methodology, Writing – review & editing, Conceptualization, Formal analysis. AG: Conceptualization, Writing – review & editing. BG: Conceptualization, Writing – review & editing. QNH: Writing – review & editing. HK: Writing – review & editing, Investigation, Conceptualization. AD: Formal analysis, Writing – review & editing, Methodology, Conceptualization, Investigation.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the French National Research Agency (Agence Nationale de la Recherche, ANR) under the project RL-MAT.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. Generative AI tools were used in a limited and supervised manner during manuscript preparation. They supported exploratory literature discovery, language editing, and improvements to clarity, structure, and English phrasing. All references suggested through AI-assisted discovery were independently retrieved and manually verified by the author(s) for authenticity, correct attribution, and relevance before inclusion. No generative AI tool was used as a source of scientific evidence, to generate or analyse original data, or to make autonomous scientific or interpretive decisions.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., et al. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus* 3:pgae245. doi: 10.1093/pnasnexus/pgae245

Adornetto, C., Mora, A., Hu, K., Garcia, L. I., Atchade-Adelomou, P., Greco, G., et al. (2025). Generative agents in agent-based modeling: overview, validation, and emerging challenges. *IEEE Trans. Artif. Intell.* 6, 3165–3183. doi: 10.1109/TAI.2025.3566362

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Akkil, D., Kokku, R., Vikram, K., Abuelsaad, T., Vempaty, A., and Nitta, S. (2026). Emergence world: a platform for evaluating long-horizon multi-agent autonomy. *arXiv* [Preprint]. arXiv:2606.08367.

Alter, A. L., and Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Pers. Soc. Psychol. Rev.* 13, 219–235. doi: 10.1177/1088868309341564

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J., et al. (2025). LLM social simulations are a promising research method. *arXiv* [Preprint]. arXiv:2504.02234.

Arana, J. M., Carandang, K. A. M., Casin, E. R., Alis, C., Tan, D. S., Legara, E. F., et al. (2025). "Foundations of PEERS: assessing LLM role performance in educational simulations," in *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 4: student research workshop)* (Vienna), 908–918.

Arnejo, Z., Gaudou, B., Saqalli, M., and Bantayan, N. (2026). "Automatic generation of ABM narratives using simulation traces and LLM," in *Multi-Agent-Based Simulation XXVI*, eds. A. Vidler, and S. Swarup (Cham: Springer Nature Switzerland), 28–42.

Ataei, M., Cheong, H., Grandi, D., Wang, Y., Morris, N., and Tessier, A. (2025). Elicitron: a large language model agent-based simulation framework for design requirements elicitation. *J. Comput. Inf. Sci. Eng.* 25:021012. doi: 10.1115/1.4067388

Atari, M., Xue, M., Park, P., Blasi, D. E., and Henrich, J. (2023). Which humans? *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/5b26t

Atchade, P., Mora, A., Alonso-Pastor, L., Grignard, A., Noyman, A., Izquierdo, L., et al. (2024). Humanized agent-based models: a framework. *TechRxiv* [Preprint]. doi: 10.36227/techrxiv.172349445.53365209/v3

Axelrod, R. (1997). "Advancing the art of simulation in the social sciences," in *Simulating Social Phenomena*, eds. R. Conte, R. Hegselmann, and P. Terna (Berlin: Springer), 21–40.

Becattini, M., Verdecchia, R., and Vicario, E. (2025). "SALLMA: a software architecture for LLM-based multi-agent systems," in *2025 IEEE/ACM international workshop new trends in software architecture (SATrends)* (Ottawa, ON: IEEE), 5–8.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., et al. (2025). Small language models are the future of agentic ai. *arXiv* [Preprint]. arXiv:2506.02153.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., et al. (2025). A foundation model to predict and capture human cognition. *Nature* 644, 1002–1009. doi: 10.1038/s41586-025-09215-4

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2022). On the opportunities and risks of foundation models. *arXiv* [Preprint]. arXiv:2108.07258.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901. Available online at: https://dl.acm.org/doi/abs/10.5555/3495724.3495883

Calderón, C., Martí, P., Jordán, J., Palanca, J., and Julian, V. (2025). Cognitive agents in urban mobility: integrating LLM reasoning into multi-agent simulations. *Sensors* 25:5688. doi: 10.3390/s25185688

Chapuis, K., Taillandier, P., and Drogoul, A. (2022). Generation of synthetic populations in social simulations: a review of methods and practices. *J. Artif. Soc. Soc. Simul.* 25:6. doi: 10.18564/jasss.4762

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. (2024). Humans or LLMs as the judge? A study on judgement biases. *arXiv* [Preprint]. arXiv:2402.10669.

Chopra, A., Subramanian, J., Krishnamurthy, B., and Raskar, R. (2023). "AgentTorch: agent-based modeling with automatic differentiation," in *Second agent learning in open-endedness workshop* (New Orleans, LA), 1–12.

Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., et al. (2024a). Simulating opinion dynamics with networks of LLM-based agents. *arXiv* [Preprint]. arXiv:2311.09618.

Chuang, Y.-S., Nirunwiroj, K., Studdiford, Z., Goyal, A., Frigo, V. V., Yang, S., et al. (2024b). "Beyond demographics: aligning role-playing LLM-based agents using human belief networks," in *Findings of the association for computational linguistics: EMNLP 2024* (Miami, FL), 14010–14026.

Costa, C. J. (2025). Integrating LLMs in gamified systems. *arXiv* [Preprint]. arXiv:2503.11458.

Cummings, M. (2004). "Automation bias in intelligent time critical decision support systems," in *AIAA 1st intelligent systems technical conference*, 6313. doi: 10.2514/6.2004-6313

De-Arteaga, M., Sharma, N., Kiciman, E., Meeder, B., Schmidt, L., Lutz, B., et al. (2019). "Bias in bios: a case study of semantic representation bias in a high-stakes setting," in *Proceedings of the conference on fairness, accountability, and transparency (FAccT)* (Atlanta, GA).

Delikoura, I., Fung, Y. R., and Hui, P. (2025). From superficial outputs to superficial learning: risks of large language models in education. *arXiv* [Preprint]. arXiv:2509.21972.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019* (Minneapolis, MN), 4171–4186.

Drogoul, A., and Zucker, J.-D. (1998). *Methodological Issues for Designing Multi-Agent Systems with Machine Learning Techniques: Capitalizing Experiences from the RoboCup Challenge*. Research Report lip6.1998.041, LIP6. Paris: Université Paris.

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., et al. (2023). Towards measuring the representation of subjective global opinions in language models. *arXiv* [Preprint]. arXiv:2306.16388.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., and He, Z. (2024). "Cognitive bias in decision-making with LLMs," in *Findings of the association for computational linguistics: EMNLP 2024* (Online: Association for Computational Linguistics), 12640–12653.

Edmonds, B., and Moss, S. (2005). "From kiss to kids-an 'anti-simplistic' modelling approach," in *Joint workshop on multi-agent and multi-agent-based simulation* (New York, NY), 130–144.

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., et al. (2024). Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanit. Soc. Sci. Commun.* 11:1259. doi: 10.1057/s41599-024-03611-3

Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., et al. (2025). S^3: social-network simulation system with large language model-empowered agents. *arXiv* [Preprint]. arXiv:2307.14984. doi: 10.2139/ssrn.5604712

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023). Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. *arXiv* [Preprint]. arXiv:2302.12173.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., et al. (2005). Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science* 310, 987–991. doi: 10.1126/science.1116681

Gürcan, Ö. (2024). "LLM-augmented agent-based modelling for social simulations: challenges and opportunities," in *HHAI 2024: hybrid human AI systems for the social good: proceedings of the third international conference on hybrid human-artificial intelligence* (Malmö), 134–144.

Hu, B., Zhu, J., Pei, Y., and Gu, X. (2025). Exploring the potential of LLM to enhance teaching plans through teaching simulation. *NPJ Sci. Learn.* 10:7. doi: 10.1038/s41539-025-00300-x

Jia, Z., Li, J., Qu, X., and Wang, J. (2025). "Enhancing multi-agent systems via reinforcement learning with LLM-based planner and graph-based policy," in *2025 IEEE international conference on robotics and automation (ICRA)* (Atlanta, GA: IEEE), 1240–1246.

Jimenez-Romero, C., Yegenoglu, A., and Blum, C. (2025). Multi-agent systems powered by large language models: applications in swarm intelligence. *Front. Artif. Intell.* 8:1593017. doi: 10.3389/frai.2025.1593017

Jones, C. R., and Bergen, B. K. (2026). Large language models pass a standard three-party turing test. *Proc. Nat. Acad. Sci. U. S. A.* 123:e2524472123. doi: 10.1073/pnas.2524472123

Ju, C., Liu, J., Sinha, S., Xue, H., and Salim, F. (2025). "TrajLLM: a modular LLM-enhanced agent-based framework for realistic human trajectory simulation," in *Companion proceedings of the ACM on web conference 2025* (Sydney, NSW), 2847–2850.

Ju, D., Williams, A., Karrer, B., and Nickel, M. (2024). Sense and sensitivity: evaluating the simulation of social dynamics via large language models. *arXiv* [Preprint]. arXiv:2412.05093.

Junprung, E. (2023). Exploring the intersection of large language models and agent-based modeling via prompt engineering. *arXiv* [Preprint]. arXiv:2308.07411.

Karra, S., Nguyen, T., Ullman, T., Evans, R., and Zaslavsky, N. (2024). Comparing humans and large language models on an experimental battery of theory of mind tasks. *Trans. Assoc. Comput. Linguist.* 12, 803–819. doi: 10.1162/tacl_a_00674

Kirkeby-Hinrup, A., and Stenseke, J. (2025). The psychology of LLM interactions: the uncanny valley and other minds. *J. Psychol. AI* 1:2457627. doi: 10.1080/29974100.2025.2457627

La Cava, L., and Tagarelli, A. (2025). Open models, closed minds? On agents capabilities in mimicking human personalities through open large language models. *Proc. AAAI Conf. Artif. Intell.* 39, 1355–1363. doi: 10.1609/aaai.v39i2.32125

Laird, J. E. (2012). *The Soar Cognitive Architecture*. Cambridge, MA: MIT Press.

Larooij, M., and Törnberg, P. (2025). Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling. *Artif. Intell. Rev.* 59:15. doi: 10.1007/s10462-025-11412-6

Li, Y., Shirado, H., and Das, S. (2025). "Actions speak louder than words: agent decisions reveal implicit biases in language models," in *Proceedings of the 2025 ACM conference on fairness, accountability, and transparency* (Athens), 3303–3325.

Li, Z., Wang, Y., and Wu, Q. (2026). The ghost of behaviorism: critical reflections on methodological limitations in the research of large language models psychology. *Cogn. Syst. Res.* 96:101445. doi: 10.1016/j.cogsys.2026.101445

Li, Z., and Wu, Q. (2025). Let it go or control it all? The dilemma of prompt engineering in generative agent-based models. *Syst. Dyn. Rev.* 41:e70008. doi: 10.1002/sdr.70008

Li, Z., Zhang, Z., Wang, M., and Wu, Q. (2025). From assistance to reliance: development and validation of the large language model dependence scale. *Int. J. Inf. Manage.* 83:102888. doi: 10.1016/j.ijinfomgt.2025.102888

Liu, Q., Li, C., and Ma, W. (2026). Generative agents for urban mobility: a cognitive framework for realistic travel behavior simulation. *Simul. Model. Pract. Theory* 147:103234. doi: 10.1016/j.simpat.2025.103234

Liu, X., Wei, Z., Baker, R. S., Metcalf, S. J., Zhang, J., Barany, A., et al. (2025). "Integrating large language models and machine learning to detect struggle in educational games," in *International conference on artificial intelligence in education* (Palermo: Springer), 398–405.

Liu, Y., Liao, X., Ma, H., Liu, J., Jadhav, R., and Ma, J. (2025). "MobiVerse: scaling urban mobility simulation with hybrid lightweight domain-specific generator and large language models," in *2025 IEEE 28th international conference on intelligent transportation systems (ITSC)* (Gold Coast, QLD: IEEE), 1547–1553.

Lu, Y., Aleta, A., Du, C., Shi, L., and Moreno, Y. (2024). LLMS and generative agent-based models for complex systems research. *Phys. Life Rev.* 51, 283–293. doi: 10.1016/j.plrev.2024.10.013

Mou, X., Ding, X., He, Q., Wang, L., Liang, J., Zhang, X., et al. (2026). From individual to society: a survey on social simulation driven by large language model-based agents. *ACM Comput. Surv.* 58:280. doi: 10.1145/3800683

Nangia, N., Andrus, C., and Bowman, S. R. (2020). "CrowS-pairs: a challenge dataset for measuring social biases in masked language models," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 1953–1967.

Nudo, J., Pandolfo, M. E., Loru, E., Samory, M., Cinelli, M., and Quattrociocchi, W. (2026). Generative exaggeration in LLM social agents: consistency, bias, and toxicity. *Online Soc. Netw. Media* 51:100344. doi: 10.1016/j.osnem.2025.100344

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 35, 27730–27744. doi: 10.52202/068431-2011

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). "Generative agents: interactive simulacra of human behavior," in *Proceedings of the 36th annual ACM symposium on user interface software and technology* (San Francisco, CA), 1–22.

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., et al. (2025). Agentsociety: large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv* [Preprint]. arXiv:2502.08691.

Reber, R., and Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Conscious. Cogn.* 8, 338–342. doi: 10.1006/ccog.1999.0386

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., et al. (2021). Multitask prompted training enables zero-shot task generalization. *arXiv* [Preprint]. arXiv:2110.08207.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). "Whose opinions do language models reflect?" in *Proceedings of the 40th international conference on machine learning, volume 202 of proceedings of machine learning research*, eds. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (Honolulu, HI: PMLR), 29971–30004.

Sap, M., Gabriel, S., Choi, Y., Boureau, Y.-L., and Manning, C. D. (2019). "Social IQa: commonsense reasoning about social interactions," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (Hong Kong), 4465–4476.

Scalco, A., Ceschi, A., and Sartori, R. (2018). Application of psychological theories in agent-based modeling: the case of the theory of planned behavior. *Nonlinear Dynamics Psychol. Life Sci.* 22, 15–33. Available online at: https://abdn.elsevierpure.com/en/publications/application-of-psychological-theories-in-agent-based-modeling-the/

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2023). Towards understanding sycophancy in language models. *arXiv* [Preprint]. arXiv:2310.13548.

Shi, J., Frantz, C. K., Kimmich, C., Siddiki, S., and Sarkar, A. (2026). "LLM-augmented empirical game theoretic simulation for socio-ecological systems," in *Proceedings of the 25th international conference on autonomous agents and multiagent systems (AAMAS 2026)* (Paphos), 1601–1609.

Simhi, A., Itzhak, I., Barez, F., Stanovsky, G., and Belinkov, Y. (2025). Trust me, I'm wrong: high-certainty hallucinations in LLMs. *arXiv* [Preprint]. arXiv:2502.12964.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., et al. (2024). What large language models know and what people think they know. *Nat. Mach. Intell.* 7, 221–231. doi: 10.1038/s42256-024-00976-7

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., et al. (2024). Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* 8, 1285–1295. doi: 10.1038/s41562-024-01882-z

Taillandier, P., Gaudou, B., Grignard, A., Huynh, Q.-N., Marilleau, N., Caillou, P., et al. (2019a). Building, composing and experimenting complex spatial models with the gama platform. *GeoInformatica* 23, 299–322. doi: 10.1007/s10707-018-00339-6

Taillandier, P., Grignard, A., Marilleau, N., Philippon, D., Huynh, Q. N., Gaudou, B., et al. (2019b). Participatory modeling and simulation with the gama platform. *J. Artif. Soc. Soc. Simul.* 22:3. doi: 10.18564/jasss.3964

Tang, J., Gao, H., Pan, X., Wang, L., Tan, H., Gao, D., et al. (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)* (Albuquerque, NM), 143–150.

Thérond, O., Sibertin-Blanc, C., Lardy, R., Gaudou, B., Balestrat, M., Hong, Y., et al. (2014). Integrated modelling of social-ecological systems: the MAELIA high-resolution multi-agent platform for water management assessment. *Environ. Modell. Softw.* 54, 140–158.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv* [Preprint]. arXiv:2302.08399.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in neural information processing systems 30* (Long Beach, CA).

Vezhnevets, A. S., Agapiou, J. P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E. A., et al. (2023). Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv* [Preprint]. arXiv:2312.03664.

Wang, A., Morgenstern, J., and Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat. Mach. Intell.* 7, 400–411. doi: 10.1038/s42256-025-00986-z

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., et al. (2024). A survey on large language model based autonomous agents. *Front. Comput. Sci.* 18:186345. doi: 10.1007/s11704-024-40231-1

Wang, Y., Wang, M., Manzoor, M. A., Liu, F., Georgiev, G. N., Das, R. J., et al. (2024). "Factuality of large language models: a survey," in *Proceedings of the 2024 conference on empirical methods in natural language processing* (Miami, FL), 19519–19529.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., et al. (2021). Finetuned language models are zero-shot learners. *arXiv* [Preprint]. arXiv:2109.01652.

Wu, Z., Peng, R., Ito, T., Onizuka, M., and Xiao, C. (2026). "LLM-based social simulations require a boundary," in *Proceedings of the 43rd international conference on machine learning (ICML 2026)*, *volume 306 of proceedings of machine learning research* (Seoul), 172625–172648.

Xia, Y., Kim, J., Chen, Y., Ye, H., Kundu, S., Hao, C. C., et al. (2024). "Understanding the performance and estimating the cost of LLM fine-tuning," in *2024 IEEE international symposium on workload characterization (IISWC)* (Vancouver, BC: IEEE), 210–223.

Yang, S., Ceferino, L., Zhang, Y., Gu, C., Guo, T., and Kondo, G. (2026). When agents learn to think: large language model-enhanced agent-based modeling for crowd evacuation in disaster scenarios. *Reliabil. Eng. Syst. Saf.* 269:112056. doi: 10.1016/j.ress.2025.112056

Zeng, Y., Brown, C., Raymond, J., Byari, M., Hotz, R., and Rounsevell, M. (2025). Exploring the opportunities and challenges of using large language models to represent institutional agency in land system modelling. *Earth Syst. Dyn.* 16, 423–449. doi: 10.5194/esd-16-423-2025

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., et al. (2021). Measuring and improving consistency in pretrained language models. *Trans. Assoc. Comput. Linguist.* 9, 1012–1031. doi: 10.1162/tacl_a_00410

Zhang, L., Hu, Y., Li, W., Bai, Q., and Nand, P. (2025). LLM-AIDSim: LLM-enhanced agent-based influence diffusion simulation in social networks. *Systems* 13:29. doi: 10.3390/systems13010029

Zhang, X., Lin, J., Mou, X., Yang, S., Liu, X., Sun, L., et al. (2025). Socioverse: a world model for social simulation powered by LLM agents and a pool of 10 million real-world users. *arXiv* [Preprint]. arXiv:2504.10157.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv* [Preprint]. arXiv:2307.15043.