Position: The Identification Crisis in LLM Social Simulation is a Rank-Constrained Mechanism Decomposition Problem

Anonymous Authors[1]

Abstract

This position paper argues that the current evidentiary crisis in LLM-based social simulation is most precisely understood as a *rank-constrained decomposition problem*: the observed outcome from a simulation is a superposition of contributions from training-prior retrieval, prompt-induced role compliance, and genuine interactional emergence, and the published evidence is rank-deficient with respect to separating these components. Recent work reports high predictive fidelity ($r = 0.85$ across 476 effects), while a parallel critical literature shows synthetic respondents fail regression, prompt-sensitivity, and temporal-stability tests. We frame this disagreement as a non-uniqueness phenomenon: in the absence of structural rank constraints on the joint mechanism representation, multiple decompositions of the same observation matrix are equally consistent with the data, exactly the way that an unconstrained matrix factorization admits infinitely many factor pairs. The fix is conceptual before it is technical: simulations become evidence only when the design imposes enough rank structure (placebo conditions, ablation grids, cross-model factors) that the mechanism factors become uniquely separable. We audit the principal LLM social-simulation literature through this rank-constrained decomposition lens, formalize a six-item identification checklist as a set of rank-restoring design moves, and argue that the low-rank-representations community is uniquely positioned to give this evidentiary problem the formal apparatus it currently lacks.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

A central methodological problem in modern AI evaluation is when the output of a learned system can be treated as evidence about the phenomenon it simulates. We argue that the recent evidentiary crisis in LLM-based social simulation is, at its core, the same problem that the low-rank representations community has long studied in a different guise: when can a high-dimensional observation be uniquely decomposed into a small number of structured factors?

A research program treating large language models as substitutes for human subjects in social science has emerged in the post-GPT-4 period. Park, O'Brien, Cai, Morris, Liang, and Bernstein (2023) introduced "generative agents" — LLM-driven agents in a 25-agent sandbox — and argued that observed behaviors constitute "believable simulacra of human behavior." Horton (2023, NBER WP 31122) introduced "Homo Silicus" and replicated Charness–Rabin (2002), Kahneman–Knetsch–Thaler (1986), and Samuelson–Zeckhauser (1988). Argyle, Busby, Fulda, Gubler, Rytting, and Wingate (2023, *Political Analysis* 31:337–351) introduced "silicon sampling" tracking ANES distributions. Hewitt, Ashokkumar, Ghezae, and Willer (2024) report $r = 0.85$ across 70 pre-registered survey experiments (476 effects, 105,165 participants), with $r = 0.90$ on post-cutoff studies. Anthis et al. (2025, ICML PMLR 267) argue "LLM social simulations are a promising research method."

A parallel critical literature reports the opposite picture. Bisbee, Clinton, Dorff, Kenkel, and Larson (2024, *Political Analysis* 32:401–416) find regression coefficients differ systematically from human baselines, distributions are compressed yielding inflated power calculations, response distributions shift with prompt wording, and the same ChatGPT 3.5 Turbo drifts measurably over three months. Boelaert, Coavoux, Ollion, Petev, and Präg (2025, *Sociol. Methods Res.* 54:1156–1196) document strong machine bias varying randomly across topics. Tjuatja et al. (2024) show LLMs do not consistently reproduce human response biases. Santurkar et al. (2023, ICML) show aligned LLMs disproportionately reflect specific demographic subsets.

Why these wings cannot be reconciled by more data. The two findings — high r and systematic structural failures —

are not contradictory; they are the diagnostic signature of a rank-deficient identification design. Predictive fidelity is a one-dimensional summary (correlation, MSE, hit rate) of an observation matrix that has at least three latent contributing mechanisms. Demonstrating that the one-dimensional summary is high tells us about the rank-1 projection, not about whether the underlying decomposition is unique. The pro-simulation results are real measurements at a low-rank projection; the critical results document structure orthogonal to that projection. They cannot be reconciled by more data at the same projection.

The rank-constrained decomposition framing. Let $\mathbf{Y} \in \mathbb{R}^{n \times d}$ be the matrix of simulation outcomes (one row per agent or trial, one column per outcome dimension). We argue $\mathbf{Y}$ admits a structural decomposition

$$\mathbf{Y} = \mathbf{F}_{\text{train}} + \mathbf{G}_{\text{prompt}} + \mathbf{H}_{\text{emerge}} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{F}_{\text{train}}$ captures the contribution from training-prior retrieval, $\mathbf{G}_{\text{prompt}}$ captures prompt-induced role compliance, $\mathbf{H}_{\text{emerge}}$ captures genuine interactional emergence, and $\boldsymbol{\varepsilon}$ is sampling noise. Without further structural constraints (factor independence, rank bounds on individual components, orthogonality conditions across designed contrasts), this decomposition is non-unique exactly as an unconstrained matrix factorization $\mathbf{Y} = \mathbf{A}\mathbf{B}^\top$ is non-unique up to invertible $\mathbf{Q}$: $\mathbf{A}\mathbf{B}^\top = (\mathbf{A}\mathbf{Q})(\mathbf{Q}^{-1}\mathbf{B}^\top)$. The pro-simulation result establishes that some decomposition produces high predictive fidelity; it does not establish that the decomposition with high $\mathbf{H}_{\text{emerge}}$ is the one obtaining.

Position. *The published LLM social-simulation literature reports rank-1 summaries (correlations, marginal accuracies) of an outcome matrix admitting at least a rank-3 mechanism decomposition. Without design moves that impose enough structural constraints to make this decomposition unique — placebo conditions zeroing out specific factors, ablation grids identifying $\mathbf{G}_{prompt}$, post-cutoff stimuli bounding $\mathbf{F}_{train}$, cross-model factors isolating model-specific from mechanism-general structure — the literature cannot license the substitution claims it is being used to support. The right intervention is not to abandon simulation but to import the rank-restoring design discipline that the low-rank representations community has long required for any factorization to be interpretable.* We call on workshops at the intersection of representation learning and AI evaluation to adopt rank-constrained identification disclosure as a publication standard.

Why the low-rank representations community. The mathematical content of identification — when a set of design contrasts plus rank constraints make a decomposition unique — is exactly the content the CoLoRAI community has formalized for tensor decomposition uniqueness, sketching theory, expressivity bounds on factorizations, and circuit identifiability. The evaluation community has the empirical problem; this community has the formal vocabulary. The bridge is overdue.

2. The Landscape of LLM Social Simulation

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

Generative agent sandboxes. Park et al. (2023): 25 LLM-driven agents in a small-town environment, with party-coordination, friendship-formation, and observe-plan-act cycles framed as evidence that LLMs produce believable social dynamics. The methodological move is ethnographic: report agent behaviors illustrated by transcripts. Subsequent work extends sandboxes to financial markets, platform governance, polarization, and organizational decisions.

Silicon-sampling survey replications. Argyle et al. (2023): GPT-3 conditioned on ANES backstories, reproducing ANES marginals and association patterns ("algorithmic fidelity"). Qu and Wang (2024, *HSSC* 11:1095) extend cross-nationally with 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, ICML PMLR 202:337–371): dictator-game, status-quo-bias, Ultimatum, Garden Path, Milgram, Wisdom of Crowds. Both report qualitative agreement in most cases. Aher et al. make a methodologically honest move that is underweighted in subsequent citation: a documented "hyper-accuracy distortion" in the Wisdom of Crowds replication, where the LLM simulation is too accurate relative to human crowds — diagnostic of retrieval from aggregated ground-truth in training data, rather than simulation of heterogeneous cognitive biases.

Predictive treatment-effect simulation. Hewitt, Ashokkumar, Ghezae, and Willer (2024) assemble 70 pre-registered nationally representative U.S. survey experiments (476 effects, 105,165 participants) and prompt GPT-4: $r = 0.85$ overall, $r = 0.90$ post-cutoff. Anthis et al. (2025, ICML PMLR 267) cite Hewitt et al. as central validation.

The critical wing. Bisbee et al. (2024) using ChatGPT 3.5 Turbo elicit feeling-thermometers for 11 sociopolitical groups under various persona conditions and compare to ANES 2016–2020 baselines: marginals close, regressions divergent, distributions compressed (inflated power), prompt-wording sensitivity, three-month drift. Replicated on ChatGPT 4.0 and Falcon-40B-Instruct: not model-specific. Boelaert et al. (2025) document systematic political-ideological bias in generative AI to opinion polls. Santurkar et al. (2023) show aligned LLMs over-represent specific demo-

graphic subsets. Hartmann, Schwenzow, and Witte (2023, arXiv:2301.01768) document left-libertarian skew in ChatGPT. Tjuatja et al. (2024, arXiv:2311.04076) show LLMs do not reproduce human response biases. The critical wing has collectively established prompt sensitivity, temporal instability, compressed variance, ideological skew, cross-cultural degradation, and failure to reproduce response biases. What it has not yet done is frame these findings under a unifying theoretical apparatus. *The rank-constrained decomposition framing supplies that apparatus.*

3. The Identification Problem as Rank Recovery

We make Equation 1 concrete at the row level. Let Y_i be the observed behavior of simulated agent i. Let $X_i = (R_i, C_i, H_i)$ denote the prompt configuration: R_i the role label ("You are a 45-year-old conservative voter"), C_i the context, H_i the interaction history. Then:

$$Y_i = f_{\text{train}}(X_i) + g_{\text{prompt}}(R_i, C_i) + h_{\text{emerge}}(X_{-i}, H_i) + \varepsilon_i. \tag{2}$$

The pro-simulation interpretation of high predictive fidelity assumes h_{emerge} is large and meaningfully tracks human social dynamics. The critical-wing interpretation of systematic failures (compressed variance, prompt sensitivity, drift) is consistent with f_{train} and g_{prompt} dominating. *Predictive fit alone does not separate these mechanisms because the decomposition is unconstrained.*

Why prediction does not identify. A model can be highly predictive of $\mathbf{Y}$ without identifying any mechanism: a sufficiently expressive function class fits the marginal distribution. Identification asks the structural question: does the joint design of prompts and observations make the mapping from latent mechanisms to outcomes injective? The decomposition in Equation 1 is rank-3 in mechanism space; $r = 0.85$ to a one-dimensional human-effect target is rank-1 information about that decomposition. Two more rank-1 design contrasts are needed before the mechanism factors become uniquely identifiable.

Three rank-restoring design contrasts.

1. **Placebo prompts** (zero out g_{prompt} contrast): identical agents, randomized interactions, or shuffled demographics. If the core metric does not collapse, the metric was not driven by the labeled mechanism.
2. **Post-cutoff stimuli** (bound f_{train}): use stimuli published after the model's training cutoff. Effect sizes here bound the contribution of retrieval. The remaining gap between pre- and post-cutoff r is the structural identification of f_{train}.
3. **Cross-model factors** (isolate model-specific from mechanism-general structure): replicate on at least three architecturally independent models. Findings consistent across models bound f_{train} to be a property of the simulation paradigm, not a model-specific quirk.

These three contrasts are exactly the rank-restoring moves the low-rank-representations community uses to make tensor or matrix decompositions uniquely identifiable. The CoLoRAI community has formalized when such design contrasts are sufficient (Kruskal-style conditions, sketching guarantees, expressivity-rank lower bounds). The simulation literature has not.

4. A Mechanism-by-Mechanism Audit

Park et al. (2023). Reports observed agent behaviors via transcripts. Does not include prompt-ablation conditions, placebo conditions, or post-cutoff stimuli. The "believable simulacra" framing identifies h_{emerge} implicitly. The design imposes no rank constraint separating prompt-script execution (g_{prompt}, with elaborate role specifications, daily plans, observation-reflection memory) from emergence (h_{emerge}). Cross-model robustness is absent (single model). The paper produces a rank-1 demonstration; the rank-3 decomposition remains unidentified.

Argyle et al. (2023). Conditions GPT-3 on ANES respondent backstories and reports marginal-distribution match and pattern-of-association match ("algorithmic fidelity"). The training-data overlap is structural: ANES is a public survey extensively analyzed in publicly available political science scholarship; the model's training corpus likely contains substantive ANES-derived statistical regularities. Without post-cutoff stimuli, f_{train} is not bounded. Qu and Wang (2024) replicate cross-nationally and find substantial degradation outside the WEIRD sample. The cross-cultural degradation pattern is what an unconstrained f_{train}-driven decomposition predicts: where training data are sparse, the apparent "simulation" fails.

Hewitt, Ashokkumar, Ghezae, Willer (2024). The strongest predictive-fidelity result: $r = 0.85$ overall, $r = 0.90$ post-cutoff. Even on the subset that bounds f_{train}, the result remains. *We concede this is the strongest observed lower bound on* $\mathbf{H}_{emerge}$ + ($\mathbf{F}_{train}$ *from training-distribution properties beyond direct stimulus retrieval*). It is not a separation of these. Possibilities consistent with the data: (a) the model has learned general regularities about how American survey respondents react to political messaging, sufficient to predict effects without simulating mechanism; (b) the model has learned to imitate the implicit response distribution embedded in published survey-experiment literature; (c) the model is genuinely simulating respondent reasoning. All three are rank-1 consistent with $r = 0.90$.

Bisbee et al. (2024). The most sophisticated critical study. The compressed variance finding is the rank diagnostic the field has been waiting for: if the observed $\mathbf{Y}$ has lower effective rank than human-baseline data, this is a quantitative measure of $\mathbf{H}_{\text{emerge}}$ failure to reproduce the dimensionality of human heterogeneity. Their three-month drift result is also a rank diagnostic: if the same prompt produces different $\mathbf{Y}$ at t and $t+90$ days, the mapping from prompt to outcome is not the function the substitution claim presupposes.

Aher et al. (2023). Their hyper-accuracy distortion in the Wisdom of Crowds replication is, in our framing, a direct diagnostic of $\mathbf{F}_{\text{train}}$ dominance: an LLM crowd is too accurate because it retrieves aggregated ground truth that is in its training corpus, not because it simulates the cognitive biases of an actual heterogeneous human crowd. Within a pro-simulation paper, this is an instance of the rank-decomposition pathology we describe.

5. Why the Hewitt Result is Not Enough

The Hewitt et al. (2024) result is genuinely impressive and is, fairly, the strongest case for treating LLM simulation as evidence. We engage it directly.

What it establishes. A frontier model conditioned on documented experimental protocols predicts, with $r = 0.85$–0.90, the human-measured average treatment effects of preregistered survey experiments. This is a rank-1 prediction from a high-dimensional input (full protocol text + sociodemographic targeting) to a low-dimensional target (mean treatment effect, possibly with subgroup heterogeneity).

What it does not establish. That the model is simulating the cognitive process generating the human responses, as opposed to extrapolating from regularities about how documented survey-experiment protocols typically produce documented effects. The post-cutoff $r = 0.90$ rules out direct stimulus retrieval but does not rule out distributional retrieval at the protocol-template level. Effect sizes in social science are heavily structured by protocol genre; a sufficiently large model trained on the meta-distribution of survey-experimental protocols can predict effect sizes well without ever simulating the cognition of a respondent.

The decisive test the literature has not run: *Hewitt-style protocols in domains where social-science conventions about protocol-effect relationships do not exist*. If r remains 0.85 in a domain genuinely untouched by the model's training-distribution-of-protocols, the substitution claim is bounded. If r collapses, then the existing $r = 0.85$ is bounded to be a measurement of model-internal regularities about a literature, not a measurement of human cognition. This is a classic out-of-distribution test in low-rank decomposition: the apparent rank-1 fit holds in-domain not because the underlying rank-3 decomposition is identified, but because the in-domain projection is well-spanned by training data.

6. The Identification Checklist as Rank Constraints

We propose a six-item checklist as a condition for publication of LLM-simulation-based social-scientific claims at NeurIPS, ICML, and adjacent venues. Each item is, in the language of this workshop, a rank-restoring design constraint that contributes to making Equation 1 uniquely identifiable.

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 20% requires caveats about g_{prompt} dominance. *Rank role:* bounds $\mathbf{G}_{\text{prompt}}$.

2. **Placebo prompt test.** Design a condition where the claimed mechanism should produce no effect: agents made identical, interactions randomized, demographics shuffled. Report the core metric. If it does not collapse, the mechanism is not identified. *Rank role:* zeroes out the labeled component.

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. *Rank role:* verifies that semantic content, not surface form, drives the projection.

4. **Training-data leakage audit.** Disclose model training cutoff and study publication dates. Report separate results for fully post-cutoff vs. pre-cutoff studies. For novelty claims, demonstrate that stimuli are not closely matched by public training corpora. *Rank role:* bounds $\mathbf{F}_{\text{train}}$ from direct retrieval.

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. *Rank role:* separates model-idiosyncratic structure from mechanism-general structure.

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. *Rank role:* prevents g_{prompt}-driven artifacts from masquerading as h_{emerge}.

Items 1–4 are adoptable immediately through reporting requirements; 5–6 require additional design but are necessary

for substitution claims. Together, they impose enough structural constraints on Equation 1 to make the mechanism factors approximately uniquely identifiable, in the same sense that orthogonality, sparsity, or non-negativity constraints make matrix factorizations identifiable in classical low-rank settings.

7. Stakes and Conclusion

When LLM social simulations are treated as substitutes for human-subject evidence, the rank constraints required to license substitution have not been imposed. The error is not in simulation per se; it is in interpreting rank-1 predictive fidelity as if it were rank-3 mechanism identification. The fix is design-level: import the rank-restoring contrast structure that the low-rank representations community has long understood is necessary for any factorization to be uniquely interpretable.

Two communities are needed for this fix. The simulation community has the empirical phenomena and the substantive stakes. The low-rank-representations community has the formal apparatus for thinking about uniqueness, identifiability, and the design contrasts required to recover structured factors from observation matrices. We position this paper as a bridge inviting that collaboration.

References

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

Anthis, J. R., et al. (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. 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 and NYU, 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.

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 ACM 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.

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).

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.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions. *Linear Algebra and its Applications*, 18(2), 95–138.

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., & Faloutsos, C. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Processing*, 65(13), 3551–3582.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., & Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15, 2773–2832.

Vergari, A., Choi, Y., Liu, A., Teso, S., & Van den Broeck, G. (2021). A compositional atlas of tractable circuit operations. *NeurIPS 34*.

A. Connection to Tensor Decomposition Identifiability

The rank-constrained decomposition framing of Equation 1 maps onto the established formal apparatus of tensor decomposition uniqueness. Treating the simulation outcome as a tensor $\mathcal{Y} \in \mathbb{R}^{n \times d \times m}$ indexed by agent, outcome dimension, and design contrast, the mechanism-decomposition $\mathcal{Y} = \mathcal{F}_{\text{train}} + \mathcal{G}_{\text{prompt}} + \mathcal{H}_{\text{emerge}} + \mathcal{E}$ becomes uniquely identifiable under analogues of Kruskal's condition: roughly, when the design contrasts m provide sufficient orthogonality across the three mechanism modes. The six-item checklist in Section 6 can be read as a domain-specific instance of a Kruskal-style uniqueness condition: items 1–3 supply contrasts identifying $\mathbf{G}_{\text{prompt}}$, item 4 supplies a contrast bounding $\mathbf{F}_{\text{train}}$, item 5 supplies cross-mode orthogonality. Formalizing the precise correspondence is a research direction we view as tractable and worth pursuing in collaboration with the CoLoRAI community.