Position: LLM Social-Simulation Agents in the Wild Cannot Serve as Social Scientific Evidence Without an Identification Strategy

Anonymous Author(s)
Affiliation
Address
email

Abstract

This position paper argues that LLM social simulations cannot substitute for human-subject evidence without an identification strategy. Recent work shows high predictive fidelity — GPT-4 simulations correlate strongly with human treatment effects across hundreds of contrasts, including a post-training-cutoff subset — while critical work shows synthetic respondents fail regression, prompt-sensitivity, and temporal-stability tests. These findings are not contradictory: prediction asks whether outputs match observed outcomes, while substitution asks whether the simulation identifies the social data-generating process. We formalize the problem as observational equivalence among three mechanisms: training-prior retrieval, prompt-induced role compliance, and genuine interactional emergence. The same outcome distribution can be rationalized by multiple combinations of these mechanisms, so predictive fit alone cannot identify emergence. We audit the principal LLM social-simulation literature through this lens, concede the strongest predictive-fidelity result, and show why it cannot license replacing respondents or experiments. NeurIPS should require an identification standard: simulations may generate hypotheses or forecasts, but they become evidence only when their identifying assumptions are explicit, testable, and stress-tested. Positioned for AIWILD, the paper treats LLM social-simulation pipelines as agentic AI deployed in the wild of social-scientific decision-making, where reliability and trustworthiness require not only predictive fit but identification of the agentic data-generating process; without it, agentic outputs masquerade as human-subject evidence and propagate into policy decisions through a failure mode that is invisible to standard agent benchmarks.

1 Introduction

A striking development of the post-GPT-4 period has been the emergence of a research program that treats large language models as substitutes for human subjects in social science. Three papers anchor this program. Park, O'Brien, Cai, Morris, Liang, and Bernstein (2023), published at UIST with accompanying preprint arXiv:2304.03442, introduced "generative agents" — LLM-driven software agents populating a 25-agent sandbox environment — and argued that observed behaviors including invitation-spreading, relationship-forming, and party-coordination constitute "believable simulacra of human behavior." Horton (2023), released as NBER Working Paper 31122 and developed in subsequent work with Filippas and Manning (2024) at the ACM Conference on Economics and Computation, introduced "Homo Silicus" and used GPT to replicate results from Charness and Rabin (2002), Kahneman, Knetsch, and Thaler (1986), and Samuelson and Zeckhauser (1988), reporting

Submitted to ICML 2026 Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD). Do not distribute.

qualitative agreement with the human findings. Argyle, Busby, Fulda, Gubler, Rytting, and Wingate (2023), in *Political Analysis* 31(3): 337–351, introduced "silicon sampling" — conditioning an LLM on demographic backstories drawn from real survey respondents to generate "silicon samples" that track ANES vote-choice distributions.

In the years since, the literature has grown considerably. Aher, Arriaga, and Kalai (2023), published in the *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202: 337–371), introduced "Turing Experiments" and replicated Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and Wisdom of Crowds findings. Hewitt, Ashokkumar, Ghezae, and Willer (2024, working paper dated August 8, 2024), widely circulated in subsequent simulation work, report that across 70 pre-registered survey experiments comprising 476 treatment effects and 105,165 human participants, GPT-4-derived predictions correlate with actual treatment effects at $r = 0.85$, with $r = 0.90$ on studies published after the model's training cutoff. Anthis, Liu, Richardson, Kozlowski, Koch, Evans, Brynjolfsson, and Bernstein (2025), published as a Position Paper in the *Proceedings of the 42nd International Conference on Machine Learning* (PMLR 267, arXiv:2504.02234), argue that "LLM social simulations are a promising research method." As these results travel beyond methodological papers, the evidentiary question becomes sharper: when does simulation become evidence rather than exploration?

A parallel critical literature exists. Bisbee, Clinton, Dorff, Kenkel, and Larson (2024), in *Political Analysis* 32(4): 401–416, find that regression results from synthetic samples differ systematically from human baselines, that minor prompt rewordings shift response distributions, and that the same model drifts measurably over three months. Boelaert, Coavoux, Ollion, Petev, and Präg (2025) in *Sociological Methods & Research* 54(3): 1156–1196 document what they term "machine bias" in LLM responses to opinion polls — a strong bias with low variance that varies randomly across topics — and conclude that models cannot currently substitute for human survey respondents. Tjuatja, Chen, Wu, Talwalkar, and Neubig (2024), arXiv:2311.04076, test multiple models against documented human response biases in survey design and find that especially instruction fine-tuned models do not consistently reproduce the human-like response biases that would be required for survey substitution to be valid. Santurkar, Durmus, Ladhak, Lee, Liang, and Hashimoto (2023) show that aligned LLMs disproportionately reflect the opinions of a specific demographic subset.

The field has therefore bifurcated into a pro-simulation wing that emphasizes fidelity results and a critical wing that emphasizes systematic failures, and the two wings have not engaged each other at the level of methodological foundations. We argue that the resolution requires importing a distinction from econometrics: the distinction between *prediction* and *identification*. A model can be highly predictive without identifying any causal mechanism; conversely, identification is possible with poor prediction. The pro-simulation wing has established prediction fidelity in certain cases. It has not systematically established identification. And the question of whether LLM simulations can substitute for human subjects in *causal* social science — which is what the policy uses of this research presuppose — turns on identification, not prediction.

Our position, stated precisely: **LLM social simulation outputs observed in the published literature are the joint product of three mechanisms — retrieval from training priors, prompt-induced role-script compliance, and genuine interactional emergence — which are observationally equivalent under the evidence currently reported, and in the principal papers we audit, no design jointly separates these three mechanisms through the identification strategy required to distinguish them. Until such identification is provided, LLM simulation results should be treated as hypotheses or forecasts, not as substitutes for human-subject causal evidence.** Prediction validates a mapping; identification validates a data-generating claim. We call on NeurIPS to adopt identification-strategy disclosure as a condition for publication of simulation-based social-science claims.

Positioning within AIWILD. The AIWILD workshop frames the safety, security, and reliability of agentic AI systems deployed in open-ended real-world settings. LLM social simulation is one such deployment: 25-agent generative-agent sandboxes, multi-agent "Homo Silicus" economic environments, and silicon-sample survey panels are all multi-agent LLM systems acting in the wild of social-scientific decision-making, with downstream consequences that include policy briefs, regulatory submissions, and empirical claims published in venues whose readers cannot inspect the simulation pipeline. Standard agent benchmarks measure task success on bounded tasks; they do not measure whether an agentic output qualifies as evidence about a real social data-generating process.

The failure mode we identify is therefore an AIWILD-class failure: a class of agent deployments in which the agents pass standard fidelity checks (Hewitt et al. $r = 0.85$ across 476 effects) yet remain unsafe in the specific sense that their outputs cannot be distinguished from training-prior retrieval or prompt-script compliance, so any downstream policy decision built on them is built on an unobserved mixture of three mechanisms with different reliability properties. The identification standard we propose is the analogous safety standard for this class of agentic deployments: a disclosure regime that lets a reader determine, before relying on an agent-generated empirical claim, which mechanism the claim identifies and which it does not.

2 The Landscape of LLM Social Simulation

The pro-simulation literature has developed along four partially overlapping paths.

Generative agent sandboxes. Park et al. (2023) is the foundational paper: 25 LLM-driven agents in a small-town environment coordinating a Valentine's Day party, forming friendships, observing-planning-acting over simulated time. The framing is that emergent behaviors constitute evidence that LLMs can produce believable social dynamics. The methodological move is ethnographic: the paper reports what the agents did, illustrated with natural-language transcripts. A follow-on literature has extended sandboxes to financial markets, platform governance, political polarization, and organizational decision-making.

Silicon-sampling survey replications. Argyle et al. (2023) is the anchor: GPT-3 conditioned on demographic backstories from ANES respondents, producing simulated samples that track ANES marginal distributions and patterns of association — "algorithmic fidelity." Qu and Wang (2024, *Humanities and Social Sciences Communications* 11:1095) extend this cross-nationally using World Values Survey data and find substantial degradation outside Western, English-speaking, developed contexts.

Economic and psychological experiment replications. Horton (2023) and Aher, Arriaga, Kalai (2023) are the anchor papers. Horton replicates dictator-game and status-quo-bias experiments; Aher et al. replicate Ultimatum, Garden Path, Milgram, and Wisdom of Crowds. Both report qualitative agreement in most cases. *Aher et al. make a methodologically honest move that is underweighted in subsequent citation:* they document a "hyper-accuracy distortion" in the Wisdom of Crowds replication, where the LLM simulation is too accurate relative to human crowds — an outcome inconsistent with true simulation of heterogeneous cognitive biases and diagnostic of retrieval from aggregated ground-truth sources in training data. This single result, within a pro-simulation paper, is a direct instance of the identification failure we describe.

Predictive treatment-effect simulation. Hewitt, Ashokkumar, Ghezae, and Willer (2024) assemble an archive of 70 pre-registered nationally representative U.S. survey experiments (476 treatment effects, 105,165 participants) and prompt GPT-4 to simulate responses. The headline result: $r = 0.85$ correlation between simulation-derived and actual treatment effects, $r = 0.90$ for studies published after GPT-4's training cutoff. Anthis et al. (2025) at ICML 2025 cite Hewitt et al. as central validation for the "promising research method" framing.

The critical wing. Bisbee et al. (2024) *Political Analysis* 32(4): 401–416 is the most methodologically sophisticated critique. Using ChatGPT 3.5 Turbo, they elicit feeling-thermometer scores for 11 sociopolitical groups under various persona conditions and compare to ANES 2016–2020 baselines. They find: marginal means are close but regression coefficients differ substantially; distributions are compressed relative to human data, yielding systematically higher statistical power calculations than warranted; response distributions shift with minor prompt wording changes; and the "same" ChatGPT 3.5 Turbo drifts measurably over a three-month period. They replicate portions on ChatGPT 4.0 and Falcon-40B-Instruct, confirming that the issues are not model-specific. Boelaert et al. (2025) document systematic political-ideological bias in generative AI responses to opinion polls. Santurkar et al. (2023) at ICML establish that aligned LLMs disproportionately represent specific demographic subsets. Hartmann, Schwenzow, and Witte (2023) document left-libertarian ideological skew in ChatGPT. Tjuatja et al. (2024) show that LLMs do not reproduce human-like response biases in survey design. The critical wing has collectively established prompt sensitivity, temporal instability, compressed variance, systematic ideological skew, cross-cultural degradation, and failure to reproduce human response biases. What this wing has not yet done is frame these findings under a unifying theoretical apparatus.

3 The Identification Problem

We formalize the distinction between prediction and identification in a simple decomposition. Let Y_i be the observed behavior of simulated agent i — a response, action, or interaction outcome. Let $X_i = (R_i, C_i, H_i)$ be the prompt configuration, where R_i is the role label ("You are a 45-year-old conservative voter…"), C_i is the context (task description, environmental details), and H_i is the interaction history (prior turns, other agents' actions). The observed simulation output can be schematically written as:

$$Y_i = f_{\text{train}}(X_i) + g_{\text{prompt}}(R_i, C_i) + h_{\text{emerge}}(X_{-i}, H_i) + \varepsilon_i$$

where:

- f_{train}: contribution of training-data priors — the model's internalized statistical regularities from pretraining and alignment data, selectively activated by the prompt;
- g_{prompt}: contribution of prompt-induced role compliance — the model's instruction-following response to the explicit role label and task framing;
- h_{emerge}: contribution of genuine interactional emergence — patterns that arise from the multi-agent interaction dynamics and are not reducible to prior retrieval or prompt compliance;
- ε_i: irreducible stochasticity from sampling temperature.

The additive form is heuristic: in practice the three mechanisms interact non-linearly through the model's attention dynamics, and the decomposition should be read as identifying three sources of variance whose contributions are not separately observable from the output alone, not as a linear-regression specification.

Observational Equivalence Proposition. *For any observed distribution of simulated outcomes $P(Y \mid X, R, C)$, multiple triples $(f_{train}, g_{prompt}, h_{emerge})$ can rationalize the same distribution. Therefore, predictive fit alone does not identify the variance share attributable to emergence.* The proposition is not surprising in econometric terms — it is the standard non-identification result for additively-separable systems without exclusion restrictions — but it is what most pro-simulation evaluations elide. The proposition implies that any claim of the form "the simulation produced a result statistically close to the human baseline, therefore it has captured human-like reasoning" is non sequitur unless an exclusion restriction is provided.

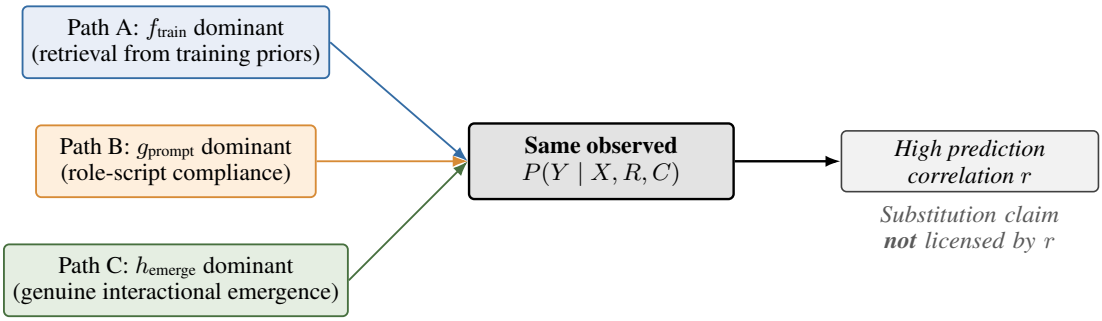

Figure 1: Observational equivalence: three distinct data-generating paths can produce the same observed distribution of simulated outcomes, hence the same prediction correlation. Substitution validity requires identifying which path generated the output; predictive fit alone cannot.

Two observations are foundational. First, **each of the three terms can, in principle, produce outputs highly correlated with human baselines**. Training-data priors encode documented human regularities (the Ultimatum Game offer distribution, the Milgram obedience rate, the approximate direction of partisan voting patterns) because these regularities were themselves documented in text that became training data. Prompt-induced compliance can produce human-like outputs when the prompt instructs "behave like a typical American voter." Genuine emergence can also produce human-like outputs, by assumption. The three mechanisms are observationally compatible with similar Y_i distributions.

Second, **the researcher does not directly observe the decomposition**. The researcher observes Y_i and X_i but not f_{train}, g_{prompt}, or h_{emerge} individually. Training is a black-box process whose output is only indirectly accessible. The prompt's causal pathway to behavior runs through the entire trained weights; there is no way to turn off f_{train} while leaving g_{prompt} operative.

This is the structural feature that renders the identification problem irreducible absent additional experimental design. Prediction fidelity (high correlation between simulated and human Y_i) is compatible with any apportionment of variance among f_{train}, g_{prompt}, h_{emerge}. And the three apportionments carry drastically different implications:

- If the observed fidelity is primarily explained by f_{train}, the simulation is functioning as a retrieval tool; its outputs are no more informative about new human behavior than a well-indexed database of previous social-science findings.
- If the observed fidelity is primarily explained by g_{prompt}, the simulation is an instruction-following artifact; its outputs may reflect the researcher's implicit theory encoded in the prompt rather than a property of human behavior.
- Only if the observed fidelity is primarily explained by h_{emerge} does the simulation generate genuinely new social-structural information. This is the case the pro-simulation literature needs, but has not yet established.

The analogy to econometric identification is direct. An OLS regression of Y on X produces a coefficient that is the joint product of causal effect, selection bias, and omitted-variable bias; the three components cannot be separated by examining the coefficient alone. The resolution in the credibility-revolution literature (Angrist and Pischke 2009) was methodological — instrumental variables, regression discontinuity, and difference-in-differences designs, each providing an exclusion restriction. The LLM simulation literature is in the pre-credibility-revolution state: researchers report prediction correlations without exclusion restrictions. The field needs its own identification apparatus.

Each of the three mechanisms admits a specific empirical probe. Probing f_{train} (training retrieval): compare simulation outputs on tasks whose answers are in training data versus tasks composed after the model's training cutoff; differential performance localizes the contribution of retrieval. Aher et al. (2023) inadvertently ran this test, with their Wisdom-of-Crowds result diagnostic of retrieval dominance. Probing g_{prompt} (prompt compliance): compare simulation outputs under systematically varied prompts that preserve the substantive task but alter the role-label or framing. Bisbee et al. (2024) run this test, finding that "minor changes in prompt wording" shift response distributions; the variance attributable to prompt wording is a lower bound on g_{prompt}. Probing h_{emerge} (genuine emergence): compare simulation outputs to a placebo condition in which the interaction structure is removed (all agents made identical, or interactions randomized). If core metrics collapse under the placebo, apparent "emergence" is not emergence; if they persist, the emergence claim is at least consistent with data. Among the principal papers we audit, this test is not run systematically.

4 A Mechanism-by-Mechanism Audit

We illustrate the decomposition by auditing principal pro-simulation papers for which mechanisms each design fails to rule out.

Horton (2023), Homo Silicus. *Observed*: GPT replicates the qualitative direction of effects in Charness–Rabin, Kahneman–Knetsch–Thaler fairness, and Samuelson–Zeckhauser status-quo bias. *Unexcluded mechanism*: training-prior retrieval (f_{train}). The replicated experiments are among the most heavily documented in behavioral economics: many standard textbooks, undergraduate courses, and policy-facing summaries report the directional findings. These findings are widely represented in the public textual environment from which frontier models are trained. The "replication" is compatible with a well-indexed retrieval system. *Test that would distinguish*: experiments composed after the training cutoff with novel stimuli, no close matches in publicly indexed sources, and no obvious pre-cutoff analogs. Horton (2023) does not provide this test.

Argyle et al. (2023), Silicon Sampling. *Observed*: silicon samples track ANES vote-choice and word-association distributions. *Unexcluded mechanisms*: training-prior retrieval (f_{train}) and prompt compliance (g_{prompt}) jointly. ANES 2016–2020 data are public, heavily discussed in academic and journalistic text, and well-represented in the public textual environment. The prompt also explicitly

encodes demographic attributes with known correlations to voting ("Ideologically, I describe myself as conservative. Politically, I am a strong Republican."). The model is instructed to produce outputs consistent with these attributes; the public textual environment contains the correlation. Both mechanisms pull toward the reported fidelity. *Test*: conditioning on demographic attributes that do *not* have well-known correlations to the outcome variable, or on attribute combinations rare in training data.

Park et al. (2023), Generative Agents. *Observed*: 25-agent sandbox produces coordinated behaviors (party invitation chains, friendship formation, job-routine adherence). *Unexcluded mechanisms*: training-prior retrieval (f_{train}) and prompt compliance (g_{prompt}) jointly. The *prompt* contains detailed agent memories ("Maria is a student majoring in Physics at Oak Hill College") and a structured planning-and-reflection scaffolding. The *public textual environment* contains extensive fiction, simulation-game walkthroughs, and social-interaction dialogue covering structurally similar coordination patterns. The design does not include an interaction-removed placebo or prompt-script ablation; therefore, observed coordination cannot be separated from role-script compliance combined with retrieval of well-represented social-interaction scripts. *Test*: adversarial scenarios where training data and prompt-induced narrative push in opposite directions, or where the coordination structure is novel and not present in fiction corpora. Park et al. do not run such tests; the evaluation methodology is qualitative and ethnographic.

Aher, Arriaga, Kalai (2023), Turing Experiments. *Observed*: replication of Ultimatum, Garden Path, Milgram outcomes; hyper-accuracy in Wisdom of Crowds. *Unexcluded mechanism*: training-prior retrieval (f_{train}), *explicitly admitted by the authors* for the Wisdom of Crowds case. The Wisdom of Crowds result is presented as a "distortion," but its diagnostic significance is understated. If f_{train} is not ruled out in one of four replicated experiments, the same mechanism is not ruled out in the other three either, and the apparent agreement is consistent with retrieval of documented human results rather than with simulation of human cognition.

Hewitt, Ashokkumar, Ghezae, Willer (2024). Deferred to §6.

5 What Published Papers Identify

Table 1: Identification Coverage of Principal Empirical LLM Simulation Studies

Paper	Claimed finding	f_{train} probe	g_{prompt} probe	h_{emerge} placebo	Cross-model	Post-cutoff
Pro-simulation empirical studies						
Park et al. 2023 (UIST)	Emergent coordination, 25-agent sandbox	×	×	×	×	×
Horton 2023 (NBER)	Replicates classic behavioral-econ findings	×	×	n/a	×	×
Argyle et al. 2023 (*PA*)	Silicon samples track ANES distributions	×	partial	n/a	×	×
Aher et al. 2023 (ICML)	Turing Experiment replications	partial	×	n/a	✓	×
Hewitt et al. 2024 (WP)	$r = 0.85$ across 476 effects	partial	×	n/a	×	✓
Critical-wing empirical studies						
Bisbee et al. 2024 (*PA*)	Prompt sensitivity + temporal drift	✓	✓	n/a	✓	×
Boelaert et al. 2025 (*SMR*)	Machine bias in opinion polling	✓	×	n/a	✓	×
Tjuatja et al. 2024	LLMs fail to reproduce response biases	✓	✓	n/a	✓	×
Santurkar et al. 2023 (ICML)	Ideological skew in aligned LLMs	✓	partial	n/a	✓	×

Position and commentary papers (e.g., Anthis et al. 2025, ICML 2025 Position Track) are not coded above because they advance arguments rather than empirical identification claims; they are discussed in the text rather than audited on identification probes.

Two patterns are visible. First, **among the pro-simulation empirical papers we audit, no design runs all five probes**; Hewitt et al. (2024) runs the most (one partial). The critical wing systematically runs the probes and finds the failures; the pro-simulation wing does not run them and cannot rule out the failures. Second, **the h_{emerge} placebo test is never performed in this sample**. For papers claiming emergence — Park et al. most directly — there is no comparison to a condition in which interaction structure is removed. The claim of emergent multi-agent behavior is therefore not distinguishable, on the published evidence, from the claim of narrative-scaffolded single-agent compliance replicated across 25 prompted roles.

The aggregation problem. Hewitt et al.'s $r = 0.85$ is a pooled statistic across 476 effects; Bisbee et al.'s within-experiment failures are statistics on specific regression coefficients. These two reports are not in tension: a pool of effects can be on average well-predicted while individual effects are not, particularly if the pool is dominated by effects whose direction is common knowledge. High cross-experiment correlation and poor within-experiment identification are simultaneously consistent with a population of experiments whose mean direction is in training data and whose specific magnitudes are poorly calibrated. This is the core conceptual error in citing $r = 0.85$ as evidence of simulation validity.

6 Why the Hewitt Result Is Not Enough

The strongest predictive-fidelity result in the pro-simulation wing is the 2024 working paper by Hewitt, Ashokkumar, Ghezae, and Willer (Stanford / NYU). Its central empirical result — that across 70 pre-registered nationally representative U.S. survey experiments comprising 476 treatment effects and 105,165 human participants, GPT-4-derived simulation predictions correlate with actual treatment effects at $r = 0.85$, with $r = 0.90$ on post-training-cutoff studies — is the most prominent fidelity result in the audited literature. It is cited in Anthis et al. (2025) and in subsequent simulation work, and is increasingly invoked in arguments for low-cost in-silico piloting.

We advance a three-layer rebuttal. Each layer is separately sufficient to undermine the paper's use as a warrant for substitution; collectively they establish that the $r = 0.85$ result is compatible with our position that prediction fidelity is not identification.

Correlation is the right statistic for forecasting, and the wrong statistic for substitution. The 476 treatment effects pooled in Hewitt et al.'s correlation are drawn from nationally representative survey experiments on U.S. adults, covering political persuasion, health behavior nudges, and social-policy attitude manipulations. The relevant question is what fraction of these 476 effects have directions or approximate magnitudes that are predictable from pre-2023 theory and empirical regularities. That fraction is not reported. This omission matters because a well-trained language model could obtain high cross-experiment correlation by recovering theory-consistent priors and plausible magnitudes, without simulating the causal response process. Published social-science experiments are disproportionately designed around hypotheses the authors expect to find support for, which means hypotheses consistent with prior theoretical and empirical results. The Milkman et al. (2022) 680,000-person vaccination megastudy tested nudges whose directions were predicted by behavioral-economics theory; the Voelkel et al. (2023) democratic-attitudes megastudy tested interventions whose directions were predicted by political-psychology theory. Effect magnitudes are variable and theoretically informative, but signs are largely not surprising to a well-read prior, and the public textual environment from which GPT-4 was trained likely represents the social-psychological, political-science, health-behavior, and behavioral-economics literatures extensively. The f_{train} mechanism is therefore observationally compatible with the headline $r = 0.85$ result unless effects are stratified by pre-cutoff predictability, surprisingness, and novelty. The post-cutoff subset ($r = 0.90$) does not rebut this concern: post-cutoff studies on topics with well-developed prior literatures inherit directional priors from pre-cutoff work, and the direction of effects is substantially determined by experimental-design conventions that are themselves in training data. The post-cutoff result rules out only direct memorization, not the broader prior-driven retrieval mechanism. The appropriate test is prediction on experiments whose direction is not predictable from prior literature: a falsifiable design would stratify Hewitt et al.'s 476 effects into (a) effects whose direction was predictable from pre-2023 theory and (b) effects whose direction was surprising, reversed, exploratory, or null. The retrieval hypothesis predicts r in stratum (a) substantially exceeds r in stratum (b); the genuine-simulation hypothesis predicts comparable values.

Pooling across experiments masks within-experiment failure. Hewitt et al.'s $r = 0.85$ is a cross-experiment correlation over 476 (predicted, observed) effect pairs. This is not a statement about within-experiment fidelity — about whether, for any given experiment, the LLM produces a distribution of simulated responses that matches the human distribution with appropriate variance, demographic structure, higher-order statistical properties, and stability under prompt perturbation. Bisbee et al. (2024), using ANES as the benchmark rather than a pool of megastudies, probe within-experiment fidelity directly. They find: simulated response distributions are compressed relative to human distributions, such that calculated power of downstream analyses would be systematically overstated; regression coefficients differ from human baselines, sometimes substantially in systematic ways; response distributions shift measurably with minor prompt rewording; and the same prompt, run on the nominally same model three months apart, yields significantly different results. These are failures at the single-experiment level that would render in-silico piloting of any specific experiment unreliable. The two findings are fully consistent: across many experiments, the LLM gets direction and rough magnitude right on average; within any given experiment, it does not reproduce the response structure required to substitute for human respondents. A researcher interested in piloting their specific experiment — the policy-relevant use case — should treat the within-experiment evidence as the directly relevant evidence. The language of "pilot" is itself doing substantial work: Hewitt et al.'s authors frame the paper as piloting — low-cost pre-screening of designs to inform resource allocation. This framing is reasonable for the $r = 0.85$ result. The risk in downstream use is that simulation is treated as a substitute for human subjects in actual social-scientific inference; this slippage is where the identification problem becomes consequential.

Prediction and explanation are different scientific operations. The point has been present in the methodological literature since at least Shmueli (2010) in *Statistical Science*. Prediction — producing outputs that correlate with observed outcomes — and explanation — producing a causal account of the generative mechanism — have different epistemic warrants. A model can be highly predictive while offering no explanatory insight; the history of science contains many examples (Ptolemaic epicycles were highly predictive of planetary positions while wrong about celestial mechanics). Hewitt et al. on its own terms establishes prediction; the $r = 0.85$ is a prediction statistic, and the authors do not claim a causal account of how their simulations produce their outputs. The slippage occurs when downstream users treat the prediction result as warrant for the substitution claim. Substitution is an explanatory claim: it asserts that the simulation accesses the same causal structure as human respondents, such that inferences drawn from simulation generalize to inferences about humans. Prediction correlation does not establish this. The substitution claim requires an argument at the level of mechanism — an exclusion restriction allowing the researcher to say that a specific prompt variation affects a specific output only through the hypothesized mechanism, not through retrieval or compliance. Such restrictions can in principle be constructed; they require designs like those sketched in Section 3. They have not been constructed at scale in the literature we audit.

7 The Identification Checklist for LLM Social Simulation

We propose a six-item checklist as a condition for publication of LLM-simulation-based social-scientific claims at NeurIPS and adjacent venues.

The Checklist.

1. **Prompt ablation with variance decomposition**. Systematically remove or replace prompt components (role label, demographic backstory, context, history) and report the fraction of outcome variance attributable to each, using η^2 or equivalent. Prompt-induced variance above a threshold (e.g., 20%) requires caveats about g_{prompt} dominance.
2. **Placebo prompt test**. Design a condition in which the claimed mechanism should produce no effect — agents made identical, or interactions randomized, or demographics shuffled — and report the core metric under it. If it does not collapse, the mechanism is not identified.
3. **Counterfactual perturbation grid**. Perturb at least four prompt dimensions and report a sensitivity matrix; a well-identified claim is robust to wording-level perturbations of constant semantic content.
4. **Training-data leakage audit**. Disclose the model's training cutoff and the publication date of each replicated study. Report separate results for fully post-cutoff versus pre-cutoff

studies. For novelty claims, demonstrate stimuli are not closely matched by public training corpora.

5. **Cross-model robustness**. Replicate core findings on at least three architecturally independent models. Model-specific findings must be restated as model properties, not as "LLM simulation" as a category.
6. **Pre-registration of sign-flip scenarios**. Pre-register prompt variants that would produce different or opposite results on the hypothesized mechanism, and report all variants.

Adoption. Items 1–4 are adoptable immediately through reporting requirements; items 5–6 require additional design but are necessary for substitution claims. Cost calibration and a worked example appear in Appendix.

8 Stakes and Conclusion

When LLM social simulations are treated as substitutes for human-subject evidence, the identification required to license substitution has not been established. The error is not in simulation per se; it is in treating simulation as human-subject data when the identifying assumptions distinguishing training-prior retrieval, prompt compliance, and emergence remain undisclosed. NeurIPS should require explicit identification-strategy disclosure for publication of substitution claims; this will redirect, not slow, the field.

Implication for AIWILD. Read as a statement about agentic deployments in the wild, the argument above identifies an under-recognized class of safety failure. Multi-agent LLM systems can be reliable in the standard agent-evaluation sense (high task-success rates, calibrated outputs, low explicit hallucination on bounded benchmarks) and still be unsafe in the deployment-as-evidence sense, because no agent benchmark currently asks whether the agent's output identifies a target social mechanism rather than reproducing a prior. The identification checklist (§7) is offered as an AIWILD-class safety requirement: a disclosure regime that the workshop community can adopt independently of any single venue, applicable wherever an agent's output is intended to substitute for human-subject evidence in the wild rather than to demonstrate task competence in a controlled benchmark.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proc. 40th International Conference on Machine Learning* (PMLR 202, pp. 337–371).

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J., Brynjolfsson, E., & Bernstein, M. (2025). Position: LLM social simulations are a promising research method. *Proc. 42nd International Conference on Machine Learning* (PMLR 267). arXiv:2504.02234.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4), 401–416. DOI 10.1017/pan.2024.5.

Boelaert, J., Coavoux, S., Ollion, É., Petev, I., & Präg, P. (2025). Machine bias: How do generative language models answer opinion polls? *Sociological Methods & Research*, 54(3), 1156–1196.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3), 817–869.

Filippas, A., Horton, J. J., & Manning, B. S. (2024). Large language models as simulated economic agents. *Proc. 25th ACM Conference on Economics and Computation*.

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. arXiv:2301.01768.

Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R. (2024). Predicting results of social science experiments using large language models. Working paper, Stanford University and New York University, August 2024.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from *Homo Silicus*? NBER Working Paper No. 31122.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, 76(4), 728–741.

Milkman, K. L., et al. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *PNAS*, 119.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proc. 36th Annual ACM Symposium on User Interface Software and Technology* (UIST '23), pp. 1–22. arXiv:2304.03442.

Qu, Y., & Wang, J. (2024). Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11, 1095.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), 7–59.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proc. 40th International Conference on Machine Learning* (PMLR 202, pp. 29971–30004).

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.

Tjuatja, L., Chen, V., Wu, S. T., Talwalkar, A., & Neubig, G. (2024). Do LLMs exhibit human-like response biases? A case study in survey design. arXiv:2311.04076.

Voelkel, J. G., et al. (2023). Megastudy identifying effective interventions to strengthen Americans' democratic attitudes. Working paper.

A Cost Calibration of the Identification Checklist

For a 25-agent sandbox running on frontier API models, the checklist increases the number of required runs but not the conceptual complexity of the design. The added cost is non-trivial, but it is small relative to the cost of a well-powered human-subject experiment. A paper failing the checklist should not make substitution claims without explicit limitations; partial compliance can be acceptable only when the paper limits itself to hypothesis generation or forecasting.

B A Worked Example of an Identified Multi-Agent Simulation Study

To illustrate that the checklist is tractable, not prohibitive, we sketch a study design that would satisfy all six items and produce publishable simulation-based social-scientific findings under the proposed standard.

Research question. Under what prompt conditions do multi-agent LLM simulations produce coordination outcomes that are not reducible to training-data retrieval or single-agent prompt compliance? Specifically: given a novel (post-training-cutoff) coordination task with features not present in the pretraining corpus, do LLM agents in multi-agent interaction produce coordination patterns that a single agent, prompted identically but without interaction, does not produce?

Task. A novel coordination game designed by the researchers in 2026, using unfamiliar labels ("blumph" and "grund" instead of "cooperate" and "defect"), with a payoff structure not reducible to known game-theoretic canonical forms, accompanied by an instructional preamble screened through public web search, benchmark-contamination checks, and embedding-based near-duplicate search, with no close public matches found. The task is run on at least two frontier models with documented training cutoffs.

Arms.

- A1: Multi-agent (25 agents) with full interaction history and role heterogeneity.
- A2: Multi-agent with all roles identical (h_{emerge} placebo).
- A3: Multi-agent with interactions randomized (h_{emerge} placebo, stronger).
- A4: Single agent, prompt-equivalent to A1 without other-agent observations ($g_{\text{prompt}} + f_{\text{train}}$ baseline).
- A5: Single agent, adversarially varied prompt wordings (prompt-sensitivity probe).

Pre-registration. Outcomes (coordination index, within-group variance, cross-iteration stability) pre-registered on OSF before data collection. Primary hypothesis: coordination index in A1 $>$ A2, A3, A4 by a pre-specified threshold. If A1 $\approx$ A4, the emergence claim is rejected in favor of the single-agent compliance hypothesis.

Outcome interpretability. If A1 coordination exceeds A2–A5 by the pre-specified threshold, and this holds on both models, and A5 wording perturbations do not eliminate the effect, the result supports h_{emerge}: genuine multi-agent emergence under conditions where f_{train} and g_{prompt} are controlled. This is a publishable simulation-based social-scientific finding. If A1 coordination does not exceed A4, or is eliminated by A5 wording perturbation, or is model-specific, the result rejects the emergence claim and is still publishable — as an identified-null result — with very different policy implications than a non-identified weak-positive result. Either outcome is informative. Non-identified results are not.

C Anticipated Objections and Responses

Objection 1: "You are applying econometric standards to a machine-learning field; the disciplinary mismatch is the problem."

Response: The disciplinary mismatch is the *symptom*, not the problem. When ML research produces outputs whose intended use is social-scientific inference — substitution for human subjects, policy piloting, population-level extrapolation — it inherits the methodological obligations of social science. Either the research should restrict its claims to in-domain ML benchmarks, or it should adopt the methodological standards appropriate to the claims it makes. We are asking for consistency: the claims and the standards should be matched.

Objection 2: "Identification strategies slow research; LLMs move fast; standards will be obsolete before they take effect."

Response: This is a rhetorically popular but empirically weak argument. Clinical trial standards (CONSORT) did not prevent rapid drug discovery; they changed the *format* of reporting and raised the quality of evidence. Identification-strategy standards in econometrics did not kill applied microeconomics; they made the resulting findings more trustworthy, not less frequent. Methodological rigor and research pace are not in the tradeoff they are assumed to be.

Objection 3: "Hewitt, Horton, Argyle, Park, and Anthis are careful researchers who do not themselves make naive substitution claims; why target their work, and won't this discourage legitimate research?"

Response: We have explicitly endorsed LLM simulation as a *predictive and piloting* tool, and have endorsed the pro-simulation papers as legitimate research that passed peer review. We attribute downstream slippage to downstream users rather than to the original researchers. The original papers' framing, however, is not neutral: framings invoking simulacra of human behavior, claims that simulations advance understanding of society, and branding such as "Homo Silicus" invite the substitution reading even as the technical text is more careful. Calling for identification-strategy disclosure is a call for more careful framing that would foreclose the slippage — a methodological complement to the pro-simulation work, not a wholesale dismissal.

Objection 4: "Your position assumes that 'social scientific evidence' has a stable meaning against which to judge simulation outputs; social science itself is methodologically plural."

Response: Social science is methodologically plural, and different traditions hold different standards. The identification-strategy framework we invoke is specifically the one that has become dominant in applied microeconomics, policy evaluation, and quantitative political science — the fields whose methodological standards are most directly relevant to the downstream uses (policy piloting, treatment-effect estimation, survey response prediction) the simulation literature has targeted. The critique we offer applies specifically to simulation results making quantitative causal claims.