When Can We Trust Experiments on Digital Twins? A Potential Outcomes Framework for Causal Inference with LLM Simulations*

Patryk Perkowski

Sy Syms School of Business, Yeshiva University

November 1, 2025

Click here for the latest version.

Abstract

Researchers are increasingly using large language models (LLMs) and digital twins – LLM-based simulations of individuals – to collect data and run experimental studies. Yet, causal inference using digital twins is complicated by (i) prediction bias from LLMs, (ii) prompting effects in how the treatment is administered to the LLM, and (iii) stochasticity in response generation. This paper develops a parsimonious framework rooted in the Rubin causal model for explaining how these challenges impact inference using digital twins. The framework models how LLMs generate potential outcomes using their training data, the prompting of treatment arms, and random noise in the output generation process. These components jointly determine the identification assumptions required for valid inference. Under constant treatment effects and an additive LLM data generating function, I show that inference is possible and derive the necessary and sufficient condition of equal prompting bias across treatment arms. However, even minor tweaks to the data-generating property of the LLM alter the identification assumptions required for valid inference. These results highlight that valid inference with digital twins depends on the data-generating properties of the LLM such that identical research designs may require different identification assumptions depending on the LLM. Overall, this manuscript provides a formal foundation for understanding whether and when AI-based simulations can serve as substitutes for human subjects, or when they still need us in the loop.

Keywords: Generative AI, Digital Twins, Causal Inference, Potential Outcomes

*Patryk Perkowski (patryk.perkowski@yu.edu).

1 Introduction

Researchers are increasingly training large language models (LLMs) to serve as substitutes for human subjects in experimental data collection and analysis. Pioneered by Horton (2023)'s work showing that LLMs can simulate human behavior in economic scenarios, this approach has evolved from using off-the-shelf LLMs to creating and experimenting on "digital twins", LLM-based replicas of study participants trained on individual-level response data (Park et al., 2024; Peng et al., 2025). This methodology is expanding across both academia and industry. For the former, researchers have evaluated how well digital twins can do strategy (Choudhury et al., 2025), replicate experiments in psychology and management (Cui et al., 2025), and predict voting behavior of the Federal Reserve's Federal Open Market Committee (Kazinnik and Sinclair, 2025). For the latter, new startups offer services that allow firms to simulate the behavior of their stakeholders and test key strategic decisions prior to launch. Expected Parrot, a member of the 2025 Y-Combinator class, describes its mission as "We help companies simulate their stakeholders and explore pricing, product, marketing, communications and other high-impact business decisions at scale."[1] Together, these developments signal the emergence of digital-twin experimentation as a new paradigm for experimental data collection, one that blurs the line between human-subjects research and computation.

The potential benefits of this approach are substantial. First, digital twins may relax the Fundamental Problem of Causal Inference (Holland, 1986). While human experiments can only observe one potential outcome per participant, digital twins can generate multiple potential outcomes for a given subject through LLM prompting. Second, data collection via digital twins is cheaper, more efficient, and benefits substantially from economies of scale compared to collecting data from human subjects. Third, experimentation on digital twins expands the feasible set of research studies, potentially allowing studies that are restricted by academic norms or regulations (like the norms against deception in experimental economics (Charness et al., 2022)) to be run safely. Experimentation on digital twins allows researchers and firms alike to bypass these constraints and run studies more efficiently and ethically.

Despite these benefits, causal inference using digital twins is complicated by several data-generating features of LLMs. First, there is the challenge of prediction bias, where the response generated by a LLM deviates systematically from a participant's true response (Qu and Wang, 2024; Peng et

[1] See https://www.ycombinator.com/companies/expected-parrot.

al., 2025).[2] Biased predictions of individual's potential outcomes poses identification challenges for causal inference: prediction inaccuracy may prevent the unbiased estimation of potential outcomes and downstream treatment effects. Second, there is the challenge of prompt bias, where seemingly minor differences in the formulation of a prompt may systematically distort the LLM's output in a way that does not reflect substantial differences in the treatment arms (Carlson and Burbano, 2025).[3] Prompt bias makes it difficult to disentangle whether an observed treatment effect is due to the actual treatment, or prompting choices made by the researcher. A third challenge is the non-deterministic nature of LLM output generation, whereby the same prompt returns different output responses because of the sampling of probability distributions and choices of the sampling algorithm. This leads to variation in output responses even conditional on the same LLM prompt, complicating inference and the calculation of standard errors. Given these challenges, prior work has found that digital twins can reproduce treatment effects in some instances but not others, yet without a formal framework for when or why this occurs.

In this manuscript I apply the Neyman-Rubin potential outcomes framework, a workhorse model for causal inference (Rubin, 1974; Gerber and Green, 2012; Imbens and Rubin, 2015) to begin to address these challenges. I consider the case where a researcher seeks to estimate a treatment effect on a human population but cannot do so because of cost, ethics, or other feasibility concerns. Instead, she trains digital twins on the sample of interest, using methods like in Park et al. (2024) and Peng et al. (2025). She then uses the responses from these digital twins to simulate the potential outcomes of human subjects under all treatment conditions, and uses this data to estimate the estimand of interest in the digital twin sample. I investigate whether and under which conditions she can use digital twins to make claims about the estimand of interest in the human sample.

I formalize this problem by modeling how LLMs generate potential outcomes as a function of the challenges identified above. I specify various data-generating processes for LLMs and show how to derive the necessary identification conditions depending on the researcher's estimand of interest. The

[2]Part of this prediction inaccuracy can reflect group-bias, where LLMs are less accurate in predicting responses for specific subgroups. For example, Peng et al. (2025) show that digital twin predictions are more accurate for respondents who are more educated, higher income, and ideologically moderate. Qu and Wang (2024) find that LLMs are more accurate in predicting public opinion responses in Western countries, English-speaking ones, and developed nations. For a review of the literature on bias and fairness in LLMs, see Gallegos et al. (2024).

[3]Prompt bias can arise from framing effects, where differences in the wording or valence of a prompt influence the LLM's output (Salinas and Morstatter, 2024), or from prompt artifact effects, where the ordering, labeling, or formatting of options influence outcomes (Brucks and Toubia, 2025). Salinas and Morstatter (2024) show how minor perturbations to a prompt, such as an additional space at the end, can change the LLM's output, while Brucks and Toubia (2025) show evidence that the ordering and labeling of responses influences LLM choices.

key insight is that valid inference depends on the functional form of the LLM's data-generating process such that different specifications (and different LLMs) will require fundamentally different identification assumptions even conditional on the same study design.

I first consider a baseline case where the LLM generates potential outcomes as the sum of the true (human) potential outcome, participant-specific prediction bias, treatment-specific prompting bias, and noise. The estimand of interest for the researcher is the average treatment effect in the human population, and I begin with the case of constant treatment effects in the human population.

In this scenario, the simulated digital twin estimator can return an unbiased and consistent estimate of the average treatment effect in the human sample under additional assumptions. Valid inference here requires the prompt bias to be equal across treatment arms, an assumption I call the Treatment-Invariant Simulation Assumption (TISA). TISA ensures that prompting bias is equal across experimental conditions, and the prompt does not lead to differential changes in subject's potential outcomes across experimental conditions. Prediction bias, on the other hand, cancels out when estimating the average treatment effect under a constant treatment effect in the human sample.[4]

I next show that even minor changes to the data generating properties of the LLM lead to important modifications in the identification assumptions required for valid inference. I consider the case where the prediction bias and the prompting bias interact in the LLM's data-generating process. Here, certain types of participants are more or less sensitive to prompt bias in the LLM's data-generating function. This contrasts to the previous case where all participants experience the same prompt bias. The necessary identification assumptions here require a new assumption– the Interaction-Invariant Simulation Assumption (IIISA). Valid inference here requires zero covariance (across the population) between individual-level simulation bias and the difference in prompting bias between treatment arms. In other words, the LLM's tendency to over or under-estimate one's potential outcome must be on average independent of how the LLM interprets the treatment prompts.

What impact would this have on identification? If a researcher assumes TISA when the true data-generating property is interactive, the resulting estimator will generally be biased because it ignores the product of the prediction bias and the difference in prompt bias across treatment arms. On the other hand, if a researcher assumes IISA when the true data-generating property is additive, estimation

[4]When the researcher is interested in heterogeneous treatment effects and the estimand of interest is the conditional average treatment effect, an additional assumption is required. In addition to TISA, it must be the case that the simulation noise is independent conditional on the covariate of interest. In the case of constant treatment effects, the looser assumption of zero conditional mean error is required.

remains unbiased but the unnecessary interaction term introduces noise and increases variance in the estimator. This asymmetry shows the importance of aligning the identification assumptions with the data-generating properties of the LLM.

Ultimately, deriving the identification assumptions for all potential data generating properties of LLMs is beyond the scope of this manuscript. The true data-generating process of LLMs is complex and often described as a "black box"[5]. Ongoing work on AI explainability seeks to uncover the mechanisms behind LLM response generation (see, for example, Bilal et al. (2025) and Luo and Specia (2024)). Moreover, as LLMs further evolve, their data generating properties will change, which will require new identification assumptions and empirical tests.

My goal here is to provide researchers and practitioners a principled way to think about causal inference for LLM-based experiments. I use tractable functional forms that capture the key sources of bias in current LLM models. Although these assumptions are restrictive, they serve two complementary goals. First, they help illustrate how current biases from the expansive literature on LLMs influence causal inference. Second, they provide researchers with a formal framework for reasoning about causal inference using digital twins.

The results highlight both opportunities and tensions for using digital twins for causal inference. Valid LLM-based causal inference requires characterizing the data-generating properties of the LLM and deriving identification assumptions from that function. Otherwise, LLM-based experiments risk conflating the true treatment effect with artifacts from how the LLM generates potential outcomes. The identification assumptions are specific to the data-generating properties of the LLMs, and the same experimental prompts with the same twin training data can require different identification assumptions depending on the LLM used. This framework thus serves as both a guide to practitioners and as a caution about the hidden complexities of using LLMs and digital twins for causal inference. This manuscript offers a shared foundation rooted in the Rubin framework for thinking about these and related challenges.

This manuscript contributes primarily to the nascent literature using digital twins as experimental subjects. A growing body of empirical papers have examined how well digital twins and LLMs more generally reproduce human responses (Park et al., 2024; Peng et al., 2025; Choudhury et al., 2025; Cui et al., 2025; Kazinnik and Sinclair, 2025). The majority of this work is empirical, and I add a formal

[5] https://www.wired.com/story/anthropic-black-box-ai-research-neurons-features

orientation to this question. In related work, Gui and Toubia (2025) show that blinding in digital twin experiments violates the unconfoundedness assumption for experiments. While their focus is on the design stage of experiments, I focus here on inference and derive identification assumptions from various LLM data-generating processes. Both, however, are complementary to a more formal approach for digital twin experimentation.

This manuscript also contributes to the expansive literature on empirical methods for causal inference. Since LaLonde (1986)'s classic study comparing experimental versus observational estimators, researchers have introduced new tools for separating correlation from causation. While LaLonde (1986) asked when observational methods reproduce the same treatment effects as experimental studies, I ask a related question: when do experimental studies on digital twins reproduce the same treatment effects as experimental studies on human participants? This manuscript is also related to the cross-disciplinary literature on using human twins for estimating causal effects (Ashenfelter and Krueger, 1994; McGue et al., 2010; McAdams et al., 2021)[6] and also for using surrogate outcomes to proxy for longer-term treatment effects (Athey et al., 2019)[7] Experimentation on digital twins builds upon the experimental tradition in empirical social science research, the twin-based approach to confounding, and the surrogate index tradition in econometrics.

2 Inference under an additive data-generating function

I begin by formalizing the human ground truth using a potential outcomes framework. My goal in this section is to understand the conditions under which digital twin simulations from large language models can recover the average treatment effect for human subjects. To begin I assume constant treatment effects, but relax this assumption later in the section.

Let $Y_i(t)$ be the potential outcome of (human) subject i under treatment assignment $t \in \{0, 1\}$. I assume a simple additive structure for potential outcomes:

$$Y_i(t) = \theta_i + t * \tau + \epsilon_i(t)$$

[6]Since human twins allow researchers to "control" for unobservable differences in genetics, they present one approach for reducing unobserved heterogeneity when random assignment is not possible.

[7]In the digital twins set-up, digital twin act as surrogates for their human counterpart.

Here, θ_i represents stable, individual-specific characteristics (such as economic preferences or demographic characteristics) that are unaffected by treatment and are fixed across potential outcomes. τ denotes the constant causal effect of treatment. In the baseline case, I assume that the treatment affects all subjects equally so that $\tau_i = Y_i(1) - Y_i(0) = \tau$ for all i. $\epsilon_i(t)$ is an idiosyncratic error term with mean zero. This set-up implies that a subject's potential outcome is given by their baseline personality or preferences (θ_i), the true impact of treatment (τ), and random fluctuations ($\epsilon_{i,}(t)$).

The estimand of interest is the average treatment effect, $\tau = E_i[Y_i(1) - Y_i(0)]$. However, the researcher cannot directly estimate τ because she does not observe either potential outcome for (human) subjects. Thus, estimating τ requires simulating both counterfactuals using a digital twin.

To assuage this problem, the researcher seeks to simulate human behavior using a large language model (LLM). For each subject i, the researcher creates a digital twin that produces simulated potential outcomes under each treatment condition. Let $\hat{Y}_i(t)$ measure the simulated outcome for participant i under treatment $t \in \{0, 1\}$. I model the simulated outcome as a noisy approximation of the participant's true potential outcome:

$$\hat{Y}_i(t) = Y_i(t) + \eta_i + \beta_t + \xi_i(t). \tag{1}$$

The simulated outcome includes two types of biases that may arise from the LLM's simulation process. First, η_i captures an individual-specific simulation bias that is constant across treatment arms. η_i arises because the LLM may systematically over- or under-estimate potential outcomes for some subjects. For example, Peng et al. (2025) show that digital twin predictions are more accurate for respondents who are more educated, higher income, and ideologically moderate. These biases can reflect disparities in the LLM's training data and fine-tuning process: LLMs trained mostly on text produced by certain demographic groups may encode more accurate representations for those groups. Such differences across participants would be captured in the individual-specific simulation bias term, η_i.

The second type of simulation bias, β_t, arises from how the treatment is prompted to the LLM. β_t captures a treatment-specific simulation bias that captures systematic distortions in how the LLM interprets the treatment prompt itself. This can be due to prompt framing effects, where the valence of a prompt influences the LLM's output, or from prompt artifact effects, where the LLM's output is influenced by the ordering, labeling, or formatting of options. Such effects are captured in β_t. This

treatment-specific simulation bias mimics demand or framing effects in human experiments, where differences in the presentation of treatment arms can systematically impact responses independent of the actual treatment itself.

Lastly, $\xi_i(t)$ captures random noise in the LLM's simulation of individual i's responses to treatment t. This component captures random variation in the LLM's output that is not driven by the individual's traits or systematic prompt effects. Instead, this variation arises due to randomness in the model's output generation process. I assume this random noise is mean zero.

The researcher now can estimate the average treatment effect using simulated outcomes from digital twins:

$$\hat{\tau}^{LLM} = E_i[\hat{Y}_i(1) - \hat{Y}_i(0)]. \tag{2}$$

Substituting the simulation model from equation 1, we get:

$$\begin{aligned}\hat{\tau}^{LLM} &= E_i\big[Y_i(1) - Y_i(0) + (\beta_1 - \beta_0) + (\xi_i(1) - \xi_i(0))\big] \\ &= \underbrace{E_i[Y_i(1) - Y_i(0)]}_{\tau} + \underbrace{E_i[\beta_1 - \beta_0]}_{\text{Prompt bias}} + \underbrace{E_i[\xi_i(1) - \xi_i(0)]}_{\text{Simulation noise}}.\end{aligned} \tag{3}$$

Under what conditions will the estimator in equation 3 return an unbiased estimate of the true average treatment effect in the human population? The expected value of the estimator is:

$$E[\hat{\tau}_i^{LLM}] = \tau + (\beta_1 - \beta_0) + E_i[\xi_i(1) - \xi_i(0)]. \tag{4}$$

Since $E_i[\xi_i(1)] = E_i[\xi_i(0)] = 0$, this simplifies to:

$$E[\hat{\tau}^{LLM}] = \tau + (\beta_1 - \beta_0). \tag{5}$$

The simulated estimator $\hat{\tau}^{LLM}$ is unbiased for τ if and only $\beta_1 = \beta_0$, which is the required assumption for valid inference.

Treatment-Invariant Simulation Assumption (TISA): the LLM does not systematically favor one treatment condition over the other in its simulation of potential outcomes. In

other words, $\beta_1 = \beta_0$.

TISA helps ensure that the prompting of the LLM does not favor one treatment condition over the other. This assumption has a natural analog to causal inference in humans, where measurement error cannot be correlated with treatment assignment. Here, the measurement error specifically comes from prompting bias. When TISA does not hold, the simulated ATE is a combination of both the true human ATE plus differences in how the prompts were delivered. Moreover, this assumption requires that researchers exercise care and caution when selecting treatment prompts, which I explore in a future section.

I next characterize the sampling variability of the simulated estimator $\hat{\tau}^{LLM}$. The precision of $\hat{\tau}^{LLM}$ depends on the idiosyncratic noise $\xi_i(t)$ that is generated by the LLM. Under TISA, the estimator becomes:

$$\hat{\tau}^{LLM} = E_i[Y_i(1) - Y_i(0) + \xi_i(1) - \xi_i(0)]. \tag{6}$$

Assume that for each individual i, the simulation noise $\xi_i(1)$ and $\xi_i(0)$ are independent, mean-zero, and have finite variances σ_1^2 and σ_0^2, respectively. Then the variance of $\hat{\tau}^{LLM}$is given by:

$$Var(\hat{\tau}^{LLM}) = \frac{1}{n}[\sigma_1^2 + \sigma_0^2]. \tag{7}$$

As N increases, so does the precision of the estimator.[8]

Overall, these results indicate that when LLMs generate potential outcomes in an additive way, causal inference is possible. When the estimand of interest is the average treatment effect under constant effects, TISA and assumptions regarding simulation noise permit valid inference. In Appendix section A.1, I extend this model to account for heterogeneous treatment effects and show that two conditions are required for identification of the conditional average treatment effect. First, TISA must hold. Second, the noise must be mean-zero conditional on the desired covariate profile. In other words, the LLM should not simulate different levels of noise for subgroups, say for men versus women. These are natural extensions of the constant treatment effects case, and researchers can estimate both the average treatment effect and heterogeneous treatment effects under such assumptions. The core is an

[8]This implies the consistency of the estimator. Under TISA and the assumptions on the simulation noise, $\hat{\tau}^{LLM} \xrightarrow{p} \tau$ by the law of large numbers, and so $\hat{\tau}^{LLM}$ is a consistent estimator of τ. By the central limit theorem, $\sqrt{n}(\hat{\tau}^{LLM} - \tau) \xrightarrow{d} N(0, \sigma^2)$, where $\sigma^2 = Var(\xi_i(1) - \xi_i(0)) = \sigma_1^2 + \sigma_0^2$.

additive data-generating function of the LLM.

3 Inference under a data-generating process with interactive biases

In this previous section, I derive the identification assumptions required when the LLM's data generating process is the sum of the subject's true potential outcome, both bias terms, and the noise term. In this section, I consider the case where the prediction bias and the prompting bias interact in the LLM's data-generating process. Here, the magnitude of the LLM's response to a given prompt is not constant across the subject pool; instead, it depends on the underlying characteristics of the participant. In other words, certain types of participants are more or less sensitive to prompt bias in the LLM's data-generating function.

More formally, the LLM now generates potential outcomes in the following way:

$$\hat{Y}_i(t) = Y_i(t) + \eta_i * \beta_t + \xi_i(t). \tag{8}$$

The researcher generates both potential outcomes using equation 8 and estimates the average treatment effect using the difference-in-means estimator $\hat{\tau}^{LLM} = E_i[\hat{Y}_i(1) - \hat{Y}_i(0)]$.

Substituting the simulation model from equation 8 gives us:

$$\begin{aligned}
\hat{\tau}^{LLM} &= E_i\big[\hat{Y}_i(1) - \hat{Y}_i(0)\big] \\
&= E_i\big[Y_i(1) + \eta_i\beta_1 + \xi_i(1) - Y_i(0) - \eta_i\beta_0 - \xi_i(0)\big] \\
&= \underbrace{E_i[Y_i(1) - Y_i(0)]}_{\tau} + \underbrace{E_i[\eta_i(\beta_1 - \beta_0)]}_{\text{Interaction bias}} + \underbrace{E_i[\xi_i(1) - \xi_i(0)]}_{\text{Simulation noise}}.
\end{aligned} \tag{9}$$

Under the assumption of mean-zero noise across treatment arms ($E_i[\xi_i(1) - \xi_i(0)] = 0$), the expected value of $\hat{\tau}^{LLM}$ simplifies to:

$$E[\hat{\tau}^{LLM}] = \tau + E[\eta_i(\beta_1 - \beta_0)]. \tag{10}$$

Equation 10 implies that the LLM-based estimate of the average treatment effect will be unbiased for the true average treatment effect τ if and only if the interaction term between the individual-level simulation bias and the treatment-specific prompting bias has mean zero:

$$E[\eta_i(\beta_1 - \beta_0)] = 0. \tag{11}$$

This gives the identification assumption for constant treatment effects under an interactive LLM data-generating process:

> **Interaction-Invariant Simulation Assumption (IISA)**: On average, the interaction between individual-level simulation bias and the difference in prompting bias between treatment arms in the subject population is zero. In other words, $\text{Cov}(\eta_i,\ \beta_1 - \beta_0) = 0$.

Under ISSA, the digital twin's prediction bias is not systematically related to the difference in prompt effects across treatment arms. While some participants may experience large interactions between prediction bias and prompting bias, the extent of these interactions does not covary with the magnitude or direction of prompt-induced biases.[9] In other words, the LLM's tendency to over- or under-estimate one's potential outcome must be independent, on average, of how it responds to differences in the treatment prompts.[10]

I next characterize the sampling variability of the simulated estimator $\hat{\tau}^{LLM}$ under IISA. The precision of $\hat{\tau}^{LLM}$ depends not only on the idiosyncratic noise $\xi_i(t)$ generated by the LLM, but also on the heterogeneity in the interaction term $\eta_i(\beta_1 - \beta_0)$. Substituting the data-generating function from Equation 8, the estimator can be expressed as:

$$\hat{\tau}^{LLM} = E_i\big[Y_i(1) - Y_i(0) + \eta_i(\beta_1 - \beta_0) + \xi_i(1) - \xi_i(0)\big].$$

Assume that the simulation noise $\xi_i(1)$ and $\xi_i(0)$ are independent, mean-zero, and have finite variances σ_1^2 and σ_0^2, respectively. Further assume that the interaction term $\eta_i(\beta_1 - \beta_0)$ is mean-zero under IISA,

[9] While the LLM's data generating function allows interactions between the prediction and prompting biases, ISSA restricts this relationship at the population level.

[10] IISA is less restrictive than TISA. Under TISA, we assume prompt biases are identical across treatment arms (i.e., $\beta_1 = \beta_0$). Under ISSA, prompt biases may differ across treatment conditions and even interact with individual-level prediction bias, but these interactions must cancel out in expectation.

and has variance $\sigma^2_{\eta\beta}$. Then the sampling variance of $\hat{\tau}^{LLM}$ is given by:

$$Var(\hat{\tau}^{LLM}) = \frac{1}{n}\big[Var(Y_i(1) - Y_i(0)) + \sigma^2_{\eta\beta} + \sigma^2_1 + \sigma^2_0\big].$$

As n increases, the precision of the estimator increases and the sampling variance converges to zero.[11]

In Appendix section A.2, I derive the identification assumptions for heterogeneous treatment effects under this data-generating process. Identification here requires (i) IISA at the subgroup level (i.e., $E[\eta_i(\beta_1 - \beta_0) \mid X_i = x] = 0$), and (ii) the same conditional mean-zero noise restrictions as in the case of an additive data generating process (i.e., $E[\xi_i(1) - \xi_i(0) \mid X_i = x] = 0$).

4 Conclusion

The debate on the future of work and AI substitution has now reached human subjects in academic and business research. While experimentation with digital twins allows a way to bypass the Fundamental Problem of Causal Inference, it introduces new prediction and prompting biases that complicate inference. In this paper, I develop a formal framework to begin to address these challenges. Modeling the data-generating properties of the LLM allows researchers to derive the necessary identification assumptions. My results help to clarify the conditions under which experimentation via digital twins will recover true treatment effects in human subjects. This paper represents a first step to a more formal orientation on this timely topic. Future research can extend this both methodologically (for example, new estimators driven by more sophisticated data-generating functions) and empirically (for example, empirical tests of the identifying assumptions using placebo tests and sensitivity analysis). Only by grounding digital-twin applications in a formal causal framework can we determine whether AI can substitute for human participants in scholarly and business research.

[11] Under IISA and the stated assumptions on the noise and interaction terms, $\hat{\tau}^{LLM} \xrightarrow{p} \tau$ by the law of large numbers, implying that $\hat{\tau}^{LLM}$ is a consistent estimator of τ. By the central limit theorem, $\sqrt{n}(\hat{\tau}^{LLM} - \tau) \xrightarrow{d} N(0, \sigma^2)$, where $\sigma^2 = Var(Y_i(1) - Y_i(0)) + \sigma^2_{\eta\beta} + \sigma^2_1 + \sigma^2_0$.

References

Ashenfelter, Orley and Alan Krueger, "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review*, 1994, *84* (5), 1157–1173.

Athey, Susan, Raj Chetty, Guido W. Imbens, and Hyunseung Kang, "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," NBER Working Paper 26463, National Bureau of Economic Research 2019.

Bilal, Ahsan, David Ebert, and Beiyu Lin, "LLMs for Explainable AI: A Comprehensive Survey," Technical Report 2025.

Brucks, Melanie and Olivier Toubia, "Prompt architecture induces methodological artifacts in large language models," *PLOS ONE*, 2025, *20* (4), e0319159.

Carlson, Natalie and Vanessa Burbano, "The Use of LLMs to Annotate Data in Management Research: Foundational Guidelines and Warnings," *The Wharton School Research Paper*, September 3 2025. Forthcoming at *Strategic Management Journal*.

Charness, Gary, Anya Samek, and Jeroen van de Ven, "What is considered deception in experimental economics?," *Experimental Economics*, 2022, *25* (2), 385–412.

Choudhury, Arnab, Bruce Kogut, and Patryk Perkowski, "Strategic Agents," 2025. Work in progress.

Cui, Ziyan, Ning Li, and Huaikang Zhou, "A Large-Scale Replication of Scenario-Based Experiments in Psychology and Management Using Large Language Models," *Nature Computational Science*, 2025, pp. 627—-634.

Gallegos, Ignacio O., Ryan A. Rossi, Joshua Barrow, Md Mahfuzur Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Rui Zhang, and Nesreen K. Ahmed, "Bias and fairness in large language models: A survey," *Computational Linguistics*, 2024, *50* (3), 1097–1179.

Gerber, Alan S. and Donald P. Green, *Field Experiments: Design, Analysis, and Interpretation*, New York: W. W. Norton Company, 2012.

Gui, George and Olivier Toubia, "The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective," *arXiv preprint*, 2025.

Holland, Paul W., "Statistics and causal inference," *Journal of the American Statistical Association*, 1986, *81* (396), 945–960.

Horton, John J., "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?," NBER Working Paper 31122, National Bureau of Economic Research April 2023.

Imbens, Guido W. and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press, 2015.

Kazinnik, Sophia and Tara Sinclair, "FOMC In Silico: A Multi-Agent System for Monetary Policy Modeling," 2025. Working Paper.

LaLonde, Robert J., "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 1986, *76* (4), 604–620.

Luo, Haoyan and Lucia Specia, "From Understanding to Utilization: A Survey on Explainability for Large Language Models," 2024.

McAdams, Tom A., Fruhling V. Rijsdijk, Helena M. S. Zavos, and Jean-Baptiste Pingault, "Twins and Causal Inference: Leveraging Nature's Experiment," *Cold Spring Harbor Perspectives in Medicine*, 2021, *11* (6), a039552.

McGue, Matt, Merete Osler, and Kaare Christensen, "Causal Inference and Observational Research: The Utility of Twins," *Perspectives on Psychological Science*, 2010, *5* (5), 546–556.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein, "Generative Agent Simulations of 1,000 People," 2024.

Peng, Tianyi, George Gui, Daniel J. Merlau, Grace Jiarui Fan, Malek Ben Sliman, Melanie Brucks, Eric J. Johnson, Vicki Morwitz, Abdullah Althenayyan, Silvia Bellezza, Dante Donati, Hortense Fong, Elizabeth Friedman, Ariana Guevara, Mohamed Hussein, Kinshuk Jerath, Bruce Kogut, Kristen Lane, Hannah Li, Patryk Perkowski, Oded Netzer, and Olivier Toubia, "A Mega-Study of Digital Twins Reveals Strengths, Weaknesses and Opportunities for Further Improvement," 2025. Working Paper.

Qu, Y. and J. Wang, "Performance and biases of Large Language Models in public opinion simulation," *Humanities and Social Sciences Communications*, 2024, *11* (1), 1095.

Rubin, Donald B., "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 1974, *66* (5), 688–701.

Salinas, Abel and Fred Morstatter, "The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance," 2024.

Appendix: For Online Publication Only

A Additional derivations

A.1 Identification of heterogeneous treatment effects under an additive data generating function

I extend the model in section 2 to account for heterogeneous treatment effects. In many settings, researchers are interested in not just estimating the overall average treatment effect, but also how the average treatment effect varies across different types of participants. In this section, I derive the conditions under which experimentation on digital twins can be used to estimate conditional average treatment effects.

Let $Y_i(t)$ be the potential outcome of (human) subject i under treatment assignment $t \in \{0, 1\}$. I now model potential outcomes as:

$$Y_i(t) = \theta_i + t * \tau_i + \epsilon_i(t) \tag{12}$$

Here, τ_i represents the individual-specific treatment effect for subject i $(= Y_i(1) - Y_i(0))$, which can vary across subjects.

While the researcher may still be interested in estimating the ATE, $\tau = E_i[\tau_i]$, she may also seek to estimate the treatment effect for individuals with some common characteristics. In this case, the estimand of interest is the conditional average treatment effect (CATE), denoted by $\tau(x) = E[Y_i(1) - Y_i(0) \mid X_i = x] = E_i[\tau_i \mid X_i = x]$, where, X_i is a vector of covariates for a participant. The researcher uses an LLM and simulates both potential outcomes for each individual using the same structure as in section 2, such that $\hat{Y}_i(t) = Y_i(t) + \eta_i + \beta_t + \xi_i(t)$.

The research then uses the simulated CATE estimator:

$$\hat{\tau}^{LLM}(x) = E_i[\hat{Y}_i(1) - \hat{Y}_i(0) \mid X_i = x]. \tag{13}$$

Substituting the simulation model from equation 1 and simplifying, we get:

$$\hat{\tau}^{LLM}(x) = \tau(x) + (\beta_1 - \beta_0) + E[\xi_i(1) - \xi_i(0) \mid X_i = x] \tag{14}$$

The estimator $\hat{\tau}^{LLM}(x)$ will return an unbiased estimate of the true CATE $\tau(x)$ if two conditions are met. First, TISA is required $(\beta_1 - \beta_0)$. Second, the simulation noise must be conditionally mean-zero: $E[\xi_i(1) - \xi_i(0) \mid X_i = x] = 0$.This is a stronger assumption than in the case of estimating the average treatment effect in section 2. To see why, recall that in the constant treatment effect case, we have $E[\xi_i(1) - \xi_i(0)] = 0$ such that the simulation noise cancels out on average across the entire sample. But that is not to say for any specific subgroup with profile $X_i = x$ that the errors must be zero. This assumption helps to ensure that the LLM does not simulate different levels of noise across the subgroups.

The sampling variability of the simulated CATE estimator $\hat{\tau}^{LLM}(x)$ is as follows. Under TISA and conditionally mean-zero simulation noise, the estimator becomes: $\hat{\tau}^{LLM}(x) = \tau(x) + E[\xi_i(1) - \xi_i(0) \mid X_i]$. Assume that for each individual i, the simulation noise $\xi_i(1)$ and $\xi_i(0)$ are independent conditional on X, conditionally mean-zero, and have finite conditional variances $\sigma_1^2(x)$ and $\sigma_0^2(x)$, respectively. Moreover, let n_x be the number of participants with $X_i = x$. The variance of $\hat{\tau}^{LLM}(x)$ is given by: $VAR(\hat{\tau}^{LLM}(x)) = \frac{1}{n_x}[\sigma_1^2(x) + \sigma_0^2(x)]$. The simulated estimator's precision increases as n_x increases. Moreover, this implies the consistency of the simulated estimator. Under TISA and the assumptions on the simulation noise, $\hat{\tau}^{LLM}(x) \xrightarrow{p} \tau(x)$ by the law of large numbers, and so $\hat{\tau}^{LLM}(x)$ is a consistent estimator of $\tau(x)$. By the central limit theorem, $\sqrt{n_x}(\hat{\tau}^{LLM}(x) - \tau(x)) \xrightarrow{d} N(0, \sigma^2(x))$, where $\sigma^2(x) = Var(\xi_i(1) - \xi_i(0) \mid X_i = x) = \sigma_1^2(x) + \sigma_0^2(x)$. This variance $o^2(x)$ represents simulation-based uncertainty from the LLM and captures sampling variability from how the LLM generates responses conditional on $X_i = x$.

A.2 Identification of heterogeneous treatment effects under an interactive data generating function

In this section, I derive the conditions under which experimentation on digital twins can be used to estimate conditional average treatment effects when the prediction bias and prompting bias interact in the LLM's data-generating function, as in equation 8.

The CATE is given by $\tau(x) = E[Y_i(1) - Y_i(0) \mid X_i = x] = E[\tau_i \mid X_i = x]$, where X_i is a vector of

covariates for participant i.

The LLM generates simulated potential outcomes according to the interactive data-generating process:

$$\hat{Y}_i(t) = Y_i(t) + \eta_i \beta_t + \xi_i(t). \tag{15}$$

The simulated CATE estimator is then given by:

$$\hat{\tau}^{LLM}(x) = E_i[\hat{Y}_i(1) - \hat{Y}_i(0) \mid X_i = x]. \tag{16}$$

Substituting the data-generating process from Equation 15, we obtain:

$$\begin{aligned} \hat{\tau}^{LLM}(x) &= E_i\big[Y_i(1) - Y_i(0) + \eta_i(\beta_1 - \beta_0) + \xi_i(1) - \xi_i(0) \mid X_i = x\big] \\ &= \tau(x) + E[\eta_i(\beta_1 - \beta_0) \mid X_i = x] + E[\xi_i(1) - \xi_i(0) \mid X_i = x]. \end{aligned} \tag{17}$$

The estimator $\hat{\tau}^{LLM}(x)$ is unbiased for $\tau(x)$ if two conditions hold. First, *Interaction-Invariant Simulation Assumption (IISA)* must hold at the subgroup level, such that $E[\eta_i(\beta_1 - \beta_0) \mid X_i = x] = 0$. Second, the simulation noise must be conditionally mean-zero: $E[\xi_i(1) - \xi_i(0) \mid X_i = x] = 0$. Under these two conditions, $\hat{\tau}^{LLM}(x)$ identifies $\tau(x)$.

I next investigate the sampling variability of the simulated CATE estimator, Under conditional IISA and conditionally mean-zero simulation noise, we get:

$$\hat{\tau}^{LLM}(x) = \tau(x) + E[\xi_i(1) - \xi_i(0) \mid X_i = x].$$

Assume that the simulation noise terms $\xi_i(1)$ and $\xi_i(0)$ are independent conditional on X_i, conditionally mean-zero, and have finite variances $\sigma_1^2(x)$ and $\sigma_0^2(x)$, respectively. In addition, let the interaction term $\eta_i(\beta_1 - \beta_0)$ have finite conditional variance $\sigma_{\eta\beta}^2(x)$, and let n_x be the number of participants with $X_i = x$. Then the sampling variance of $\hat{\tau}^{LLM}(x)$ is given by:

$$Var(\hat{\tau}^{LLM}(x)) = \frac{1}{n_x}\big[Var(Y_i(1) - Y_i(0) \mid X_i = x) + \sigma_{\eta\beta}^2(x) + \sigma_1^2(x) + \sigma_0^2(x)\big].$$

As n_x increases, the estimator's precision increases and the sampling variance converges to zero.[1]

[1] Under IISA and the assumptions on the simulation noise and interaction variance, $\hat{\tau}^{LLM}(x) \xrightarrow{p} \tau(x)$ by the law of large numbers, and so $\hat{\tau}^{LLM}(x)$ is a consistent estimator of $\tau(x)$. By the central limit theorem, $\sqrt{n_x}(\hat{\tau}^{LLM}(x) - \tau(x)) \xrightarrow{d} N(0, \sigma^2(x))$, where $\sigma^2(x) = Var(Y_i(1) - Y_i(0) \mid X_i = x) + \sigma^2_{\eta\beta}(x) + \sigma^2_1(x) + \sigma^2_0(x)$.