GENERATIVE AI AND THE SOCIAL SCIENCES: RETHINKING METHODOLOGY WITH LARGE LANGUAGE MODELS AND SYNTHETIC POPULATIONS.

Abstract

The advent of large language models (LLMs) promises a new paradigm shift in social science methodology with unprecedented powers to simulate human populations, create synthetic survey data, and model complex social dynamics.This paper explores the epistemological underpinnings and methodological considerations of generative AI in social research, engaging with this topic critically.We believe that the transformative promise of LLM is that they can be used to tackle persistent data issues such as missing observations, privacy concerns and counterfactual inaccessibility, but that requires radical rethinking of traditional statistical assumptions.Based on advances in synthetic population generation, agent-based modeling, and benchmark evaluation frameworks, we highlight three main challenges: the false sense of probabilistic inference when using LLM outputs, demographic representativeness and diversity collapse and the trade-off between individual-level prediction and population-level validity.We perform the synthesis of new solutions, such as parametric social identity injection, evolutionary generation of the social identity, and domain-specific training approaches.The paper ends by suggesting a pragmatic methodological approach where LLM can be utilized and used as a high capacity pattern matcher to generate hypothesis and run controlled experiments, but not a replacement for probabilistic inferences from representative samples.

Keywords: Generative AI, Large Language Models, Synthetic Populations, Social Science Methodology, Agent-Based Modeling, Computational Social Science

1. Introduction

Basic data problems have long plagued social science research.However, variables of interest (such as wealth, subjective well-being, cognitive ability) are often hard to measure accurately, longitudinal datasets are often not long enough to capture life course processes, surveys often suffer from nonresponse and panel attrition that systematically distorts population estimates, and

privacy concerns limit access to rich administrative data (Xie et al., 2026) .Most importantly, social scientists do not have the ability to directly observe counterfactual results necessary for inferring causes (Morgan &Winship, 2014).

The rise of large language models (LLM) has created a lot of interest in the social science fields as a way to overcome these limitations (Horton et al., 2023; Argyle et al., 2023).More recently, researchers have exploited these models for simulations of human populations, behaviours, and social interactions (Binz& Schulz, 2023; Park et al., 2023; Hewitt et al., 2024), which has led to a speculation that a paradigm shift is imminent in the methodology of the social sciences (Madden, 2025) .

A methodological frontier but one that has epistemological pitfalls.However, Madden warns that asking LLMs as if they were humans that have been sampling from stable real-world distributions may result in incorrect conclusions.LLMs don't explicitly model uncertainty or variation in phenomena they simulate and aren't probabilistic inference engines in the strict sense of the word.They are not strict Bayesian probabilistic inference engines, in that they do not explicitly model uncertainty or variation in phenomena that they simulate, and they generate text probabilistically through auto-regressive token prediction.

In this paper the following three central questions are explored:

1. What are the statistical and epistemological constraints to the use of LLMs as synthetic agents in social research?
2. How can synthetic populations be created to be representative of the human population?
3. What methodology may help to ensure the valid use of generative AI in social science?

2. Epistemological Foundations: LLMs as Synthetic Social Agents

2.1 The Promise and Peril of Homo Silicus

The idea of using computational agents to model human behavior is old: in economics, agent-based models have been used to model economic behavior (Epstein & Axtell, 1996); and in psychology, cognitive architectures have been used to model human cognition (Anderson et al.,1998).LLMs are a step-up in terms of quality; they can produce human-like text without the need to program them with a specific set of behavioral rules (Horton et al., 2023).

In recent work LLM-as-respondent has been used as a cheap alternative to human sampling.Argyle et al. (2023) showed that GPT-3 was able to produce subpopulations that match human survey distributions when given demographic backstories.Park et al. (2024) proposed a survey automation framework which demonstrated that the generated responses from LLM were similar to the patterns found in the U.S. General Social Survey.For social science experiments, GPT-4 has been seen to predict the treatment outcomes with greater accuracy than human experts in some cases (Hewitt et al., 2024).

But Binz et al. (2025) made bold claims that "foundation models can predict and simulate human behavior in any experiment expressible in language," which has garnered sharp methodological criticism.The statistical underpinnings behind these statements are still subject to dispute, as we will show (Figure 1).

2.2 The Illusion of Probabilistic Inference

One of the fundamental epistemological risks is that of misinterpreting the LLM's output (Figure 1).Traditional social science methodology is based on drawing probabilistic inferences from samples to populations, and generally assuming exchangeability or conditional independence (Gelman& Hill, 2007; Pearl, 2009).In essence, LLM's challenge these assumptions.

It was shown by Falck et al. (2024) that LLMs fail to satisfy the martingale property that is required for coherent Bayesian inference.This condition guarantees that forecasts remain the same when the order of observation is changed.As a true Bayesian posterior should, predictions systematically deviated from one another between GPT-4, Llama-2 and Mistral-7B, as reported by Falck et al. (2024).They called this "introspective hallucination": the model would ask itself questions, and add additional data elements to its context, which it would then use to modify its own predictions.

The discovery has far-reaching implications.If an LLM is employed for simulation of a population, it might produce one agent after another, therefore, its predictions might be influenced by the arbitrary generation order.This order dependence would make it difficult to trust the data generated by the LLM for any research design in which the order of observations is known to be unimportant (Madden, 2025).

2.3 Pattern Matching vs. Statistical Inference

The apparent success of predictive models by LLM is because they are the best pattern matching systems that are trained on large amounts of data (Brown et al., 2020) (Figure 1).Predicting behaviour or attitude when using an LLM is not an example of statistical inference, but an example of interpolating from learned patterns of natural language (Bommasani et al., 2021).

This distinction is important because it is difficult to see the training data without their own complex selection biases.WEIRD people (Western, Educated, Industrialized, Rich, Democratic) tend to be overrepresented in the internet, in text corpora that have been digitized, and in socially dominant perspectives (Henrich et al., 2010).These biases are then imparted to LLM's by the model as well, potentially exacerbating bias in synthetic data (Bisbee et al., 2024; Santurkar et al., 2023).

Furthermore, the alignment of LLMs to human preferences using the Reinforcement Learning from Human Feedback (RLHF) tuning process tends to yield "agreeable" outputs that "flatten out behavioral variation" (Yun et al., 2025).This systematic mode collapse is no simple phenomenon of noise, but rather an intrinsic characteristic of the current architectures that actively reduce the diversity needed for social science research (Paglieri et al., 2026).

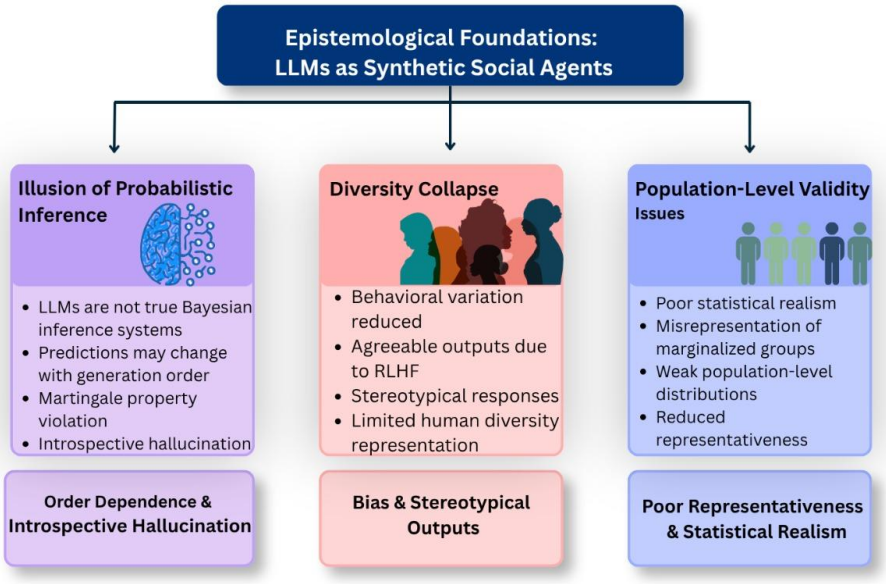

Figure 1. Epistemological foundations of LLMs as synthetic social agents

3. Current Applications and Empirical Findings

3.1 Survey Simulation and Public Opinion Research

The domain of social science where LLM is most developed is survey simulation (Figure 2).This has been pioneered by Argyle et al. (2023) who trained GPT-3 on rich demographic backgrounds and ran silicon samples against the distributions of human surveys over several surveys.Their results showed a very high level of accuracy on a number of items, but also regular differences with respect to politically charged issues.Park et al. (2024) expanded this framework by automating the survey design, administration, and analysis, and validated with the General Social Survey.Using artificial intelligence-generated media personas, Cui, (2025) replicated most of the key results from media-effects studies, along with revealing systematic limitations.

In a recent study, Chen et al. (2025) assessed the prediction performance of LLMs on the results of economic field experiments but identified significant prediction failure with regard to sensitive social categories, such as gender, ethnicity and social norms.The pattern indicates that historically marginalized groups may be misrepresented systematically by LLMs.

3.2 Multi-Agent Social Simulations

In addition to single-agent survey data, researchers have integrated LLMs with other agents to investigate interactional dynamics, such as in multi-agent tasks (Figure 2).In addition to single-agent survey data, researchers have incorporated LLMs into multi-agent scenarios to examine interactional dynamics, including multi-agent tasks.Park et al. (2023) developed "generative agents" that have memory, planning, and reflection capabilities and generate believable micro and meso level social behavior in sandbox environments.These agents coordinated, shared and established social relationships without explicit programming.

Kozlowski &Evans (2025) created and applied generative agent-based modeling to model societal reactions to COVID-19 policies.MouriZadeh Khaki & Choi, (2026) evaluated LLM agents for economic games and Cross et al. (2026) investigated status signaling phenomena.Matyas et al. (2026) simulated processes of psychological attitude change.

To systematically examine when LLM systems produce realistic social networks, Kilaru et al. (2026) identified prompt architecture as a substantive sociological variable.The model scale as

well as the inbreeding homophily and connectivity patterns changed, introducing qualitative differences, if not noise.Importantly, LLM-generated networks performed competitively with real social networks both on clustering and modularity, while systematically introducing demographic biases that are above empirical levels, consistently across each of these metrics.

3.3 Social Digital Twinning

The most ambitious applications involve full ‘social digital twins' simulated populations that are simulated to be a clone of real societies in order to test policy options (Figure 2).Gürcan et al. (2025) introduced a Social Digital Twinning platform, combining real-world data with LLM-aided agent-based simulation, enabling stakeholders to interactively experiment with intervention measures.They presented a demonstration on youth school dropout in Norway, highlighting what is possible when practicing evidence-based policymaking through policy simulation.

In the same way, researchers at urban analytics have created synthetic data pipelines that simulate human reaction to urban environments, allowing big-scale perception studies without the need for field data collection (Jiang et al., 2025).These applications also highlight important issues of whose views are being represented, and who is missing from simulations based on biased data.

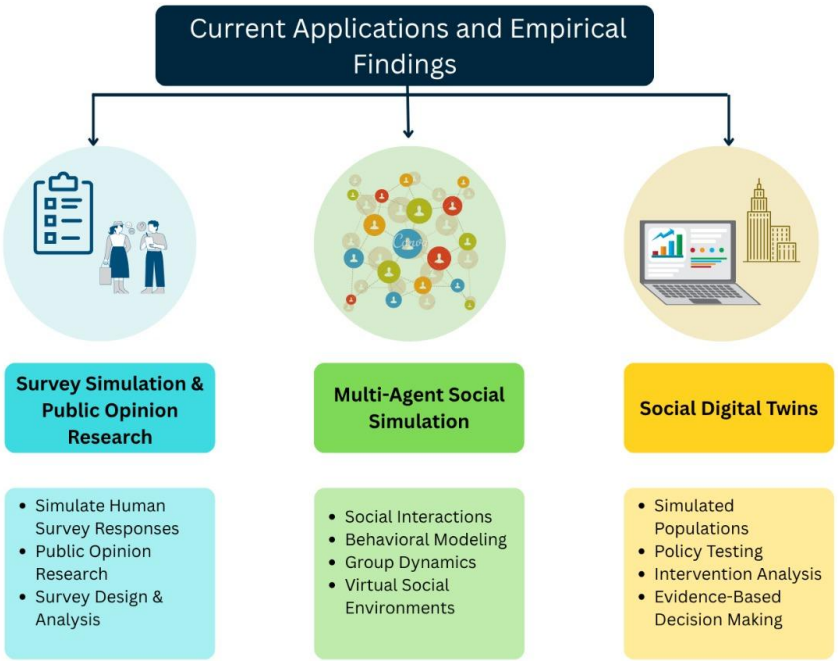

Figure 2. Current applications and empirical findings of LLM

4. The Challenge of Synthetic Population Generation

4.1 Diversity Collapse and Representational Bias

One of the common issues with LLM-based social simulation is the phenomenon of "diversity collapse," which results in the generated populations converging towards stereotypical, agreeable, or majority answers, and omitting the spectrum of human variation (Bisbee et al., 2024; Anthis et al., 2025).

The reason for this problem is several.RLHF tuning aims to maximize responses that humans consider good explicitly, systematically disadvantaging responses that deviate from the majority (Santurkar et al., 2023).Secondly, training data is biased towards the majority views and minority and outlier views are underrepresented.Third, conventional prompting methods do not produce diversity; just request an LLM to "generate diverse personas" results in populations confined to stereotypical responses (Paglieri et al., 2026).

The implications of social science are serious.Density matching a method for optimizing synthetic populations to replicate the average response pattern can ignore rare, consequential outliers that are most critical to understanding social dynamics and/or stress testing interventions (Paglieri et al., 2026).As mentioned in the blog of IAAC (2025), synthetic populations are "infused with the biases of the data and models on which they are based, and can be used to perpetuate dominant views while minimizing or even silencing the voices of historically marginalized groups.

4.2 Population-Level Statistical Realism

Xie et al. (2026) created a systematic benchmark for assessing population-level statistical realism of LLM-generated social science data, named SSDataBench.They evaluate five basic pattern types: univariate distributions, bivariate associations, multivariate prediction of outcomes, life event sequence distributions, and associations between life event sequences and covariates.

Xie et al. (2026) systematically identified representational limitations of 15 state-of-the-art LLMs across four long- and three cross-sectional surveys on demographics, socioeconomic status, marriage, health, abilities, and attitudes.In sparse conditioning regimes, current LLMs fall

back to “typological reasoning” that reduces the diversity of real-world data into overly stereotypical patterns, such as collapsed univariate distributions, overgeneralized bivariate associations, and failure to account for life-course dependencies.

Most importantly, these restrictions are not scale dependent.There are no capacity benchmarks for the performance, and newer versions of the same model do not consistently outperform the older versions (Xie et al., 2026).Importantly, it suggests a structure problem: The advancement of LLMs is largely driven by case-wise prediction accuracy goals that can actively reinforce typological tendencies that are not necessarily population distributions.

4.3 Emerging Solutions

A few interesting solutions are available to solve these problems (Figure 3).

Figure 3. Interesting emerging solution to solve LLMs problems

Parametric Social Identity Injection (PSII): Hu et al. (2025) and the PSII framework of Wang et al.(2026) embed demographic and value-oriented vectors into LLM hidden states.This does not simply prompt the model to output social identity, but it explicitly represents social identity in the model's model space, where synthetic agents are free to produce social identity that embodies inter-group heterogeneity and intra-group variability.The framework creates narrative

personas from social media data, applies quality assessment to select high fidelity profiles and employs importance sampling to ensure the resulting distribution of the profiles is aligned with the reference psychometric distributions, such as the Big Five personality traits.

Evolutionary Persona Generation: Paglieri et al. (2026) proposed Persona Generators which reverse the optimization direction from density matching to support coverage, meaning that synthetic populations cover all the space of possible traits, opinions and preferences.Evolved generators substantially improve over baselines on six diversity metrics with the help of an AlphaEvolve iterative improvement loop using LLM mutation operators.Notably, generated populations are able to successfully mitigate the problem of LLM mode collapse, resulting in more accurate representations of true human trait distributions than even baseline models that are tuned to human statistics.

Domain-Specific Training: Xie et al. (2026) present initial evidence that training can be very powerful when it is specific to social science data, for increasing realism at the population level.This indicates a trajectory away from generic social science models towards more specialized social science foundation models that are trained to model statistical patterns, not to maximize case-wise prediction.

4.4 Evaluation Frameworks

Synthetic data must be evaluated in several dimensions.Jiang et al. (2025) tested four generative approaches (regression-based Synthpop, deep learning CTGAN and TVAE, and LLM-based REaLDTabFormer) on a large voter file dataset, measuring their utility, fidelity, and privacy.They found systematic trade-offs, with Synthpop achieving the best fidelity overall, TVAE achieving best performance downstream, but worse general fidelity and potential for overfitting, REaLDTabFormer performing similarly and CTGAN performing best for privacy protection.

If we consider social science applications, we believe that the main validity criterion should be population-level statistical realism, or the ability to replicate real social systems' distributions, associations, and structural relationships (Xie et al., 2026).The standard is based on traditional survey research principles that date back over 80 years (Groves et al., 2011).

5. Rethinking Methodology: A Pragmatic Framework

5.1 Appropriate Scope Conditions

Based on the analysis above, we suggest a pragmatic reframing of social science methodology in the era of generative AI.Based on the analysis above, a pragmatic reframing of social science methodology in the era of generative AI is proposed.LLMs are not probabilistic inference systems, but rather "high-capacity pattern matchers for quasi-predictive interpolation under explicit scope conditions", as Madden (2025) states.

Table 1 Summary of appropriate and inappropriate applications

Appropriate Applications	Inappropriate Applications
Hypothesis generation and prototyping	Population parameter estimation
Controlled experimentation with known biases	Inference about marginalized populations
Intervention exploration in silico	Substitute for representative sampling
Augmenting incomplete data with validation	Causal identification without ground truth
Cross-cultural pattern exploration	Generalizing beyond training distribution

5.2 Methodological Guardrails

We propose 5 practical guardrails for rigorous research:

1. Preregistered Human Baselines: The findings of LLM should be preregistered and compared with human baselines to gauge expectations of effect sizes and variability (Madden, 2025)
2. Reliability-Aware Validation: Researchers should validate the output generated by LLMs with ground-truth human data, such as out-of-distribution tests and sensitivity analyses (Xie et al., 2026).
3. Subgroup Calibration: Assess performance across demographic subgroups to look for systematic misrepresentation using synthetic data (Argyle et al., 2023; Kilaru et al., 2026).
4. Order-Insensitive Sampling: If applying sequential populations, explicitly check for order dependence, diagnostics include Falck et al. (2024)'s martingale tests.
5. Diversity Auditing: Report Diversity Metrics along with aggregate metrics, such as support coverage and tail representation (Paglieri et al., 2026).

5.3 Hybrid Methodological Designs

The most promising approach is the hybrid designs, which combine both synthetic and real data (Jiang et al., 2025; Gürcan et al., 2025)LLM-based simulation is not intended to replace traditional methods, but to provide an alternative (Figure 4).

- Develop and pilot test surveys and experimental designs
- The uncertainty propagation of data augmentation for missing observations is discussed.
- What if policy simulations (counterfactual generation)
- Sensitivity analysis for robustness under alternative distributions in the population

This "hybrid" strategy combines the epistemic strengths of traditional inference with the special capabilities of generative AI for exploring and augmenting.

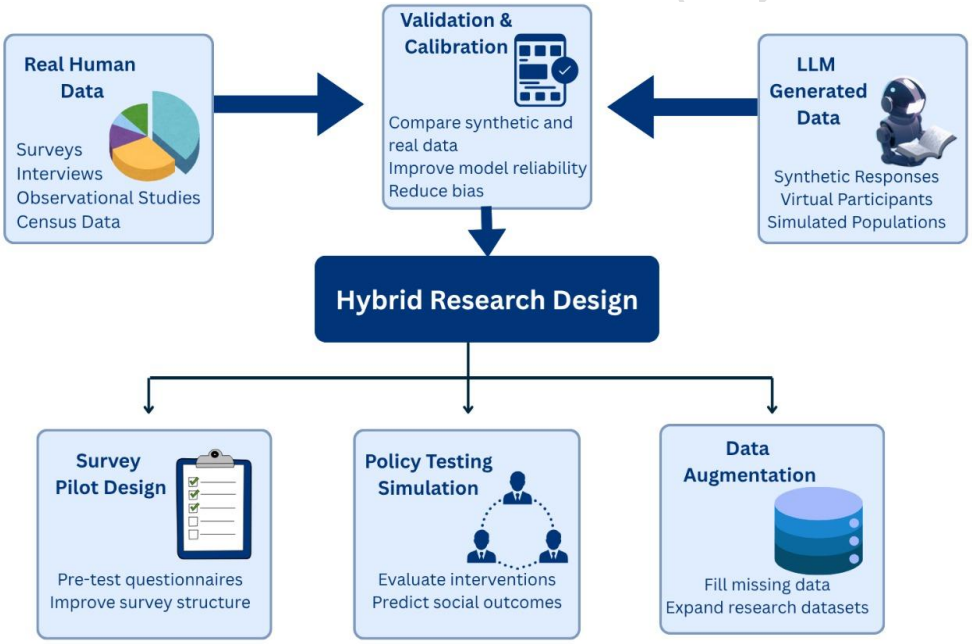

Figure 4. Schematic representation of hybrid methodological designs

6. Future Directions

Finally, three research priorities are outlined:First, the development of special social science foundation models explicitly trained to model statistical patterns at the population-level instead of to optimize case-wise prediction (Xie et al., 2026).These models would need to be trained to

new goals, evaluated to new standards, and the architecture of the model would need to be innovative.Second, the development of professional guidelines and reporting standards for LLM-based social simulation a process that parallels the development of guides and standards for survey research and experimental design.These should include that transparency and validation, uncertainty communication, diversity auditing should be covered.Third, ongoing interdisciplinary exchanges between computer scientists building generative models and social scientists who have an understanding of human populations.The problems that we have come up with are not simply technical, but epistemological in nature, in that joint attention to the nature of social scientific inference, the validity of measurements, and the theory of statistics is required.With the ongoing development of generative AI technology, the question isn't “Will social scientists use these tools?” but “How will they use these tools responsibly?”Robust methodological frameworks that recognize the potential and constraints of LLMs can be put in place to enable the benefits of generative AI while maintaining the epistemic integrity of social science.

7. Conclusion

Generative AI and large language models may herald a new era of social science methodology by providing new capabilities to simulate human populations and investigate social dynamics in silico.But this potential has to be changed in basic assumptions of traditional statistics and new validation frameworks have to be created.We find three main problems: the epistemological divide between LLM pattern matching and probabilistic inference, the longstanding concern about demographic representativeness and diversity collapse, and the conflict between optimizing for individual-level prediction and population-level validity.New technologies like parametric social identity injection, evolutionary persona generation and domain-specific training offer a way forward.

References

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J.,& Bernstein, M. (2025). Llm social simulations are a promising research method. *arXiv preprint 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.

Anderson, J. R., Bothell, D., Lebiere, C., &Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341-380.

Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

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.

Berti, P., Dreassi, E., Leisen, F., Pratelli, L., &Rigo, P. (2025). A probabilistic view on predictive constructions for Bayesian learning. *Statistical Science*, *40*(1), 25-39.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., & Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, *644*(8078), 1002-1009.

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.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., &Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Chen, Y., Hu, Y., & Lu, Y. (2025). Predicting field experiments with large language models. *arXiv preprint arXiv:2504.01167*.

Cui, J. (2025). Human-AI in Digital Media: Effects on Digital Twins and the Moderating Role of Personalized Advertising. *Media, Communication, and Technology*, *1*(1), p35-p35.

Cross, L., Grau-Moya, J., Cunningham, W. A., Vezhnevets, A. S., &Leibo, J. Z. (2026). A Generative Model of Conspicuous Consumption and Status Signaling. *arXiv preprint arXiv:2603.13220*.

De Finetti, B. (1974). Theory of probability: Wiley classics library.

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

Falck, F., Wang, Z., & Holmes, C. C. (2024, April). Are large language models bayesian? a martingale perspective on in-context learning. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Fortini, S., &Petrone, S. (2023). Prediction-based uncertainty quantification for exchangeable sequences. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2247).

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gürcan, Ö.,Falck, V., Rousseau, M. G., & Lima, L. L. (2025, June). Towards an llm-powered social digital twinning platform. In *International Conference on Practical Applications of Agents and Multi-Agent Systems* (pp. 118-130). Cham: Springer Nature Switzerland.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., &Tourangeau, R. (2011). *Survey methodology*. John Wiley & Sons.

Henrich, J., Heine, S. J., &Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, *33*(2-3), 61-83.

Hewitt, L., Ashokkumar, A., Ghezae, I., &Willer, R. (2024). Predicting results of social science experiments using large language models. *Preprint*.

Horton, J. J., Filippas, A., & Manning, B. S. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (No. w31122). National Bureau of Economic Research.

Hu, Z., Lian, J., Xiao, Z., Xiong, M., Lei, Y., Wang, T.,&Xie, X. (2025). Population-aligned persona generation for llm-based social simulation. *arXiv preprint arXiv:2509.10127*.

Jiang, Y., Liang, S., & Choi, J. (2025, July). Synthetic Survey Data Generation and Evaluation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1* (pp. 2292-2302).

Kilaru, S. H., Manikyala, S. T., Upadhyay, R., Ramavath, S. S. K., Nunavathu, S., &Alharthi, D. (2026). When Do LLMs Generate Realistic Social Networks? A Multi-Dimensional Study of Culture, Language, Scale, and Method. *arXiv preprint arXiv:2605.12898*.

Kozlowski, A. C., & Evans, J. (2025). Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. *Sociological Methods & Research*, *54*(3), 1017-1073.

Madden, E. R. (2025). Evaluating the Use of Large Language Models as Synthetic Social Agents in Social Science Research. *Journal of Social Computing*, *6*(4), 334-341.

Matyas, J., Cunningham, W. A., Vezhnevets, A. S., Mobbs, D., Duéñez-Guzmán, E. A., &Leibo, J. Z. (2026). Stabilising Generative Models of Attitude Change. *arXiv preprint arXiv:2604.19791*.

Morgan, S. L., &Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.

MouriZadeh Khaki, A., & Choi, A. (2026). Evaluating Fairness in LLM Negotiator Agents via Economic Games Using Multi-Agent Systems. *Mathematics*, *14*(3), 458.

Novikov, A., Vũ, N., Eisenberger, M., Dupont, E., Huang, P. S., Wagner, A. Z.,&Balog, M. (2025). Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*.

Paglieri, D., Cross, L., Cunningham, W. A., Leibo, J. Z., &Vezhnevets, A. S. (2026). Persona Generators: Generating Diverse Synthetic Personas at Scale. *arXiv preprint arXiv:2602.03545*.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1-22).

Pearl, J. (2009). *Causality*. Cambridge university press.

Poole-Dayan, E., Kessler, D. T., Chiou, H., Hughes, M., Lin, E. S., Ganz, M., & Roy, D. (2025). Applying Large Language Models to Characterize Public Narratives. *arXiv preprint arXiv:2511.13505*.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023, July). Whose opinions do language models reflect?. In *International conference on machine learning* (pp. 29971-30004). PMLR.

Wang, H., Zhou, Y., Du, B., Ai, Q., & Liu, Y. (2026). Parametric Social Identity Injection and Diversification in Public Opinion Simulation. *arXiv preprint arXiv:2603.16142*.

Xie, Y., Liang, L., Li, S., Lu, Y., Xiao, Z., Shi, M.,&Xie, Y. (2026). Evaluating the statistical realism of LLM-generated social science data. *Proceedings of the National Academy of Sciences*, *123*(19), e2538145123.

Yun, L., An, C., Wang, Z., Peng, L., & Shang, J. (2025). The price of format: Diversity collapse in llms. *arXiv preprint arXiv:2505.18949*.