
VALIDATING GENERATIVE AGENT-BASED MODELS

July 19, 2025

ABSTRACT

1 Introduction

Agent-based models (ABMs) offer a conceptual framework for simulating the actions and interactions of autonomous agents in order to understand the behavior of a system and investigate emergent phenomena. LLMs offer a powerful alternative to traditional ABMs due to their strong text generation capabilities, allowing for more realistic, rich, and detailed simulations. Recent research has discussed the power of LLM simulations due to ‘algorithmic fidelity’ Argyle et al. (2023) — the training data of these LLMs encode biases such as demographic or behavioral biases, so it could be reasonable to believe that these LLMs are able to capture a wide variety of human social reality. However, it is unclear how to evaluate and validate these simulations, particularly since there are persisting questions of algorithmic bias — for which types of people is there sufficient algorithm fidelity? Furthermore, the alignment process of LLMs introduce further distortions, e.g., mechanisms that hinder models from becoming toxic and raise questions about whether these LLMs exaggerate prosocial tendencies in solutions Chang et al. (2025). Establishing a systematic, consistent, and unified framework for validating LLM simulations would allow us to benchmark different LLM backends and quantify the added value of LLM simulations compared to existing ABMs.

Here when we mean LLM simulations or agents, we specifically mean using Large Language Models like ChatGPT to generate agents, which can be individuals or composite units, humans or other non-human entities like companies. These agents are differentiated from other assistive functions of LLMs such as content analysis because these agents themselves are objects of study. Therefore, the use of LLMs for content annotation or assistive agents is out of scope. Instead, we’re interested in LLM agents for simulating hypotheticals. We aim to situate these LLMs in versatile settings; from individual and fully isolated LLM agents simulating survey respondents to a complex network of agents interacting in simulated environments.

While there are some validation recommendations for traditional ABMs Gräbner (2018), evaluating individual LLMs in simulations is a prolific topic of study. However, there is little connection between these two bodies of literature, connecting agent-based social simulations with LLM-based simulations. Popular LLM benchmarks like BigBench and composite ones like HELM attempt to establish LLM competencies across a wide range of applications such as mathematical knowledge to morality. While these capabilities do have some bearing on the validity of LLMs in certain types of simulations, these are only one aspect. There is no holistic framework measuring the many different facets of simulations such as realism or similarity to behaviors displayed by the real entities to be simulated, particularly for a specific simulation context. To illustrate, with HELM, we can know that GPT3.5 has general knowledge skills equal to that of a [ADD]. However, if the same LLM is prompted to take on the persona of a moderately intelligent Reddit user for simulating conversation, would it achieve realistic results, i.e., would the agent display expected behaviors and not default to superhuman capabilities? More importantly, how do we know that the emergent behavior from the interaction of two agents with different intelligence levels mimic real-life behaviors? In this emerging area of Generative ABMs, while there are several studies exploring the potentials of LLM-based simulations, there are several open questions about how to establish the validity of these simulations Larooij and Törnberg (2025). Particularly, while individual papers conduct some type of validation, they are often ad-hoc and have little to do with the final simulation outcome Larooij and Törnberg (2025). An end-to-end unified framework for identifying errors and biases in LLM social simulations remains elusive.

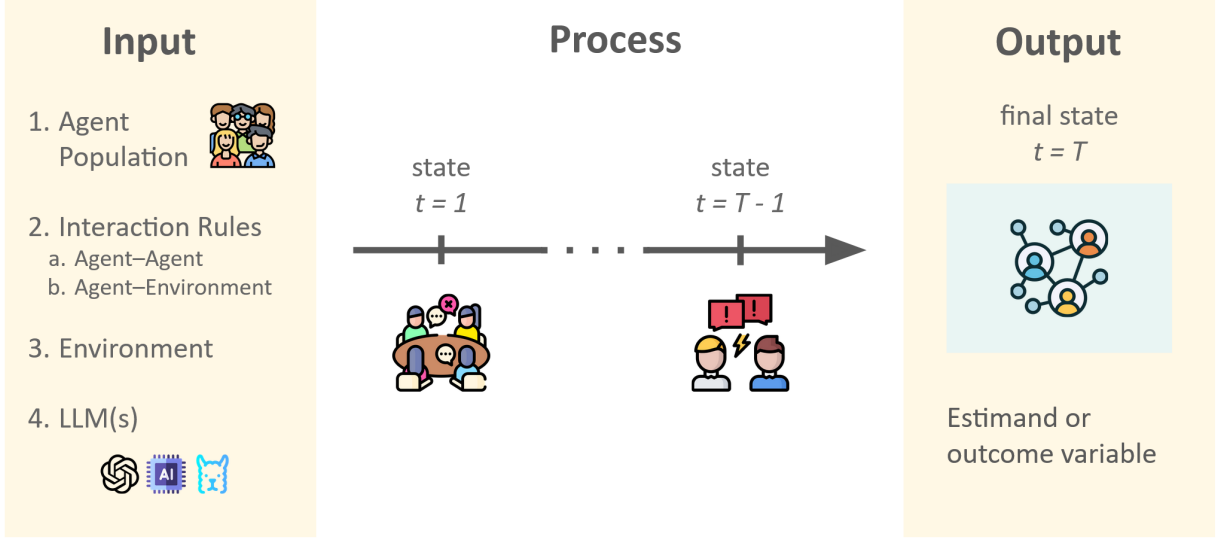


Figure 1: **Summary of Generative LLM Social Simulations and their Components.**

Therefore, in this work, we bridge the literature on validating ABMs and recent research on evaluating LLM capabilities, by specifically creating a validation framework for evaluating LLM simulations. Besides the validity concepts in ABM literature, we also incorporate validity concepts in measurement theory, particularly more recent work aimed at translating these concepts to Computational Social Science studies Sen et al. (2021). We harmonize analogous validity concepts like face validity (measurement theory) and input validity (ABM validation), and enumerate and (re)define concepts particularly salient for LLM simulations. Grounding our framework on several different types of simulation to ensure breadth and comprehensiveness, we present a hypothetical running example and use it to illustrate different types of validation required in these case studies (Section 4).

Our framework provides a blueprint for those interested in designing their own LLM simulations, allowing the comparison of different LLM backends (LLM architecture, alignment types, etc) and recommendations for minimizing threats to validity. Future work can build on our framework to design benchmarks for specific contexts for testing particular types of validity. Finally, we distill our validation framework into a checklist that is designed to help researchers conducting social simulation to reflect on and document their design choices (Section 7.1).

2 Types of Generative Simulations

To systematically categorize LLM-based social simulations, we propose a taxonomy based on three orthogonal dimensions: level of abstraction, interactivity, and progression.

Level of Abstraction. This dimension captures the granularity at which agents and social structures are modeled:

1. **Micro-Level:** Simulations at this level focus on individual agents, each typically instantiated as a separate LLM instance or prompt context. Agents operate with personal goals, beliefs, and behaviors, allowing the study of individual-level behavior, e.g., simulating survey respondents Argyle et al. (2023).
2. **Macro-Level:** These simulations model aggregated societal behavior, often bypassing individual agent reasoning. Here, the LLM may be prompted to simulate complex macro entities such as countries or firms, using high-level descriptors rather than explicit agent-based interactions .
3. **Micro-Meso-Macro:** Simulations that span multiple levels. Micro-level agents interact within meso-level social structures (e.g., institutions, groups), and emergent macro-level phenomena are either modeled or observed. This level is particularly suited for analyzing how individual behavior leads to collective outcomes and constitutes the most complex types of simulations .

Interactivity. This dimension distinguishes simulations based on the presence or absence of interactions between agents or between the agent and the environment:












	micro simulation		micro-meso-macro simulation		macro simulation	
						
	no interaction 	interaction 	no interaction 	interaction 	no interact. 	interact. 
single-step 	Argyle et al., 2023 Bisbee et al., 2024 Survey Responses	Aher et al., 2023 Shock Experiments	Aher et al., 2023 Wisdom of the Crowd			
multi-step 		Park et al., 2023 Generative Agents Ahnert et al., 2025 Persuasive Dialogues		Park et al., 2022 Törnberg et al., 2023 Rossetti et al., 2024 Simulated Social Media Chang et al., 2023 Social Networks Bailis et al., 2024 Werewolf		

Figure 2: **Taxonomy of LLM simulations and example studies.**

1. **Non-Interactive:** Agents or scenarios are modeled in isolation, with no interaction during the simulation run. These simulations often aim to explore hypothetical reasoning, individual response patterns, or single-agent decision-making under fixed contexts. Examples would include the use of LLMs for survey response simulations Argyle et al. (2023); ?.
2. **Interactive:** Agents engage with each other and/or a simulated environment. Interactions can include dialogue, influence, cooperation, or competition, and may be direct (agent-to-agent) or mediated (through the environment). Examples would include LLM-based agents interacting in virtual environment Park et al. (2022).

Progression. This dimension captures the procedural complexity of the simulation:

1. **Single-Step:** The simulation comprises a single round of LLM prompting and output. These are typically used for static analyses or snapshot evaluations, such as a one-time prediction of behavior or a collective response to a single event.
2. **Multi-Step:** Simulations proceed through multiple discrete steps or stages. Agents may form memories, revise beliefs, update goals, or adapt strategies based on prior events, allowing for recursive reasoning and temporally extended narratives.

3 Components of a Generative Simulation

The design of non-generative Agent-based Models (ABMs) can be broken down into three components – input, process, and output (see for example Gräbner (2018)). We adopt the description of these three components to include generative- or LLM elements. A simulated world consists of agents (A^1, \dots, A^N), powered by one or more LLMs, and the environment (E). Interactions between agents are decided by behavioral rules (e.g., constraints) and environmental interaction. Like ABMs, there are three parts of a simulation:

3.1 Input, i.e., the world at $t=0$

This is the initialization stage of a simulation. Here we define an **agent population**, i.e., the personas of all agents involved in the simulation and the description of the **environment** where the agents will interact. We also define rules or heuristics of **interactions** between all agents and between the agents and the environment. If interactions between agents are included, they can be *constrained* – limiting the number of agents a focal agent can interact with – or *unconstrained* – allowing every agent to interact with all other agents.

For an LLM-based simulation, we also define the LLM to be used throughout the simulation as well as its hyperparameters (model architecture, size, temperature, etc.).

3.2 Process, i.e., the world at $t=1, \dots, T-1$

Each simulation consists of several process steps where the simulation progresses. Some simulations can be *one-shot*, i.e., they do not have any clearly defined process steps, e.g., Argyle et al. (2023), or Aher et al.

3.3 Output, i.e., the world at $t=T$

It is at the output stage where we compute the **estimand**, i.e., the outcome of interest, e.g., the rate of information diffusion ?.

4 Quality Criteria for Generative Simulations

To effectively evaluate errors and biases in generative LLM social simulations, we take inspiration from the quantitative social sciences, particularly measurement theory and survey methodology as well as the literature on how to relate Agent-based models to reality Gräbner (2018). Survey methodology makes use of ‘error frameworks’ that help a survey designer systematically assess errors like sampling error or response error in the entire survey lifecycle Groves and Lyberg (2010). Another hallmark of these error frameworks are categorizing errors into **measurement errors** and **representation errors**. Measurement errors occur when incorrectly defining and operationalizing the social construct a researcher is interested in measuring, while representation errors occur when incorrectly defining and operationalizing the target population of interest, i.e., to whom the study’s findings generalize to. The canonical “Total Survey Error” framework Groves and Lyberg (2010) has been extended and adopted to systematically characterize errors in contexts beyond surveys Amaya et al. (2020), such as studies using digital traces and computational methods Sen et al. (2021) and data donations Boeschoten et al. (2020); Bosch and Revilla (2022). We adopt a similar perspective by assessing errors in each component of an LLM simulation (Section 3). Accounting for measurement and representation errors help establish the *validity* of social simulations. Measurement theory speaks of another quality dimension besides validity, i.e., **reliability** which pertains to how repeatable or stable a measurement. Thus reliability forms the third dimension of our quality framework.

To that end, We divide our evaluation framework into three main parts mirroring the components of an LLM-simulation. Each of these three sections describes potential errors of measurements, representation, and reliability, summarized in Figure 3.

Most of a simulation designer’s design choices occur during the input step, while further assessments need to be carried out during the process and output stage. To illustrate each of these choices and the quality criteria associated with them, we use a running example of a hypothetical simulation where we simulate a Reddit-like platform, specifically the r/vegan subreddit, to assess the impact of the introduction of a new moderation rule, with the following background:

Simulation level: micro-macro

Estimand: The change in toxicity level of conversations in the subreddit

Agents: Reddit users and moderators

Interactivity: yes

1. Between agents:

- (a) users can interact by upvoting or downvoting each other’s posts, commenting on them, reporting them for violating the subreddit’s rules
- (b) Moderators can remove posts and ban other users, i.e., delete posts from the environment or delete agents from the environment

2. Between agent-environment: the environment is a subreddit-like platform which hosts posts and comments from agents.

Progression: yes: sequence of content recommendations, posting actions, interactions with posts, and moderation actions in the subreddit

4.1 Input ($t = 0$)

The input component relies entirely on the specification of the initial parameters of the simulation, particularly the inputs given to the LLMs and without assessing the LLMs’ responses. The main inputs for an LLM simulation consist of the agents, the environment, the rules governing possible interactions between the agents as well as the interaction

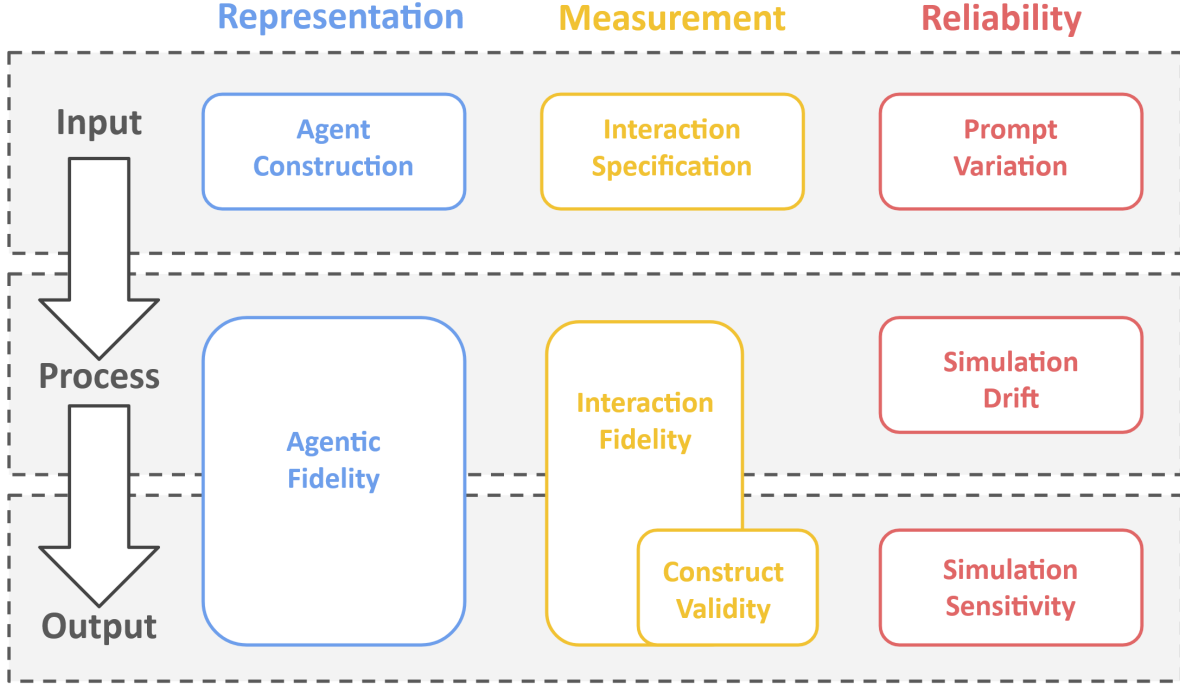


Figure 3: **Quality Framework for Generative LLM simulations.**

between agents and the environment, and the LLMs used to simulate the agents. The input stage has the following processes (not necessarily done sequentially):

Agent Construction. We first start with defining the identity, personality, and roles an LLM agent in our simulation can take. This step consists largely of so-called prompt engineering and designing prompts that lead to agents who are faithful to the simulation but can also include examples of typical agent actions, and even memories the agent has already formed – for example in the form of entries in a vector data base. There is substantial work on persona creation and the optimal ways of designing agents that are representative and realistic Moon et al. (2024), though a large portion of this work focuses on demographic personas. There is less work on how other aspects of identity, e.g., personality or occupation, can be incorporated into personas. Agent construction primarily runs into *errors of representation*, since we have to ensure that the LLM can represent the persona of interest.

For the r/vegan simulation, we create agents using real personas from the actual r/vegan subreddit. We use three different types of persona prompts for both Reddit users and moderators, albeit with different types of information — 1) username only, 2) username and the other subreddit this user is most active in, 3) username, other active subreddits and recent posting, comments, and moderating history.

Interaction Specification. If the simulation is interactive, then we define interaction rules between agents and between the agent and environment. As with the agent prompts, these interaction rules can also be prompted in different ways, e.g., with examples of real-world interactions. At this step, it is important to consider whether the interaction rules are a good proxy of real-world mechanisms as this has important bearings on the generalizability of the findings of the simulation, akin to assessing the generalizability of controlled experiments to real-world settings. Interaction specification runs into *errors of measurement*, since poorly specified interaction rules can lead to invalid measurements of the final estimand.

In our example, agents represent Reddit users and moderators. Therefore, interaction rules in the simulation reflect the interactions that are made possible by the affordances of the platform.

Agent-Environment interaction. Agents can create Reddit-like posts, reply to posts, upvote and downvote posts, as well as report posts. Privileged agents, i.e., moderators can delete posts. Once created, a post becomes part of the environment, as do any replies and votes associated with the post. The environment in turn mimics a

Reddit-like feed with algorithmic ranking of posts and comments. Similar simulations are described by Park et al. (2022) for Reddit, and Törnberg et al. (2023); Nudo et al. (2025); Yang et al. (2024); Rossetti et al. (2024) for X (formerly “Twitter”).

Agent-Agent interaction. Agents interact with each other indirectly through the environment, i.e., they can comment on each other’s posts, upvote or downvote them, or report them. Moderators can also delete posts from the platform (environment).

Parameter Variations. The input parameters of the LLM – specifically the model hyperparameters, such as temperature, as well as the prompt – can influence simulation outcomes. Sclar et al. (2023) show that LLMs have high prompt brittleness. Therefore variations of the prompt and other model hyperparameters might be necessary to assess their influence on simulation results. Ideally prompts should be varied systematically, however, the use of natural language in prompting opens up a vast prompt space that may be difficult to navigate in a principled manner. Nonetheless, the simulation context might help inform specific variations, e.g., using different demographic descriptors for demographic prompting. Prompt engineering approaches like chain-of-thought prompting, self-consistency, and others should be reported at this stage. Parameter variations can help surface *reliability* issues.

For the r/vegan simulation, we design prompt variations based on stylistic and format changes, e.g., QA-style (“username: u/bleedingheart...”) input Santurkar et al. (2023) vs. text-based instructions (“your username is u/bleedingheart...”). For the interaction instructions, we vary the order of the interactions an agent can do. Regarding variations in the other model parameters, we run simulations at different temperatures.

4.2 Process ($t = 1, 2, \dots, T - 1$)

The process component includes different simulations stages or steps (not to be confused with design steps like agent construction) of a simulation. Simulations which are one-step involve only one process step, $t = 1$, which is functionally equivalent to the final stage or the state of the simulation in the output component. Multi-step simulations involve several steps and end with the penultimate state of the simulation. The transition between the states in the process component typically models some type of temporal progression (though spatial progressions might also apply). The most crucial design choice in the process stage is whether interactions happen sequentially or simultaneously. This choice can introduce path dependencies in the simulation and can thus substantially influence simulation outcomes. Since truly simultaneous interaction is oftentimes hard to implement, a randomization scheme for the order of interactions can be used instead. If a simulation is run multiple times with randomized interaction orders, a dependence of simulation results on interaction order can be ruled out. In addition, several assessments of simulation quality can be made at every step, initialized at the first one:

Agentic Fidelity. We should first assess agentic fidelity, i.e., if the LLM has successfully adopted the persona we had provided during the agent construction phase. It is important to assess agentic fidelity independent of overall validity of the final estimand (discussed in more detail in Section ‘Output’), since validity of macroscopic measures might be affected by the (in)validity of microscopic behavior, especially since there is evidence that LLMs poorly represent certain groups Santurkar et al. (2023); Sen et al. (2025). Furthermore, different microscopic configurations might lead to the same macroscopic patterns, weakening the hypothesized link between microscopic mechanisms and macroscopic outcomes. Assessing the fidelity of agents helps to establish the plausibility of this link. As such this step constitutes a first check of face validity. Following from agent construction, issues in agentic fidelity are *errors of representation*.

There is no one single way to establish agentic fidelity. Previous research has relied on comparing the LLM agents’ behavior to human behavior – also called a “Turing test” by Argyle et al. (2023) who compare the LLMs’ answers to survey questions with real humans’ answers. Justifications of agentic fidelity can also be theoretical, e.g., using a representative survey population to create personas Törnberg et al. (2023).

Based on the 35 papers surveyed in Larooij and Törnberg, we note that only a minority of LLM simulation studies conduct any type of agentic fidelity assessment that is not directly tied to the macroscopic outcome of interest. However, agentic fidelity can be especially hard to establish when the LLM personas are based on non-demographic traits, e.g., personality (‘neurotic persona’) or participation in a social movement (‘vegan persona’).

For the r/vegan simulation, we are interested in simulating Reddit users of the r/vegan subreddit. A first ‘sniff test’ for agent fidelity could be provided by checking the responses of LLM agents to a survey that assesses the respondent’s attitude towards a vegan lifestyle. This assumes that all users of the r/vegan subreddit are interested in adopting a vegan lifestyle and would establish that the LLM agents have some understanding

of the population they are supposed to simulate. Alternatively, timelines of posts and comments of the real users active in r/vegan could be sampled and used as activity histories for agents in the simulation input stage. Following a leave-one-out design similar to the design proposed by Nudo et al. (2025), LLM agents could be tasked to regenerate a post or comment by the modeled user given the post’s or comment’s context (e.g., subreddit description or comment thread). The similarity between the original post and the regenerated post could be used as a metric to assess the fidelity of the LLM agent’s posting behavior.

Interaction Fidelity. Related to agentic fidelity, we should also assess if the interactions between agents and the interactions between the agents and the environment are realistic. Assessing this interaction fidelity is important since certain behaviors are harder to simulate for LLMs because of how they have been engineered: specifically, LLMs are subject to safety guardrails, preventing them from producing overly hateful or toxic text Nudo et al. (2025), or text with questionable or even illegal content, such as a manual for building a bomb. Furthermore, instruct-models are fine-tuned for the task of question-answering Ouyang et al. (2022), making them less suitable to simulate other types of behavior. The same principles of validation that apply to agentic fidelity apply to interactions, e.g., the Turing test when human interactions cannot be differentiated from LLM interactions. Concrete examples include comparing the characteristics of the social networks created by LLM agents against that of human social networks Chang et al. (2025), or comparing the self-reported social costs of an interaction to previous studies with human participants (Ahnert et al., 2025). Following interaction specification, issues at this stage can contribute to *errors of measurement*.

For both agentic and interaction fidelity, we recommend conducting multiple tests to establish convergent validity and triangulate. It is also important to assess the interplay between agent personas and interaction. For instance, Chang et al. find that political agents especially show higher than real-world homophily rates when forming social ties.

In our example simulation, we can compare the behavioral traces of our LLM agents with that of real-world behavior on r/vegan. Such a comparison could be manual (humans being asked to differentiate between real and LLM-generated posts and comments), automatic (comparing the semantic, topical, syntactic similarity between human and LLM-generated content), or both. We should specifically keep an eye towards behaviors that are related to our macro estimand, i.e, toxicity levels of conversations and explicitly check for this at every step of the process.

Simulation Drift. Simulation drift refers to a lack of agentic and interaction fidelity over multiple steps of the simulation. Simulation drift does not entail expected changes in the environment and agents’ behavior based on an intervention and it is important to disentangle changes in the LLMs’ behavior due to an intervention or expected changes during the progression or the simulation vs. reduced fidelity of the agents. Common reasons behind simulation drift include the LLM reverting back to its AI agent persona after a few rounds of conversation Choi et al. (2024) or the cognitive load of information accumulated during the previous steps. Simulation drift is a *reliability error*, since it threatens the stability of the simulation.

In our toy simulation, LLM agents portraying Reddit users might start replying like AI agents several steps into the simulation.

4.3 Output ($t = T$)

The final component of a simulation is the output stage, e.g., the final state of the simulation after its last progression step. At this stage, the estimand variable is measured or outcome of interest.

Construct Validity. Construct validity concerns the correct definition and operationalization of the outcome measure. This is particularly important for intangible social constructs which need to be measured through proxies, e.g., the number of slurs used as a proxy for toxicity. Much research has been dedicated to measuring and improving construct validity Jacobs and Wallach (2021). However, this process is highly context-dependent and there are few general tests that are applicable to different types of measurement. This step is essential in LLM simulations but might have little to do with the LLM agents. Researchers can establish the validity of their final measure in different ways, following best practices laid out for quantitative measurement theory Jacobs and Wallach (2021). Threats to construct validity constitute *errors of measurement*.

In our running example, we might use the Perspective API to measure the toxicity of all generated conversation threads on our simulated r/vegan subreddit as has been done in past research on Reddit discussions Xia et al.

(2020); Kumar et al. (2023), including LLM-based simulation studies Törnberg et al. (2023). However, it should be noted that the Perspective API has several drawbacks and validity issues Sap et al. (2019).

Simulation Sensitivity. The last quality concern in the simulation is related to *reliability* where the designer should establish the stability of their findings, especially since LLMs are highly stochastic Bender et al. (2021). Similar to how Machine Learning practitioners are encouraged to report the average of their experimental results across multiple runs along with measures of stability (e.g., variance, confidence intervals, etc.), we also recommend running the simulation with identical parameters end-to-end across several runs and reporting the average estimand and measures of stability for the ensemble of runs. Computing simulation stability is relatively straightforward, though resource and time intensive. Here, the number of simulation runs needed depends on the acceptable level of uncertainty in the outcome measure. It is also not clear how to establish the ideal number of runs, but machine learning experiments can provide inspiration here.

For our toy simulation, we would run the simulation 100 times and report the average Perspective API-based toxicity scores across these runs with standard deviations.

The above-mentioned framework is also distilled into a checklist that simulation designers can use to report the validation steps they implemented for their study. The full checklist can be found in Section 7.1, while filled-out examples are in Section 7.2.

5 Related Work

5.1 Agent-based Models

Components of agent-based models. Agent-based models (ABMs) are computational simulations of systems that are composed of many agents. The general aim of such models is to understand emergent, macro-scale system outcomes that arise from local, micro-scale interactions between agents. They explore the simplest set of behavioral assumptions required to generate a macro pattern of explanatory interest. ABMs provide theoretical leverage where the global patterns of interest are more than the aggregation of individual attributes, but at the same time, the emergent pattern cannot be understood without a bottom-up dynamical model (Macy and Willer, 2002). In other words, ABMs can be understood as theory so well specified that it can be coded as a simulation.

What is conceptualized as agent can be very diverse: for example, ABMs have been successfully used to model swarms of animals (Beni, 2020), the spread of an infectious disease in schools (Lasser et al., 2021), or the emergence of critical traffic scenarios with autonomous vehicles (Hallerbach et al., 2018). Next to these diverse applications, ABMs have also been used to successfully model groups of humans and emergent phenomena within them, such as segregation (Schelling, 1971), opinion polarization (Li and Xiao, 2017), and social contagion (Iacopini et al., 2019).

Agents in these models have individual characteristics and interact with other agents based on a set of interaction rules. For example, in Schelling’s model of segregation (Schelling, 1971), agents have one of a set of two possible attributes (for example, being “blue” or “red”) and exist in a two-dimensional grid-like world. They observe the attributes of the agents in their immediate neighborhood and decide to move to a new, previously unoccupied position in the grid depending on whether a certain number of their neighbors have the same attribute as themselves. Depending on the threshold for the number of neighbors required to have the same attribute, the model shows segregation of agents with different attributes into separated regions on the grid. While this model is obviously a very simplified toy example, it can still be useful to explain segregation in real neighborhoods based on attributes such as race or political orientation. Allowing agents to change their attributes (oftentimes also called “states”), depending on for example on interactions with other agents or time, introduces additional complexity into the model that empowers it to explain for example social contagion (Iacopini et al., 2019).

As already became apparent in the above-mentioned example, an ABM can be more than a collection of agents and interaction rules: usually, it also includes some metric to measure “proximity” between agents which then defines which agents can interact with each other. This metric can be induced by situating the agents in an euclidean world with one, two or three dimensions. Alternatively, it is common to combine ABMs with networks that are not necessarily embedded in an euclidean world but rather define which interactions between agents are possible based on the existence of network edges between these agents. An example is the simulation of interactions of people on a social medium, where the follower-network determines which content is seen by which agent. If the ABM introduces such a notion of proximity and restricts possible agent interactions based on it, we call the interactions “constrained”.

Next to a notion of “proximity”, ABMs oftentimes also include an environment that influences how agents behave or that agents can interact with and even change. An example would be a simulation of the spread of an infection in a society in which different containment measures are implemented at different points in time. One could also imagine that in this simulation, the implementation of containment measures depend on the number of currently infected agents, thus implementing a feedback loop between the dynamics of the spread of the infection and the state of the environment.

5.2 Validating Agent-based models

The main challenge in using computer simulations to study social phenomena is concern about their predictive validity (Bharathy and Silverman, 2010; Windrum et al., 2007; Fagiolo et al., 2006). Here, Gräbner (2018) provides a useful framework for how to relate such computational models to reality along several levels of validity. The first level, input validity, assesses the ability of the model to represent aspects of the real system correctly at time $t = 0$. This is usually achieved by comparing descriptive statistics of the computational model and the real system. For example, if a soccer game is modeled, the computational model should include 11 players and a playing field of adequate size. The second level, process validity, assesses the credibility of the mechanisms that are encoded in the model. As mechanisms are oftentimes not directly observable, this usually requires to make an argument for the plausibility of encoded mechanisms that is grounded in what is known about the real system and its behavior.

The third level, descriptive output validity, assesses whether the output of the model replicates existing observational data of the system. This is oftentimes the highest reachable validation level for research using ABMs, and even this bar is already very high: validating the output of the model against observational data requires that what is measured as model output is compatible with what is observable about the real system. If both model output and observational data are consistent, this is a relatively strong indicator that the simulation can model the system in a useful way. However, this still leaves open the possibility that different model configurations lead to the same output, as micro-specifications of agent behavior in the model can be a sufficient but not the only possible explanation. This is why the assessment of the credibility of the mechanisms that are encoded in the model is crucial to plausibly argue that the encoded mechanisms are indeed responsible for driving the observed behavior. Another pitfall when it comes to descriptive output validation is the danger of overfitting the model to observational data. ABMs often have a number of free parameters that need to be calibrated as they are not directly observable or they have no direct counterpart in the real system. For this calibration, observational data about the system is used. The same data then cannot be used to assess the fit of the model to the observational data, as good consistency between model output and observational data is a natural outcome of the fitting process. Therefore, some observational information about the system needs to be held back and not used to calibrate the model to credibly assess descriptive output validity.

The fourth and last level of validity is provided by the assessment of the predictive power of the model: how well can the model predict future states of the system? And even more extreme: does the model still predict future states of the system if the functioning of a mechanism is changed in the system and the model? Assessing predictive validity requires the acquisition of observational data about the system that is not used to parameterize or train the model but rather held back to be used for validation. Practically, this can be achieved by using time-resolved data that is separated into a subset observed earlier that is used to parameterize the model, and a subset observed later that is held back and used to validate the simulated model. In the best case, the data from the later model includes a change in the system that can be mirrored in the model by adapting the respective mechanism. If model output and observational data are still consistent, this is further evidence that the encoded mechanisms are indeed what also drives system behavior in reality.

Modeling increasingly complex systems oftentimes requires the introduction of more and more parameters and rules to ABMs. As a result, a common criticism of ABMs is that the results are “built into the model” (Waldherr and Wijermans, 2013) rather than emerging from it. This criticism is not completely unwarranted as, particularly in the context of modeling human behavior, it is oftentimes hard to cleanly justify why a particular social or psychological mechanism as driver of behavior is included while another is omitted. Due to a lack of direct observability of these mechanisms, researchers need to resort to observable proxies but even for these, observational data is generally sparse. As a result, the choice of mechanisms and their parametrization can be somewhat arbitrary and the above-mentioned steps for validation cannot be completed. Here, the use of LLMs as “generators of human behavior” could provide way forward. LLMs promise to encode a wide variety of relationships between perception (e.g., inputs, prompts) and actions (e.g., outputs, generated text). To which extent these relationships reflect real human behavior is subject to ongoing research, and ways to assess their validity in the context of ABMs is subject of this contribution. However, compared to the current baseline of hard-coded probabilistic interaction rules in ABMs, it is likely that substantial improvements regarding the realism of the simulated behaviors are possible.

5.3 Social Simulations with Large Language Models

Potential of Social Simulations. Recent work has demonstrated the potential of social simulations to partially automate core aspects of social science research. Bail (2024) envisions generative AI supporting tasks such as hypothesis generation and experimental design, while Manning et al. (2024) and Swanson et al. (2024) go further by implementing automated pipelines where LLM agents simulate study participants and conduct *in silico* experiments. These systems showcase how LLMs can simulate human behavior to explore hypotheses, rediscover known causal relationships, and evaluate interventions.

Social simulations also offer a sandbox for exploring interventions that would be infeasible or unethical to test in the real world. For example, Törnberg et al. (2023) simulate a social media platform to test how different news feed algorithms influence cross-partisan discourse, while Park et al. (2022) develop a tool for testing the impact of alternative community rules and goals. Rossetti et al. (2024) propose creating “digital twins” of online platforms, enabling researchers to simulate user engagement and policy interventions without relying on proprietary social network data.

In addition to their potential for social science research, social simulations have been proposed as benchmarks for evaluating LLM capabilities. Bailis et al. (2024) introduce a dynamic benchmark based on the social game *Werewolf*, in which LLM agents with hidden roles engage in deception and persuasion. As more capable models compete against one another, such simulations offer a moving target for evaluation, unlike static benchmarks.

Drawbacks and Threats to Validity. Despite these promising directions, several recent studies have pointed to important threats to validity in LLM-based social simulations. First, Taubenfeld et al. (2024) find that LLM agents often converge toward generic or default behaviors, even when assigned distinct personas. This undermines attempts to simulate realistic diversity in agent behavior. Second, Barrie and Törnberg (2025) raise concerns that some simulated outcomes may reflect memorized facts or narratives from the training corpus rather than emergent dynamics—especially when simulating historical or well-known social patterns. Third, Larooij and Törnberg (2025) emphasize that the black-box nature of LLMs violates core assumptions of traditional agent-based modeling: unlike ABMs with interpretable rules, LLM agents do not have transparent mechanisms linking micro-level behavior to macro-level outcomes. This opacity makes it difficult to identify what processes the model is actually simulating. Finally, social simulations are computationally expensive. As noted by Bender et al. (2021) and Larooij and Törnberg (2025), individual LLM queries are already costly, and simulations often require many iterations to assess robustness—making large-scale experiments prohibitively resource-intensive.

5.4 Evaluating Large Language Models

Widely adopted benchmarks such as MMLU, BIG-bench, HELM, and AGIEval assess LLM capabilities across domains like reasoning, knowledge, and multitask performance. However, these benchmarks exhibit integrity issues, e.g. data errors and contamination. McIntosh et al. (2024) critique 23 state-of-the-art benchmarks for prompt sensitivity, evaluator diversity, and cultural bias, arguing for more robust and adaptive evaluation frameworks. Human evaluation is still considered the gold standard of LLM evaluation for certain use cases, e.g., . An emergent paradigm—LLM-as-Judge—uses one model to evaluate another’s outputs, facilitating scalability but also raising concerns around evaluator bias and lack of transparency. Evaluating generative LLMs demands fairness and risk assessments across demographic and ethical dimensions. Bouchard (2024) introduces a structured framework (LangFair) mapping model-prompt populations to bias metrics dependent on context ([arXiv][6]). Typical bias tests include stereotype evaluations (e.g., StereoSet, CrowS-Pairs), toxicity assessments, and fact-based hallucination detection using benchmarks like TruthfulQA or SHADES for multilingual fairness.

For LLM-based social simulations, Larooij and Törnberg (2025) survey 35 social simulation papers for the type of validation techniques applied, classifying them into five main categories, including human evaluation of alignment between LLMs and reality, benchmarking against existing social patterns, and comparing against other simulation approaches. They find that a majority of current studies conduct ‘believability’-based validations, testing superficial generations of the LLMs that may have little to do with the final macro estimate.

6 Discussion

Large Language Models (LLMs) offer a novel approach to agent-based simulations, enabling rich, realistic modeling of social behavior through text generation. However questions about how to evaluate the validity of these simulations, especially given inherent biases and alignment artifacts, abound. We propose a unified validation framework that bridges traditional ABM validation methods with measurement theory and recent LLM evaluation research. We define and align key validity concepts, illustrate them through simulation examples, and present a practical checklist to guide researchers in designing and assessing LLM-based simulations.

Designing Benchmarks for GABMs. While there are several general-purpose benchmarks and evaluation suites, e.g., HELM, BigBench, none focus on the specific evaluation criteria needed for simulations. Our framework lays out these criteria, e.g., agentic fidelity, and provides a blueprint for the design of concrete simulation-focused benchmarks for LLMs. Such benchmarks would naturally be polymorphic, i.e., testing several abilities of LLMs such as their ability to simulate a particular group of people through various metrics (e.g., similar opinion generation, similar social media posting behavior) or their ability to adhere to a persona across multiple conversation rounds.

Recommendations and Best Practices. By organizing our framework into three important quality dimensions — measurement, representation, and reliability, and tying them to the different components of generative LLM simulation, we enable researchers to systematically reflect on the design process of their simulation. Certain errors might be outside a researchers control, e.g., guardrails of a commercial LLMs; however, they may still document that so that future research can try and address such errors. Finally, our framework can provide guidance on *which* problems are best solved by generative ABMs and help establish their limits, e.g., to which target population the findings of a potential simulation study would apply to.

More practically, the checklist in Section 7.1 is specifically designed to be used by designers of LLM simulations to improve documentation and reproducibility. As an example, we provide a filled out version of the checklist for the social network simulation study by Törnberg et al. (2023).

Limitations. We acknowledge that our taxonomy is quite coarse; this is to ensure an optimal trade-off between simplicity and richness. Future work can extend our modular taxonomy to include more granular dimensions, e.g., 1) extending interactivity to a spectrum with subdimensions like environment-only, peer-to-peer, group-level, etc., or 2) the cognitive abilities of agents. We also note that our framework is not designed to be comprehensive or complete — such a framework would be difficult to devise in the face of rapidly evolving language model technology. Nonetheless, we hope our framework is *modular* so that new sources of errors can be easily integrated in future.

Acknowledgments

Jana Lasser has received funding from the European Research Council (ERC) under the European Union’s Horizon Europe programme (Grant agreement No. 101160928).

References

- Ahnert, G., Wurth, E., Strohmaier, M., and Mata, J. (2025). Simulating Persuasive Dialogues on Meat Reduction with Generative Agents. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*, Copenhagen, Denmark.
- Amaya, A., Biemer, P. P., and Kinyon, D. (2020). Total error in a big data world: adapting the tse framework to big data. *Journal of Survey Statistics and Methodology*, 8(1):89–119.
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.
- Bailis, S., Friedhoff, J., and Chen, F. (2024). Werewolf Arena: A Case Study in LLM Evaluation via Social Deduction. arXiv:2407.13943 [cs].
- Barrie, C. and Törnberg, P. (2025). Emergent LLM behaviors are observationally equivalent to data leakage. arXiv:2505.23796 [cs].
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Beni, G. (2020). Swarm intelligence. *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, pages 791–818.
- Bharathy, G. K. and Silverman, B. (2010). Validating agent based social systems models. In *Proc Winter Sim Conf*, pages 441–453.
- Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., and Oberski, D. L. (2020). Digital trace data collection through data donation. *arXiv preprint arXiv:2011.09851*.

- Bosch, O. J. and Revilla, M. (2022). When survey science met web tracking: Presenting an error framework for metered data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185:S408–S436.
- Chang, S., Chaszczewicz, A., Wang, E., Josifovska, M., Pierson, E., and Leskovec, J. (2025). Llms generate structurally realistic social networks but overestimate political homophily. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 341–371.
- Choi, J., Hong, Y., Kim, M., and Kim, B. (2024). Does chat change llm’s mind? impact of conversation on psychological states of llms. *arXiv preprint arXiv:2412.00804*.
- Fagiolo, G., Windrum, P., and Moneta, A. (2006). Empirical validation of agent-based models: a critical survey. Technical report, LEM Working Paper Series. <https://web.archive.org/web/20220617211514/https://www.econstor.eu/handle/10419/89466>.
- Gräbner, C. (2018). How to relate models to reality? an epistemological framework for the validation and verification of computational models. *Journal of Artificial Societies and Social Simulation*, 21(3).
- Groves, R. M. and Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly*, 74(5):849–879.
- Hallerbach, S., Xia, Y., Eberle, U., and Koester, F. (2018). Simulation-based identification of critical scenarios for cooperative and automated vehicles. *SAE International Journal of Connected and Automated Vehicles*, 1(2018-01-1066):93–106.
- Iacopini, I., Petri, G., Barrat, A., and Latora, V. (2019). Simplicial models of social contagion. *Nat Commun*, 10:2485.
- Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.
- Kumar, D., Hancock, J., Thomas, K., and Durumeric, Z. (2023). Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference 2023*, pages 2797–2807.
- Larooij, M. and Törnberg, P. (2025). Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations. *arXiv preprint arXiv:2504.03274*.
- Lasser, J., Sorger, J., Richter, L., Thurner, S., Schmid, D., and Klimek, P. (2021). Assessing the impact of sars-cov-2 prevention measures in austrian schools by means of agent-based simulations calibrated to cluster tracing data. *medRxiv*, pages 2021–04.
- Li, J. and Xiao, R. (2017). Agent-Based Modelling Approach for Multidimensional Opinion Polarization in Collective Behaviour. *JASSS*, 20:4.
- Macy, M. W. and Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual review of sociology*, 28(1):143–166.
- Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. Working Paper 32381, National Bureau of Economic Research.
- Moon, S., Abdulhai, M., Kang, M., Suh, J., Soedarmadji, W., Behar, E. K., and Chan, D. M. (2024). Virtual personas for language models via an anthology of backstories. *arXiv preprint arXiv:2407.06576*.
- Nudo, J., Pandolfo, M. E., Loru, E., Samory, M., Cinelli, M., and Quattrociocchi, W. (2025). Generative exaggeration in llm social agents: Consistency, bias, and toxicity. *arXiv preprint arXiv:2507.00657*.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., and Bernstein, M. S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.
- Rossetti, G., Stella, M., Cazabet, R., Abramski, K., Cau, E., Citraro, S., Failla, A., Improta, R., Morini, V., and Pansanella, V. (2024). Y social: an llm-powered social media digital twin. *arXiv preprint arXiv:2408.00818*.
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.
- Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186.

- Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2023). Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.
- Sen, I., Flöck, F., Weller, K., Weiß, B., and Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, 85(S1):399–422.
- Sen, I., Lutz, M., Rogers, E., Garcia, D., and Strohmaier, M. (2025). Missing the margins: A systematic literature review on the demographic representativeness of llms. *ACL*.
- Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., and Zou, J. (2024). The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation.
- Taubenfeld, A., Dover, Y., Reichart, R., and Goldstein, A. (2024). Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, Miami, Florida, USA. Association for Computational Linguistics.
- Törnberg, P., Valeeva, D., Uitermark, J., and Bail, C. (2023). Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.
- Waldherr, A. and Wijermans, N. (2013). Communicating social simulation models to sceptical minds. *Journal of Artificial Societies and Social Simulation*, 16(4):13.
- Windrum, P., Fagiolo, G., and Moneta, A. (2007). Empirical Validation of Agent-Based Models: Alternatives and Prospects. *JASS*, 10:8. <https://web.archive.org/web/20230528061134/https://www.jasss.org/10/2/8.html>.
- Xia, Y., Zhu, H., Lu, T., Zhang, P., and Gu, N. (2020). Exploring antecedents and consequences of toxicity in online discussions: a case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–23.
- Yang, Z., Zhang, Z., Zheng, Z., Jiang, Y., Gan, Z., Wang, Z., Ling, Z., Chen, J., Ma, M., Dong, B., et al. (2024). Oasis: Open agents social interaction simulations on one million agents. *arXiv preprint arXiv:2411.11581*.

7 Appendix

7.1 Checklist

In addition to the input, process, and output components discussed in Section 4, we also include two supplementary sections on Reproducibility and Ethics.

Preliminaries

- **Simulation level:** [micro, macro, micro-macro, micro-meso-macro]
- **Estimand:**
- **Agents:**
- **Interactivity:** [yes, no]
 - Between agents?
 - Between agent-environment?
- **Progression:** [single-step, multi-step]

Input: Agent Construction

- What is the target population the agents are meant to simulate?
- What identity variables and other information make up the personas of the agents?
- Why and how were they chosen?
- Which actions can an agent take (only include actions here that are not interactions between agents or agents and the environment)?

Input: Interaction Rule Specification

- Which interactions can an agent make with other agents and which rules govern these interactions?
- If there is an environment: which interactions can an agent make with the environment and which rules govern these interactions?
- Are interactions with other agents and/or the environment happening simultaneously or sequentially?

Input: Prompt Variation Assessment

- Do you report all the prompts you have used?
- Did you assess the impact of prompt variations in inducing personas?
- Did you assess the impact of prompt variations in prompting agent interactions?

Process: Agentic Fidelity

How do you establish that the LLM agent has successfully simulated the given persona?

Process: Interaction Fidelity

How do you establish that the agent-agent behavior and agent-environment behavior is realistic?

Process: Simulation drift (only if your simulation has multiple steps)

- Did you assess agent persona drift across multiple steps?
- Did you assess interaction drift across multiple steps?

Output: Construct Validity

Construct validity: How is the estimand measured?

Output: Simulation Stability

Did you report the average estimand and variance estimates for multiple runs of the simulation?

LLM-Specific Dimensions

- Which model(s) have been used and why?
- Do you make use of the system prompt or the user prompt?
- Did you report exact model parameters and hyperparameters?
- Did you assess the impact of temperature variation?
- How do you extract and process the LLMs' responses?
- What type of computation infrastructure is used to run the simulation?

7.2 Filled-out Checklists for Existing Papers**Simulating Recommendation Algorithms** Törnberg et al. (2023)**Preliminaries**

- **Simulation level:** Micro–Macro
- **Estimand:** Measures of cross-partisan interaction and toxicity (e.g., reply rates, toxicity scores)
- **Agents:** LLM-based agents based on ANES respondents
- **Interactivity:** Yes
 - Between agents? Yes
 - Agent–environment? Yes
- **Progression:** Multi-step

Input: Agent Construction

- **Target population:** U.S. Twitter users from ANES 2020
- **Identity variables:** Demographics, political affiliation, cultural/personal interests
- **Selection rationale:** To reflect real-world diversity in behavior and social identity; based on survey calibration

Input: Interaction Rule Specification

- **Interaction rules:** Post, share, like, and comment based on personalized feed exposure
- **Agent tasks:** Posting/sharing articles, engaging with content
- **Environment interaction:** Feed determined by algorithmic filtering
- **Environment realism:** Stylized but plausible representation of social media

Input: Prompt Variation Assessment

- **All prompts reported?** Partially — examples in Appendix
- **Persona variation assessed?** No
- **Interaction prompt variation assessed?** No

Process

- **Agentic fidelity:** Survey-calibrated personas; no external validation
- **Interaction fidelity:** Not validated; realism assumed from LLM capability
- **Simulation drift:**
 - Persona drift: Not assessed
 - Interaction drift: Not assessed

Output

- **Construct validity:** Estimands derived using Perspective API.
- **Simulation stability:** Averages reported; no variance estimates provided

LLM-Specific Dimensions

- **Model(s) used:** GPT 3.5
- **System/user prompts used?** Not specified
- **Parameters/hyperparameters reported?** No
- **Temperature sensitivity tested?** No
- **Response extraction:** Not specified
- **Computation infrastructure:** Not reported

7.3 Glossary

Agent-based Model (ABM): The conceptual framework or mathematical/computational model that defines the agents, their behaviors, the environment, and the rules of interaction. It is essentially the design and structure of the system that represents how agents behave and interact. The ABM defines the theoretical setup and logic behind the simulation.

Agent: An autonomous entity with specific behaviors, attributes, and rules for decision-making within a simulation. Agents can represent individuals, groups, or entities. In a traffic simulation, each car or driver is an agent with behaviors like acceleration, braking, or lane-changing.

Environment: The virtual space in which agents operate, including both physical and social elements that influence their behavior. It provides context for the agents’ actions. Some ABMs may not distinctly model environments or allow interactions with it. Example: A city map in a simulation where agents (residents) navigate roads, interact with infrastructure, and make decisions based on their surroundings.

Interaction: The way agents influence each other or the environment through direct or indirect actions. Interactions can be physical, social, or informational. ABMs may have no interaction among the agents, some constrained form of interaction (e.g., due to simulating physical space), or fully unconstrained forms of interaction where all agents can interact with other agents. Example: In a predator-prey simulation, a predator agent interacts with a prey agent by hunting, while prey agents react by fleeing.

Micro Simulation: A simulation approach that focuses on individual-level processes and decisions, often involving detailed, bottom-up modeling of agent behaviors. Example: Modeling each household’s energy consumption in a neighborhood to assess electricity demand.

Macro Simulation: A top-down approach that focuses on system-wide outcomes or aggregated behaviors rather than individual actions, often abstracting individual detail. Macro simulations do not have agents, and therefore no interactions between agents. However, there can be interactions with the environment. Example: Modeling the overall energy demand of a city based on statistical averages of household usage.

Micro-Macro Simulation: A hybrid approach that connects micro-level behaviors of agents with macro-level outcomes. It shows how individual actions aggregate to produce broader system dynamics. Example: A labor market model where individual job-seeking behavior (micro) affects unemployment rates (macro), which in turn influences future job-seeking strategies.

Micro-Meso-Macro Simulation: Similar to micro-macro simulations, however, there is a third level between micro agents and macro-level outcomes — meso structures. Example: modeling team dynamics in an organization where micro agents are individuals.