

Verification-in-Use for LLM-Agent Simulations: Toward Robust Inference under Model Uncertainty

Hirokazu Shirado

Carnegie Mellon University

shirado@cmu.edu

Yuxuan Li

Carnegie Mellon University

yuxuanll@andrew.cmu.edu

Sauvik Das

Carnegie Mellon University

sauvik@cmu.edu

Abstract

Large language model (LLM) agents are increasingly used to simulate human behavior and social interaction in policy-relevant domains. Yet a central methodological question remains unresolved: how should such simulations be evaluated for policy use? A common answer is empirical validation—showing that a simulator reproduces observed human or real-world behavior, then using it to reason about unobserved settings. However, in many policy contexts, this target is both difficult to satisfy and stronger than what practical use requires. We argue for an alternative framework: *verification-in-use*. Rather than asking whether a simulator is globally valid, verification-in-use asks whether it supports a reliable comparison between policy alternatives under model uncertainty. We formalize this perspective as an iterative process in which a plausible model space is constrained through empirical observations, expert judgment, and stakeholder feedback, while robustness is assessed in terms of whether decision-relevant contrasts remain stable enough for downstream use. We further characterize practical failure modes of this process, including incomplete plausible spaces, stopping ambiguity, and persistent non-separability. By shifting the evaluation target from global realism to decision-oriented robustness, this paper offers a framework for using LLM-agent simulations more transparently and responsibly in policy settings.

1 Introduction

Large language model (LLM) agents are increasingly used to simulate human behavior and social interaction in policy-relevant domains such as collective decision-making, public communication, and emergency response. For example, such simulations can help policymakers compare how alternative warning messages, coordination strategies, or institutional responses could shape downstream behavior before deployment. Their appeal lies in their ability to represent open-ended, language-mediated, and context-sensitive interactions that are difficult to capture with more rigid simulation frameworks.

This growing use raises a central methodological question: *how should such simulations be evaluated for policy use?* A common answer is *validation*: showing that a simulator reproduces human or real-world behavior on observed data, then using it to reason about unobserved settings. However, in many policy contexts, this target is both difficult to satisfy and stronger than what practical use requires. A simulator may fail to establish global realism while

still be useful for comparing policy alternatives; conversely, it may appear realistic on selected benchmarks while still supporting unreliable downstream conclusions. In particular, simulation outputs may preserve the direction of differences between policies while exaggerating their magnitude, or may appear robust only because relevant sources of uncertainty were not explored.

We argue that, under model uncertainty, evaluating LLM-agent simulations for policy requires a shift from *validation* to *verification-in-use*. In particular, we propose that many policy applications should be evaluated in terms of *decision validity*: whether a simulation supports a reliable comparison between policy alternatives under plausible assumptions. Because the relevant assumptions, intervention spaces, and uncertainties are often only partially known in advance, this evaluation must proceed iteratively through refinement of the plausible model and intervention space.

Accordingly, this paper focuses on the conceptual and inferential foundations of verification-in-use. Our aim is to clarify what should count as reliable policy use of LLM-agent simulations and to provide a theoretical foundation for future workflows, tools, and case-based applications.

The paper makes three contributions. First, we distinguish validation from verification and characterize two conceptual shifts required for evaluating LLM-agent simulations. Second, we formalize verification-in-use as a framework that targets robust comparison under model uncertainty. Third, we identify practical failure modes that arise from this iterative setting—including distorted plausible spaces, ambiguous stopping, and unresolved comparisons—and propose safeguards for their transparent and responsible use in policy settings.

2 Why Global Model Validity Is the Wrong Target

A common response to the use of LLM-agent simulations is to ask whether they can be validated as faithful models of human behavior. However, three features of LLM-agent simulations make classical model validation especially challenging:

- (1) **Mechanism opacity.** In traditional agent-based models, behavioral rules are explicit and interpretable. In LLM-agent simulations, behavior emerges from a combination of pretrained model parameters, prompting strategies, memory architectures, and decoding procedures, which are not easily interpretable as a coherent causal mechanism.

This paper was prepared for PoliSim@CHI 2026: LLM Agent Simulation for Policy, a workshop at CHI 2026 (CHI Conference on Human Factors in Computing Systems), April 16, 2026, Barcelona, Spain.

- (2) **Large design space.** Seemingly minor changes in prompts, role descriptions, context windows, retrieval mechanisms, or decoding parameters can materially alter agent behavior, creating a large space of behaviorally plausible but distinct simulation specifications.
- (3) **High face validity.** LLM-generated behavior often appears plausible and human-like, increasing the risk that simulations are overtrusted even when their causal or inferential validity is weak.

These properties make *validate-then-simulate* workflows especially problematic, where an LLM is first validated on a narrow observed setting and then generalized to broader or unobserved scenarios [2]. Recent work shows that even strong predictive performance does not guarantee valid downstream inference about human behavior [6]. A model may reproduce observed responses while still failing to support valid conclusions if its outputs are driven by training-data leakage, or if its errors are systematically correlated with variables used in later analyses. As a result, local behavioral agreement is insufficient to justify broader simulation-based inference.

The problem is especially acute in the policy settings we target. In domains such as emergency response, crisis coordination, and rare-event planning, the most decision-relevant scenarios are often precisely those for which human-labeled observations are unavailable, ethically inaccessible, or have never been directly observed [3, 7]. At the same time, dismissing LLM-agent simulations altogether would foreclose an increasingly important class of tools for exploring open-ended, language-mediated, and context-sensitive social processes that are difficult to represent with more rigid modeling approaches.

In this setting, global model validity is likely the wrong epistemic target. The practical objective is often not to establish that a simulator is globally realistic, but to determine whether it supports a reliable comparison between alternatives under model uncertainty.

3 Core Shift: Decision-Oriented Iterative Verification

Given the challenges above, we argue that evaluating LLM-agent simulations for policy use requires two conceptual shifts. First, the target of evaluation should move from *model validity* to *decision validity*: whether a simulation supports a reliable comparison between policy alternatives. Second, evaluation should be understood as an *iterative process*, because the relevant assumptions, uncertainties, and candidate interventions are often only partially known in advance. Verification-in-use arises from combining these two shifts (Table 1).

Shift 1: From model validity to decision validity. Classical validation asks whether a simulator faithfully reproduces real-world behavior across contexts [2]. In many policy settings, however, such a standard is difficult to satisfy and often stronger than what actual use requires. In many applications, the practical objective is not to recover the full behavioral distribution, but to compare candidate interventions [3]. For example, when evaluating whether a new warning strategy reduces public misinterpretation, whether

an evacuation protocol reduces congestion, or whether a communication policy improves coordination relative to current practice, the key question is often not the exact predicted outcome under each condition, but whether one alternative is likely to outperform the other in a way that is robust to residual uncertainty. Here, *decision validity* refers to whether the simulation supports a reliable comparison for the decision at hand.

Shift 2: From static evaluation to iterative refinement. In actual policy settings, the relevant assumptions and intervention space are rarely fully specified in advance [4]. Some uncertainty dimensions can be anticipated from prior evidence or operational experience, but others only become visible through interaction with the simulation itself. Likewise, candidate policies often need to be revised as implausible outputs, omitted dynamics, or missing assumptions are revealed. As a result, evaluation cannot be treated as a one-shot test of a fixed simulator. It must proceed iteratively through refinement of both the plausible model space and the comparison space [8].

Operationally, verification-in-use proceeds through an iterative loop in which simulation is used to explore candidate policy comparisons, discrepancies are identified against empirical observations or stakeholder expectations, and the plausible model and intervention space is revised accordingly (Fig. 1). This process is repeated until the comparison between alternatives becomes sufficiently robust for the intended use or the remaining uncertainty is explicitly characterized.

4 Formalizing Verification-in-Use

We formalize verification-in-use as an inferential framework for comparing policy alternatives under model uncertainty. Rather than asking whether a simulator is globally valid, the objective is to assess whether a comparison between alternatives remains robust across a constrained set of behaviorally *plausible* models—that is, models not ruled out by the empirical observations, behavioral regularities, expert judgment, or stakeholder constraints available at the current stage of refinement.

4.1 Decision-Relevant Target

Let A and B denote two candidate policies. For a given model θ , define the decision-relevant contrast:

$$\Delta^{A,B}(\theta) = \phi(f_{\theta}(X^A)) - \phi(f_{\theta}(X^B)), \quad (1)$$

where X^A and X^B denote the simulation inputs corresponding to Policies A and B , respectively, $f_{\theta}(\cdot)$ denotes the simulation process under specification θ , and $\phi(\cdot)$ denotes a decision-relevant evaluation functional applied to the resulting simulation output. The inferential target is whether the comparison between Policies A and B is sufficiently stable to support downstream use. For example, in an emergency communication setting, A and B may represent two alerting strategies, and $\phi(\cdot)$ may evaluate the simulated rate of public misinterpretation.

Table 1: Two shifts in the evaluation of LLM-agent simulations.

	Static / one-shot evaluation	Iterative refinement
Model validity	Classical validation Evaluate whether the simulator reproduces observed behavior or predicts outcomes accurately.	Iterative simulator improvement Refine the simulator through repeated comparison to data or stakeholder feedback, while still targeting global realism or fidelity.
Decision validity	Static robustness analysis Assess whether a comparison between alternatives is stable across a predefined uncertainty set.	Verification-in-use Iteratively refine the plausible model and intervention space in order to support reliable decision-relevant comparison under uncertainty.

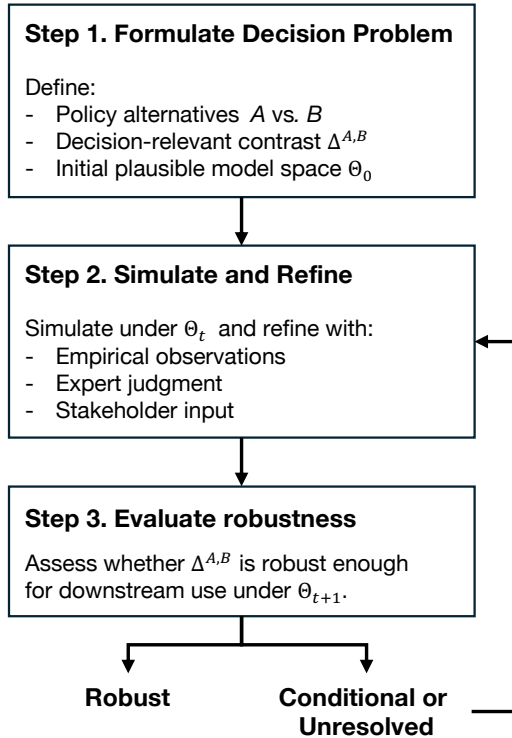


Figure 1: Operational workflow of verification-in-use. The process begins by defining a policy comparison and an initial plausible model space, then iteratively refines that space using simulation outputs and external constraints. If the comparison is sufficiently robust, it can support downstream use; otherwise, the process returns to refinement.

4.2 Plausible Model and Comparison Space

Rather than identifying a single true simulator, verification-in-use operates over a set of plausible models:

$$\Theta_t = \{\theta : \theta \text{ satisfies the constraints available at iteration } t\}. \quad (2)$$

Each $\theta \in \Theta_t$ represents one behaviorally plausible specification of the simulation process at iteration t . Crucially, Θ_t is not assumed to be fully specified in advance and may evolve through iteration

as new discrepancies, missing assumptions, or overlooked intervention dimensions are revealed.

4.3 Constraint-Directed Iteration

At each iteration t , the current plausible model set Θ_t is confronted with empirical observations, known behavioral regularities, expert judgment, or stakeholder feedback. These constraints are applied through a shared evaluation context in which both human and LLM-generated outcomes are observed for comparable inputs:

$$D_{\text{shared}} = (X, Y, f_{\theta}(X)), \quad (3)$$

where Y denotes the human outcome for input X . Beyond this shared anchor, the simulator can also be evaluated on a broader set of policy-relevant or counterfactual inputs for which only LLM-generated outcomes are available:

$$D_{\text{LLM},t} = (\tilde{X}, f_{\theta}(\tilde{X})). \quad (4)$$

Operationally, iteration refines the plausible model set and the broader surrogate simulation space as follows:

$$\Theta_t \rightarrow D_{\text{shared}} \rightarrow (\Theta_{t+1}, D_{\text{LLM},t+1}), \quad (5)$$

where t indexes the current stage of iterative refinement. In this framework, the comparison $\Delta^{A,B}(\theta)$ is evaluated using the broader surrogate dataset $D_{\text{LLM},t+1}$, while the shared evaluation context D_{shared} determines which model behaviors remain plausible.

Iteration therefore operates by revising which model specifications remain plausible and by regenerating the broader surrogate space used to assess $\Delta^{A,B}(\theta)$ beyond directly observed human cases. This process contracts the plausible set by eliminating implausible models, but it may also expand or restructure the space when discrepancies reveal omitted decision-relevant dimensions. In practice, the shared anchor may be closer to the status quo or previously observed conditions (e.g., Policy B), while this broader surrogate space is used to assess less-observed or newly proposed alternatives (e.g., Policy A). The purpose of iteration is not to converge on a single fully valid simulator, but to progressively constrain uncertainty in ways that matter for the comparison of interest.

4.4 Robustness Criterion

The inferential goal is to determine whether the comparison between alternatives is stable across Θ_t . In some cases, it may be sufficient that the direction of the comparison is invariant across

plausible assumptions. However, for many policy uses of LLM-agent simulations, this is too weak: simulations may preserve the direction of differences while exaggerating their magnitude [5]. A stronger criterion is therefore:

$$\inf_{\theta \in \Theta_t} \Delta^{A,B}(\theta) > \delta, \quad (6)$$

for some threshold $\delta > 0$ representing a decision-relevant minimum effect. Under this view, verification-in-use does not attempt to certify a simulator as globally valid. Instead, it evaluates whether the comparison between alternatives is robust enough, across the current plausible space, for the intended inferential or policy use.

5 Conditions for Reliable Verification-in-Use

Verification-in-use does not guarantee robust decision support in all settings, nor is it intended as a procedure for improving simulator fidelity. Instead, it provides a framework for assessing whether a comparison between policy alternatives remains stable across the constrained model set Θ_t .

Its reliability therefore depends on whether iterative refinement meaningfully reduces uncertainty in the policy comparison of interest and whether the remaining uncertainty is small enough to support a stable choice between alternatives. In this section, we characterize two conditions under which verification-in-use can support reliable policy comparison: (i) *decision-relevant learning*, and (ii) *decision separability*. These are not guarantees of correctness, but conditions under which iterative verification becomes decision-useful.

5.1 Condition 1: Decision-Relevant Learning

Let Θ_t denote the current set of plausible models, and let Θ_{t+1} denote the updated set after incorporating new observations or constraints. Iteration is informative if it reduces variation in the decision-relevant contrast:

$$\text{Var}_{\theta \in \Theta_{t+1}}(\Delta^{A,B}(\theta)) < \text{Var}_{\theta \in \Theta_t}(\Delta^{A,B}(\theta)). \quad (7)$$

In policy terms, refinement should reduce uncertainty in which intervention appears preferable, not merely alter the simulator or improve its apparent realism. Effective verification must therefore eliminate or revise models that differ in their implications for $\Delta^{A,B}(\theta)$, rather than only those that differ in behaviorally irrelevant ways. In practice, this condition requires three things.

Coverage of decision-relevant regularities. Constraints must target the aspects of behavior and context that actually affect the comparison between policy alternatives. This does not require full identification of the underlying mechanism, but it does require that the behavioral and environmental regularities driving $\Delta^{A,B}(\theta)$ are meaningfully constrained. Otherwise, models may agree on observed behavior while still diverging in the unobserved dynamics that determine which intervention performs better.

Error learnability. Discrepancies between simulation outputs and empirical observations must provide information that can be used

to revise the plausible set in a decision-relevant way. If observed mismatches do not indicate how models differ in their implications for $\Delta^{A,B}(\theta)$, then refinement may change the simulator without improving the reliability of the policy comparison.

Constraint alignment. Constraints must not only target relevant aspects of behavior, but do so in a way that differentiates models with respect to the decision-relevant contrast. A constraint may improve apparent realism while leaving variation in $\Delta^{A,B}(\theta)$ unchanged. Informative iteration therefore requires that constraints align with the dimensions along which plausible models disagree about the policy choice at hand.

5.2 Condition 2: Decision Separability

Define the range of policy differences over the plausible set:

$$\mathcal{I}_t = \left[\inf_{\theta \in \Theta_t} \Delta^{A,B}(\theta), \sup_{\theta \in \Theta_t} \Delta^{A,B}(\theta) \right]. \quad (8)$$

A necessary condition for robust policy comparison is that this interval does not cross zero:

$$0 \notin \mathcal{I}_t. \quad (9)$$

If \mathcal{I}_t contains zero, then there exist plausible models under which each alternative is preferred, and no stable ranking can be justified under the current evidence. In policy terms, this means the simulation may still be useful for narrowing uncertainty, but it does not yet support choosing one intervention over another with confidence. For example, if some plausible models imply that a revised emergency alert reduces misinterpretation while others imply that it increases confusion, the simulation does not yet support adopting the new policy.

6 Failure Modes and Practical Safeguards

The previous section characterized conditions under which iterative refinement can appear informative and decision-useful. But even when these conditions are met locally, the broader inferential process may still be poorly behaved. In particular, verification-in-use may converge prematurely, narrow the plausible space in path-dependent ways, or yield apparently stable but ultimately suboptimal conclusions.

These are not failures of local fit alone, but of the broader iterative inferential process. To characterize them, it is useful to contrast verification-in-use with standard Bayesian inference [1], where a fixed hypothesis space, likelihood-based updating, and posterior representation provide structural discipline for learning and decision-making. Verification-in-use does not adopt these commitments directly. Instead, the plausible model space Θ_t may itself be revised iteratively as new discrepancies, constraints, and decision-relevant uncertainties are revealed. This flexibility makes the framework better suited to real-world policy settings in which the relevant assumptions and interventions are only partially known in advance [4]. At the same time, it opens the door to distinct higher-order vulnerabilities in how the plausible space is constructed, how

iteration is governed, and how unresolved uncertainty is interpreted. We highlight three practical failure modes and corresponding safeguards.

6.1 Incomplete or Distorted Plausible Space

Verification-in-use can fail if the retained plausible model set Θ_t is too narrow, systematically distorted, or insufficiently explored. Because Θ_t is not fixed in advance and is constructed through iterative refinement, important regions of the space may never be considered. In such cases, the comparison $\Delta^{A,B}(\theta)$ may appear stable only because relevant sources of variation have been excluded.

A particularly important form of distortion is *representational flattening*: the plausible space may over-represent dominant or average behavioral patterns while under-representing subgroup heterogeneity or context-specific variation [9]. In such cases, simulation outputs may appear realistic while smoothing over the differences that determine whether an intervention succeeds or fails in practice.

Practical safeguards. This risk can be mitigated by explicitly managing the exploration of Θ_t . In practice, this includes: (i) tracking the diversity of retained assumptions, (ii) periodically re-evaluating $\Delta^{A,B}(\theta)$ under deliberately varied or previously discarded specifications, and (iii) allocating part of the iteration budget to structured exploration rather than only local refinement. In policy settings, incorporating diverse stakeholders in the refinement process can help surface missing assumptions and underrepresented perspectives. Researchers should also document which assumptions materially affect $\Delta^{A,B}(\theta)$, making explicit where conclusions depend on particular regions of the plausible space.

6.2 Stopping Ambiguity

Verification-in-use can also fail because the update from Θ_t to Θ_{t+1} does not, by itself, specify when refinement is sufficient for downstream use. Unlike standard Bayesian inference over a fixed model space, verification-in-use offers no canonical stopping rule. As a result, iteration may end because a comparison appears favorable, resources are exhausted, or no new constraints are readily available, rather than because uncertainty in $\Delta^{A,B}(\theta)$ has been sufficiently reduced. In policy settings, this creates a particular risk of treating a provisional simulation result as decision-ready.

Practical safeguards. Stopping decisions should be tied to explicit and reportable criteria appropriate to the decision context. One practical safeguard is to adopt a *preregistration-like* approach to refinement: before or during iteration, researchers should document what kinds of discrepancies, robustness patterns, or uncertainty reductions would count as sufficient for pausing, continuing, or escalating the process. In practice, this may include a predefined refinement budget, anticipated stopping criteria such as stability in the sign or lower bound of $\Delta^{A,B}(\theta)$ across recent refinements, and explicit reporting of the remaining uncertainty at the current decision point. Findings should therefore be presented as conditional on the explored space, stopping rule, and current operational context.

6.3 Persistent Non-Separability

Even after meaningful refinement and broad exploration, the comparison $\Delta^{A,B}(\theta)$ may remain non-separable across Θ_t . That is, there may exist plausible models $\theta \in \Theta_t$ under which each alternative is preferred. In such cases, no globally stable ranking can be justified under the current evidence.

This should not always be interpreted as a failure of the process. In policy applications, persistent non-separability may instead reflect meaningful context dependence or heterogeneity in effects across plausible conditions.

Practical safeguards. Persistent non-separability should be treated as a substantive result rather than a defect. In practice, this involves identifying which assumptions or contextual dimensions drive variation in $\Delta^{A,B}(\theta)$ and, where possible, reformulating the decision problem conditionally (e.g., recommending different alternatives under different conditions). When such structure cannot be identified, the appropriate outcome is to explicitly report the comparison as unresolved and use this to guide further data collection, model refinement, or redesign of the intervention space. System interfaces can support this by making retained and discarded assumptions inspectable to stakeholders.

7 Conclusion

We proposed *verification-in-use* as an alternative. Rather than asking whether a simulator is universally valid, this framework evaluates whether it supports reliable comparison between policy alternatives under uncertainty—for example, whether one intervention is likely to outperform another in ways that are robust to modeling assumptions. This shifts evaluation toward decision validity and treats simulation as part of an iterative process that combines model exploration, empirical constraints, and stakeholder input.

For policy applications, this perspective also makes key methodological risks more explicit. Because verification-in-use relaxes standard Bayesian assumptions, its reliability depends on how the plausible space is constructed, how iteration is stopped, and how unresolved uncertainty is interpreted. We highlighted three practical failure modes and corresponding safeguards to support more transparent and accountable use.

We hope this framework helps move LLM-agent simulation for policy away from one-shot realism claims and toward more transparent, iterative, and decision-oriented forms of inference. For HCI researchers and system builders, this reframes LLM-agent simulations as interactive, revisable decision-support tools whose usefulness depends on how uncertainty is surfaced, constrained, and communicated in practice. More broadly, we see this paper as a theoretical foundation for future implementation work: developing concrete workflows, interfaces, and reporting norms that support verification-in-use in practice, while accumulating domain-specific knowledge about what kinds of constraints and refinement processes enable robust inference with LLM-agent simulations.

References

- [1] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.
- [2] Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. This human study did not involve human subjects: Validating llm simulations as behavioral evidence. *arXiv preprint arXiv:2602.15785*, 2026.
- [3] Robert Lempert. Agent-based modeling as organizational and public policy simulators. *Proc. Natl. Acad. Sci. U. S. A.*, 99 Suppl 3(Suppl 3):7195–7196, May 2002.
- [4] Yuxuan Li, Sauvik Das, and Hirokazu Shirado. What makes llm agent simulations useful for policy? insights from an iterative design engagement in emergency preparedness. *arXiv preprint arXiv:2509.21868*, 2025.
- [5] Yuxuan Li, Hirokazu Shirado, and Sauvik Das. 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*, pages 3303–3325, 2025.
- [6] Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large language models: An applied econometric framework. Technical report, National Bureau of Economic Research, 2025.
- [7] Vincent Awj Marchau, Warren E Walker, Pieter Jtm Bloemen, and Steven W Popper. *Decision making under deep uncertainty: from theory to practice*. Springer Nature, 2019.
- [8] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118, 1955.
- [9] Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025.