

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

September 29, 2026

Abstract

Researchers are increasingly using large language models (LLMs) and digital twins—LLM-based simulations of individuals—to collect data and conduct experiments. Yet causal inference with digital twins is complicated by systematic prediction error, prompting effects in how treatments are administered to the LLM, and stochasticity in response generation. This paper develops a parsimonious framework, rooted in the Rubin causal model, for characterizing how these features affect causal inference with digital twins. Across three classes of data-generating processes, I derive the corresponding identification assumptions, testable restrictions that can distinguish among competing data-generating processes, and, when possible, bias-corrected estimators when the identifying assumptions fail. I further show that, even at equal sample sizes, digital-twin experiments can be more precise than human experiments when potential outcomes are highly correlated within individuals and simulation errors are highly correlated across treatment conditions. I illustrate the value of the framework using a digital-twin experiment on how hiring algorithms influence job seekers’ beliefs, showing how human validation data can be used to diagnose the simulator’s DGP, test the relevant identifying assumption, and adjust digital-twin treatment effects when that assumption fails. Overall, the paper provides a formal foundation for causal inference in LLM-based experiments and clarifies when digital twins can substitute for human subjects and when humans must remain in the loop.

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

^{*}Patryk Perkowski (patryk.perkowski@yu.edu) I thank Bruce Kogut for his invitation to participate in Columbia University’s Digital Twins Megastudy (Peng et al., 2025), which inspired this manuscript. For helpful feedback and discussions, I thank Assa Cohen, Bo Cowgill, Pablo Hernández-Lagos, and Jessica Hullman, as well as seminar and conference participants at Yeshiva University, Columbia’s Conference on Artificial Intelligence, Machine Learning, and Business Analytics, Northwestern’s Symposium on Validating Generative AI-Based Social Sciences, and USC’s AI in Management Conference. All errors are my own.

1 Introduction

The possibility of machine substitution has reached human subjects research. Researchers are increasingly training large language models (LLMs) to serve as substitutes for human subjects in survey data collection and in experimental studies. 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 participants trained on individual-level response data (Park et al., 2024; Peng et al., 2025).

This practice is growing in both academic research and in business practice. For the former, researchers have evaluated how well digital twins replicate experiments in psychology and management (Cui et al., 2025), predict voting behavior of the Federal Reserve’s Federal Open Market Committee (Kazinnik and Sinclair, 2025), and anticipate the market entry and pricing decisions of strategists (Choudhury et al., 2025). For the latter, both new and established firms offer services that allow customers 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 Similarly, Shopify has launched SimGym, a system that creates digital customers that behave like real (human) ones.² Together, these developments signal the emergence of LLM simulations as a new paradigm for experimental studies, one that blurs the line between human-subjects research and computation.

The benefits of using LLM simulations for “human”-subjects research 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

¹See <https://www.ycombinator.com/companies/expected-parrot>.

²See <https://help.shopify.com/en/manual/online-store/simgym>.

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 al., 2025).³ 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).⁴ 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

³Part 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).

⁴Prompt 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.

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 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.⁵

I then provide guidance on how to probe TISA in practice. I first show that it is impossible to test TISA using only LLM simulated data: any pattern in the simulated treatment effects can always be rationalized by either the true treatment effect or by asymmetric prompt bias. Instead, validation requires both LLM simulated data plus human data on actual potential outcomes. If the researcher observes potential outcomes for a random subset of participants, she can empirically estimate the prompt bias effect and thereby test TISA. Moreover, even if TISA does not hold, I derive a bias-corrected estimator that returns the true average treatment effect in the human sample.

While analytically convenient, the additive case imposes strong assumptions on the simulation behavior of LLMs. To that end, I consider two alternative classes of data-generating processes: one where individual prediction bias and prompting bias interact, and one where the LLM scales the

⁵When 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.

prompting bias by the level of potential outcome. I derive the identification conditions for each, show how to use human data to probe the assumption, and propose (where possible) a bias-corrected estimator if the identification assumption fails to hold.

I show that identification under an interactive data-generating process requires that individual-level prediction bias be uncorrelated with differences in prompting bias across treatment arms. While this assumption can be empirically tested by examining the relationship between baseline prediction error and the simulated average treatment effect, there is no bias-corrected estimator without additional modeling assumptions. In the case of the scaled data-generating process, identification requires that the ratio of treated and control potential outcomes must be equal to the ratio of control and treatment prompting bias. Human data allow the researcher to estimate these ratios directly, and a reweighting procedure produces unbiased treatment effects when the identification assumption does not hold.

Overall, these results show that deriving and testing identification conditions is possible in LLM-based experiments when the researcher knows the true data-generating process of the LLM. Moreover, identification assumptions are local to the given data-generating process, and different LLMs, study designs, and even outcome variables may require different identification conditions. How, then, can a researcher learn about the data-generating process in practice?

I show how human data allows the researcher to falsify amongst potential data-generating processes. Each data-generating process has testable restrictions on how the prediction error, or the difference in the LLM’s simulation of the potential outcome versus the true potential outcome, varies across participants, prompts, and treatment arms. For example, the additive case requires that the relative order of prediction errors must be the same across treatment and control prompts. The scaled version requires participants with larger potential outcomes experience larger prediction errors. Human data allows the researcher to evaluate and falsify potential data-generating processes based on their testable implications. I derive these tests for the three classes of data-generating processes considered.

Ultimately, deriving the identification assumptions for all potential data generating properties is beyond the scope of this manuscript. The true data-generating process of LLMs is complex and often described as a “black box”⁶. Ongoing work on AI explainability seeks to uncover the mechanisms that guide 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

⁶<https://www.wired.com/story/anthropic-black-box-ai-research-neurons-features>

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 the ability of LLMs to substitute for humans in experimental studies. Second, they provide researchers with a formal framework for reasoning about causal inference using digital twins and LLM simulations more broadly.

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 or the outcome of interest. 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 also speaks more broadly to the future of “human” subjects research and the relationship between human and synthetic data in experimental work. Human data collection is required not only to probe identification assumptions but also select among potential data-generating processes. LLM-based experiments are best understood as a division of labor between humans and machines: human data provides empirical grounding and validation while LLM data provides scale, speed, and precision once the appropriate causal structure has been established. To that end, it is not a question of whether LLMs will substitute for human subjects in research and business practice; instead, it is how the availability of digital twins enables new methodological innovations that unlock complementarities between the two.

Overall, this manuscript contributes to three related literatures. First, it contributes 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 results across these papers are mixed—some find that LLM simulations can reproduce treatment effects from human samples, while others do not. The majority of this work is empirical, and I add a formal orientation that allows us to understand the conditions under which such experiments will lead to unbiased estimates of the true treatment effect in human samples.

Second, I contribute to the methodological literature on large language models in human subjects research. Particularly relevant here is Ludwig et al. (2025), who develop a general econometric framework for incorporating LLM outputs into empirical analysis, with a primary focus on automated measurement and prediction. In their framework, the use of LLMs as simulated human subjects appears as a special case of measurement and requires simulated responses to reproduce human behavior on a given experiment or survey. My approach examines whether simulated outcomes preserve the counterfactual differences that identify treatment effects, even in cases where LLM predictions are systematically biased. Also related is Gui and Toubia (2025), who study the use of LLMs as simulated experimental subjects and highlight how prompt ambiguity can cause spurious confounding in LLM experiments. While their analysis focuses on diagnosing prompt failures and proposes design and prompting fixes, I characterize the general conditions under which simulation bias does or does not distort causal estimates, and clarify when such distortions can be detected or corrected using human data.

Lastly, this manuscript 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)⁷ and also for using surrogate outcomes to proxy for longer-term treatment effects (Athey et al., 2019)⁸ 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.

⁷Since 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.

⁸In the digital twins set-up, digital twin act as surrogates for their human counterpart.

2 Inference under an additive data-generating process

I begin with a benchmark case in which the LLM generates simulated outcomes through an additive data-generating process. This case is analytically transparent and provides a useful baseline for when and why experimentation via LLM simulations can recover true treatment effects in the underlying human sample. I assume constant treatment effects in the human sample but relax this assumption later in the section.

2.1 An additive model of LLM-generated outcomes

Let $Y_i(t)$ be the potential outcome of (human) subject i under treatment assignment $t \in \{0, 1\}$. I assume a constant-treatment-effects specification where potential outcomes are as follows:

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

Here, θ_i represents stable, participant-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. 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:

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

$$\hat{Y}_i(t) = f(Y_i(t), \alpha_i, \beta(t), \epsilon_i(t)). \quad (2)$$

$$\begin{aligned} \hat{\tau}^{LLM} &= E_i[\hat{Y}_i(1) - \hat{Y}_i(0)] \\ &= E_i[f(Y_i(1), \alpha_i, \beta(1), \epsilon_i(1)) - f(Y_i(0), \alpha_i, \beta(0), \epsilon_i(0))]. \end{aligned} \quad (3)$$

The simulated outcome includes two types of biases that may arise from the LLM’s simulation process. First, α_i captures an individual-specific simulation bias that is constant across treatment arms. α_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, α_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 noise arises due to randomness in the model’s output generation process. Throughout this section, I maintain the following assumption:

Assumption A1 (Mean-Zero Simulation Noise). The idiosyncratic simulation noise satisfies

$$E[\xi_i(t)] = 0 \quad \text{for each } t \in \{0, 1\}.$$

The researcher observes $\{\hat{Y}_i(0), \hat{Y}_i(1)\}$ for each subject and can now estimate the average treatment effect using simulated outcomes from digital twins:

$$\hat{\tau}^{LLM} = E_i[\hat{Y}_i(1) - \hat{Y}_i(0)]. \quad (4)$$

2.2 Identification of the average treatment effect

If we substitute the simulation model from equation 2, we get:

$$\begin{aligned} \hat{\tau}^{LLM} &= E_i[Y_i(1) - Y_i(0) + (\beta_1 - \beta_0) + (\xi_i(1) - \xi_i(0))] \\ &= \underbrace{E_i[Y_i(1) - Y_i(0)]}_{\tau} + \underbrace{E_i[\beta_1 - \beta_0]}_{\text{Prompt bias}} + \underbrace{E_i[\xi_i(1) - \xi_i(0)]}_{\text{Simulation noise}}. \end{aligned} \quad (5)$$

When will equation 5 return an unbiased estimate of the true average treatment effect in the human sample? The expected value of the estimator is:

$$E[\hat{\tau}_i^{LLM}] = \tau + (\beta_1 - \beta_0) + E_i[\xi_i(1) - \xi_i(0)]. \quad (6)$$

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

$$E[\hat{\tau}^{LLM}] = \tau + (\beta_1 - \beta_0). \quad (7)$$

The simulated estimator $\hat{\tau}^{LLM}$ is therefore unbiased for τ if and only $\beta_1 = \beta_0$, which motivates the identification assumption for the additive case:

Proposition 1 (Identification of the ATE under an additive data-generating process).

Suppose the LLM generates simulated outcomes according to $\hat{Y}_i(t) = Y_i(t) + \alpha_i + \beta_t + \xi_i(t)$ where $Y_i(t)$ denotes the true human potential outcome, α_i is the prediction bias for participant i , β_t is a treatment-specific simulation bias, and $\xi_i(t)$ is idiosyncratic simulation noise satisfying $E[\xi_i(t)] = 0$ for each t in $0, 1$. Then the simulated difference-in-means estimator $\hat{\tau}^{LLM} = E_i[\hat{Y}_i(1) - \hat{Y}_i(0)]$ is an unbiased estimator of the true average treatment effect $\tau = E[Y_i(1) - Y_i(0)]$ if and only if $\beta_1 = \beta_0$.

I refer to this condition as the **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 human experiments, where measurement error cannot be correlated with treatment assignment. The difference, though, is that measurement error here is introduced through prompting effects rather than noisy outcome measurement. When TISA does not hold, the simulated ATE will be a combination of both the true human ATE plus differences in how the prompts were delivered.

TISA helps clarify why predictive accuracy in LLM simulations is neither necessary nor sufficient for causal identification. While predictive accuracy focuses on measuring differences between the true and predicted values ($E\left[\left(\hat{Y}_i(t) - Y_i(t)\right)^2\right]$), identification relies upon differences in simulated potential outcomes ($E\left[\hat{Y}_i(1) - \hat{Y}_i(0)\right]$). An LLM may generate biased predictions of a subject’s potential outcomes but still satisfy TISA if the prompting bias is equal across treatment arms. Relatedly, an LLM that achieves high predictive accuracy on average may violate TISA if it processes treatment prompts asymmetrically. Causal inference in the additive case requires treatment prompt symmetry instead of than predictive accuracy.

2.3 Extention to heterogeneous treatment effects

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 additive data-generating function of the LLM.

2.4 Discussion

The results in this section show that causal inference using LLM simulations is in principle possible. When the LLM’s data-generating process is additive, introduces treatment-invariant bias, and has

mean zero noise, simulated experiments can recover unbiased estimates of both average and heterogeneous treatment effects from the human sample.

However, while identification with human experiments requires assumptions regarding randomization, identification here requires assumptions regarding how the LLM simulates data. The researcher cannot rely on standard experimental logic to justify TISA: random assignment of prompts does not ensure that the LLM processes those prompts symmetrically. Instead, she must examine the plausibility of TISA in a given experimental context, which motivates the following section.

3 Validation and bias correction in the additive case

In this section, I examine validation techniques for TISA in section 2. I distinguish between validation methods using only simulated LLM data (section 3.1) versus those with human data (section 3.2), and show how human data permits a bias correction under this framework (section ??).

3.1 Validation using only simulated LLM data

Suppose a researcher is interested in validating TISA without human data. Instead, she only has access to LLM-simulated data ($\hat{Y}_i(1)$ and $\hat{Y}_i(0)$). Can she use only this data to assess TISA? The answer is no.

To see why not, the simulated treatment effect can be written as:

$$\hat{Y}_i(1) - \hat{Y}_i(0) = (Y_i(1) - Y_i(0)) + (\beta_1 - \beta_0) + (\xi_i(1) - \xi_i(0)).$$

The researcher observes the left-hand side but never observes the true treatment effect in humans $Y_i(1) - Y_i(0)$. Therefore, any pattern in the simulated effects can always be rationalized by either the true treatment effect or by asymmetric prompting bias. To that end, probing the identification assumption here will require using both LLM and human data.

Although simulated LLM data cannot validate the identification assumptions, they are still useful in other ways for an experimentalist. For example, rerunning the same prompt multiple times can help the researcher estimate the variance of idiosyncratic simulation noise, which is helpful for power

calculations. However, it is impossible to validate TISA (or other identification assumptions) using only simulated data.

3.2 Validation using human data

Once human data is available, the identification challenge changes substantially: human data introduces direct information about the relationship between simulated and true potential outcomes. Even limited human data can help the researcher break the confounding between true causal effects and simulation bias.

Suppose the researcher observes human outcomes for a random subset of individuals, with some observed under treatment and others under control. She can then directly estimate treatment-specific simulation bias:

$$\beta_t = E\left[\hat{Y}_i(t) - Y_i(t)\right].$$

This has two important implications. First, human data allow the researcher to assess the plausibility of TISA by testing whether the estimated biases differ across treatment arms. Instead of assuming TISA, the researcher can probe whether the assumption holds in the given setting. Second, human data permits a bias-corrected estimator even if TISA falls. Thus, even if the researcher finds that TISA does not hold in their setting, they can apply a bias correction and recover the true average treatment effect.

3.3 Bias correction under an additive data-generating process

Under the additive data-generating process, human data allows the research to use a bias-corrected estimator when TISA is violated. This is done through the following steps:

Step 1 (Bias estimation). Using a validation sample with both human outcomes (some for treated potential outcomes and others for control) and simulated outcomes, the researcher estimates treatment-specific simulation bias:

$$\hat{\beta}_t = E_n\left[\hat{Y}_i(t) - Y_i(t)\right],$$

Step 2 (Outcome correction). The researcher constructs bias-corrected simulated outcomes:

$$\tilde{Y}_i(t) = \hat{Y}_i(t) - \hat{\beta}_t.$$

Substituting the additive simulation model gives:

$$\tilde{Y}_i(t) = Y_i(t) + (\beta_t - \hat{\beta}_t) + \xi_i(t).$$

Treatment effects can then be estimated using $\tilde{Y}_i(1) - \tilde{Y}_i(0)$.

Proposition 3 (Bias correction under an additive data-generating process). *Suppose simulated outcomes are generated according to*

$$\hat{Y}_i(t) = Y_i(t) + \alpha_i + \beta_t + \xi_i(t),$$

with $E[\xi_i(t)] = 0$. If β_t is consistently estimated using human outcome data and subtracted from simulated outcomes, then the resulting estimator of the ATE is unbiased for τ .

This result highlights an important distinction between identification and correction. In the absence of human data, TISA must be assumed to justify causal inference. With human data, the researcher can estimate prompting effects and apply a correction if TISA fails (provided the additive structure of the simulator’s data-generating process is correctly specified, which I turn to in section 5). As I show in section 4, however, this property is highly special: once bias varies across individuals, this debiasing procedure no longer works and bias correction may not be possible without stronger modeling assumptions.

4 Other data-generating processes for LLM simulations

Sections 2 and 3 derive the identification conditions and propose validation and bias-correction for an additive LLM data generating function. While analytically convenient, this model imposes strong restrictions on how the LLM simulates potential outcomes. For example, the additive case assumes a constant prompting effect that does not covary with participant characteristics or the level of the

potential outcome. Such restrictions are unlikely to hold in practice, as LLM responses are likely generated through nonlinear reasoning processes that depend on both the content of the prompt and the underlying attributes of the individual being simulated. While the additive model serves as a useful benchmark, this section considers two alternative data-generating processes for LLM-based simulations.

In this section, I begin with a meta-data-generating process. I then focus on two alternative data-generating processes. The first is an interactive model where the LLM generates potential outcomes and the individual prediction bias and treatment prompting effect interact; the second is one where the LLM scales the prompt effect by the level of the potential outcome. For each, I derive identification conditions, examine how human data helps with the identification problem, and propose (where possible) a de-biased estimator for that given data-generating process.

4.1 A meta-data-generating process for simulated potential outcomes

I model the simulated outcome generated by the digital twin of participant i under treatment $t \in \{0, 1\}$ as

$$\hat{Y}_i(t) = g(Y_i(t), \alpha_i, \beta_t, \xi_i(t)),$$

where $g(\cdot)$ is an unknown function characterizing the LLM’s simulation process. As previously, α_i captures individual-specific prediction bias, β_t captures treatment-specific prompting bias, and $\xi_i(t)$ captures idiosyncratic simulation noise. Different LLM models, training corpora, and prompt designs correspond to different functional forms for $g(\cdot)$. The researcher does not observe $g(\cdot)$ directly and must instead reason about it using a combination of modeling assumptions, human data, and diagnostic tools.

In this case, the simulated difference-in-means estimand can be written as

$$E[\hat{Y}_i(1) - \hat{Y}_i(0)] = E[g(Y_i(1), \alpha_i, \beta_1, \xi_i(1)) - g(Y_i(0), \alpha_i, \beta_0, \xi_i(0))].$$

Identification of the true average treatment effect $\tau = E[Y_i(1) - Y_i(0)]$ therefore requires restrictions on $g(\cdot)$ such that all terms involving α_i , β_t , and $\xi_i(t)$ either cancel in expectation or can be estimated and removed using human data. Without additional structure on $g(\cdot)$, we cannot identify the causal

estimand: for any observed distribution of simulated outcomes, there exist multiple combinations of human potential outcomes and simulation bias processes that generate the same data.

The remainder of this section examines tractable restrictions on $g(\cdot)$ and derives the corresponding identification conditions.

4.2 An interactive data-generating process

I first consider an interactive data-generating process where treatment-specific prompting bias interacts with individual-level prediction bias:

$$\hat{Y}_i(t) = Y_i(t) + \alpha_i \beta_t + \xi_i(t). \quad (8)$$

The interaction term $\alpha_i \beta_t$ implies that prompting bias affects individuals differently depending on how accurately the LLM represents them. This specification captures settings in which prompts do not shift predictions uniformly, but instead have differential effects across participants depending on how accurately the LLM represents them. This can occur, for example, if individuals whose backgrounds are poorly represented in the LLM’s training data are especially sensitive to how prompts are phrased.

I present the high-level takeaways here, while Appendix B contains formal derivations of these results. When the LLM simulates outcomes using equation 8 and with a mean-zero noise assumption, the simulated difference-in-means estimator gives:

$$E[\hat{\tau}_{\text{LLM}}] = \tau + E[\alpha_i(\beta_1 - \beta_0)]. \quad (9)$$

This motivates the following identification assumption:

Proposition 4 (Identification under an interactive data-generating process). *Suppose simulated outcomes are generated according to*

$$\hat{Y}_i(t) = Y_i(t) + \alpha_i \beta_t + \xi_i(t),$$

with $E[\xi_i(t)] = 0$. Then $\hat{\tau}^{LLM}$ is an unbiased estimator of τ if and only if

$$E[\alpha_i(\beta_1 - \beta_0)] = 0.$$

I refer to this condition as the **Interaction-Invariant Simulation Assumption (IISA)**. IISA requires that either (i) TISA (as in the linear case, where $\beta_1 = \beta_2$), or that (ii) individual-level prediction bias not consistently push the treatment effect in one direction or the other. The LLM can mispredict potential outcomes at the individual level, but these must cancel in expectation across the sample. Recall the identifying assumption TISA in the additive case in section 2. TISA required only that prompting bias enters as a common shift across individuals. IISA is weaker in that it allows prompting bias to differ across treatment arms, but stronger in that it restricts how such differences may covary with individual simulation bias.

As in the additive case, simulated outcomes alone do not allow for probing IISA. Instead, the researcher needs both simulated and human data. In particular, she must observe human outcomes for each participant under a treatment arm, plus the LLM’s simulated outcomes for those participants under both treatment and control prompts. This data allows her to identify which participants the LLM predicts accurately versus poorly, and to measure how simulation error varies across those participants. She can then measure baseline prediction error ($e_{i0} \equiv \hat{Y}_i(0) - Y_i(0)$) and compute simulated individual treatment effects ($\Delta \hat{Y}_i \equiv \hat{Y}_i(1) - \hat{Y}_i(0)$). If individuals with larger baseline prediction errors tend, on average, to exhibit larger or smaller simulated treatment effects, this indicates that prompting effects are disproportionately concentrated among participants the LLM captures poorly, thereby violating IISA. Under IISA, variation in baseline prediction error should be unrelated to how strongly the treatment prompt changes simulated outcomes.

Even if IISA fails, the researcher can turn to a bias-corrected estimator. Using a validation sample where she observes treated potential outcomes for some participants and control potential outcomes for others, the research can estimate $e_{i1} = \hat{Y}_i(1) - Y_i(1) = \alpha_i * \beta_1 + \epsilon_{i1}$ and $e_{i0} = \hat{Y}_i(0) - Y_i(0) = \alpha_i * \beta_0 + \epsilon_{i0}$. By mean-zero noise, we have $E[\epsilon_{i0} | T = 0] = E[\epsilon_{i1} | T = 1] = 0$. By random assignment, we have $E[\alpha_i | T = 1] = E[\alpha_i | T = 0] = E[\alpha_i]$. Thus subtracting prediction errors gives us:

$$E[e_{i1} | T = 1] - E[e_{i0} | T = 0] = E[\alpha_i(\beta_1 - \beta_0)].$$

She can thus subtract the difference in prediction errors across treated and control units ($e_{i1} - e_{i0}$) to return an unbiased estimate in the human sample.

4.3 A scaled data-generating process

I next consider a data-generating process where prompt bias scales with the level of the underlying potential outcome:

$$\hat{Y}_i(t) = Y_i(t) (1 + \beta_t) + \varepsilon_i(t). \quad (10)$$

Under this data-generating process, prompting bias enters as a proportional distortion rather than as an additive shift. This captures settings in which treatment prompts systematically exaggerate or attenuate outcomes for individuals with larger baseline responses. For example, if a treatment prompt induces more confident predictions, the LLM may disproportionately amplify outcomes for individuals whose true potential outcomes are already large, while leaving individuals with small outcomes largely unaffected.

Appendix C contains the full derivation of the results, and I highlight the main takeaways here. When the LLM simulates outcomes using equation 10 and with a mean-zero noise assumption, the simulated difference-in-means estimator gives:

$$E[\hat{\tau}_{\text{LLM}}] = \tau + E[\beta_1 Y_i(1) - \beta_0 Y_i(0)]. \quad (11)$$

Identification therefore requires that $E[\beta_1 Y_i(1) - \beta_0 Y_i(0)] = 0$, which I refer to as the **Outcome-Scale-Invariant Simulation Assumption (OSISA)**. OSISA requires that treatment-specific scaling of simulated outcomes exactly offset differences in average potential outcomes across treatment arms, so that proportional prediction error does not distort average treatment comparisons. OSISA holds if

$$\frac{E[Y_i(1)]}{E[Y_i(0)]} = \frac{\beta_0}{\beta_1}, \quad (12)$$

This means that the ratio of treated and control potential outcomes must be equal to the ratio of control and treated prompting bias. Although prediction error may increase with the level of aa

participant’s potential outcome, treatment assignment must not change how outcomes are scaled in expectation. Under OSISA, participants with larger potential outcomes are not systematically more or less mispredicted under treatment than under control in a way that shifts the average simulated treatment effect.

Like in the additive and interactive cases, solely LLM data is not enough to investigate this assumption: the left side of equation 34 only comes from actual human data. The researcher, however, can use a combination of LLM and human data to investigate this assumption. She can collect human data across treatment and control conditions to calculate the left-hand side of equation 34 ($E[Y_i(1)]$ and $E[Y_i(0)]$). Then she can use simulated and human outcomes to calculate the quantity on the right side of equation 34. Given equation 10, the researcher can estimate the treatment-specific scaling parameter β_t by regressing prediction error $e_{it} \equiv \hat{Y}_i(t) - Y_i(t)$ on the observed human outcome $Y_i(t)$ within each treatment arm. Under this model, the coefficient on the observed human sample identifies β_t . Thus, she can estimate all four quantities in equation 34 and therefore probe the assumption.

A bias-corrected estimator is available if the true data-generating process is scaled but OSISA fails. The researcher can use paired human and LLM data in the following way:

Step 1: Estimate average human potential outcomes. Using human outcome data from treatment and control groups, estimate the average potential outcomes $E[Y_i(1)]$ and $E[Y_i(0)]$.

Step 2: Estimate treatment-specific scaling parameters. Define prediction error as

$$e_{it} \equiv \hat{Y}_i(t) - Y_i(t).$$

Under the outcome-scaled data-generating process, we know that

$$e_{it} = \beta_t Y_i(t) + \varepsilon_i(t),$$

with $E[\varepsilon_i(t) \mid Y_i(t)] = 0$. Estimate β_t as the coefficient from a regression of e_{it} on $Y_i(t)$ within each treatment arm. This gives $\hat{\beta}_t$.

Step 3: Construct bias-corrected simulated outcomes. Define the bias-corrected simulated potential

outcomes as

$$\tilde{Y}_i(t) = \frac{\hat{Y}_i(t)}{1 + \hat{\beta}_t}.$$

Step 4: Estimate the corrected average treatment effect. Define the corrected simulated individual treatment effect as

$$\Delta\tilde{Y}_i = \tilde{Y}_i(1) - \tilde{Y}_i(0),$$

and estimate the bias-corrected ATE using

$$\hat{\tau}_{\text{corr}} = \frac{1}{N} \sum_{i=1}^N \Delta\tilde{Y}_i.$$

Table 1: Data-generating processes: Identification, testing, and bias correction

	Potential data-generating processes		
	Additive	Interactive	Outcome-scaled
Source of bias	Arm-level shift plus individual component that does not interact with treatment.	Individual simulation bias interacts with treatment-specific prompting asymmetry.	Treatment-specific scaling of outcomes by outcome magnitude.
Identifying assumption	TISA: $E[\beta_1 - \beta_0] = 0$. Prompting bias enters as a common shift and does not distort average treatment comparisons.	IISA: $E[\alpha_i(\beta_1 - \beta_0)] = 0$. TISA or the prediction biases cancel out across the sample.	OSISA: $E[\beta_1 Y_i(1) - \beta_0 Y_i(0)] = 0$. Outcome scaling does not distort average treatment comparisons.
Key falsification tests for the identifying assumption	The relative ordering of individuals by prediction error is invariant across treatment arms.	There is no systematic relationship between simulated treatment effects and baseline prediction error.	Prompting effects and average human outcomes satisfy $E[\beta_1 Y_i(1) - \beta_0 Y_i(0)] = 0$.
Bias correction	Possible via arm-level correction using human validation data.	Possible by estimating $e_{i1} - e_{i0}$ and rescaling simulated treatment effect.	Possible by estimating β_t and rescaling simulated outcomes.

5 Falsifying candidate data-generating processes

The results in the previous sections show that deriving identification conditions is possible when the research knows the data-generating process of the LLM. How, then, can researchers learn information about the data-generating process in practice? Two features of LLMs and this set-up make such an exercise possible.

The first feature is that human data allows the researcher to directly estimate prediction errors, which has testable restrictions depending on the data-generating process of the LLM. Given the research has access to human data, she can estimate the prediction error for a subset of participants: $e_i(t, p) \equiv \hat{Y}_i(t, p) - Y_i(t)$. Each family of data-generating processes imply different testable restrictions in how the prediction errors vary across individuals, prompts, and treatment arms. For example, the additive case (section 2) implies that while prediction errors may differ across participants, their relative ordering must be preserved across treatment and control conditions. The scaled version (section 4.3) implies that participants with larger potential outcomes will experience larger prediction errors. Thus human data allows the researcher to evaluate and falsify potential data-generating processes.

The second feature is that the researcher can combine across-participant variation in prediction error with variation across treatment arms or closely-related prompts to diagnose and rule out data-generating processes. Unlike human experiments, the researcher has direct control over these sources of variation (for example, she can use the same sample and run the same or closely-related prompts at scale). She can then use this data alongside the human data to falsify potential data-generating processes.

These two features – the ability to estimate prediction errors with human data and the ability to run repeated prompts at scale – present a way to falsify various classes of data-generating processes. Suppose the researcher observes, for a validation sample of individuals i , both human outcomes $Y_i(t)$ and simulated outcomes $\hat{Y}_i(t)$. She can then calculate $e_i(t) = \hat{Y}_i(t) - Y_i(t)$. Each data-generating process implies a specific function form for $e_i(t)$, which I display in Table 2.

Table 2: Prediction errors across potential data-generating processes

	Potential data-generating processes		
	Additive	Interactive	Outcome-scaled
LLM outcome equation	$\hat{Y}_{it} = Y_i(t) + a_i + b_t + \varepsilon_{it}$	$\hat{Y}_{it} = Y_i(t) + a_i \cdot b_t + \varepsilon_{it}$	$\hat{Y}_{it} = Y_i(t) (1 + b_t) + \varepsilon_{it}$
Prediction error	$e_{it} = a_i + b_t + \varepsilon_{it}$	$e_{it} = a_i \cdot b_t + \varepsilon_{it}$	$e_{it} = b_t \cdot Y_i(t) + \varepsilon_{it}$
Key testable restrictions	• Treatment or prompt changes induce only common shifts in prediction error. • Covariate-error relationships are stable across conditions.	• Treatment or prompt changes how individuals are mispredicted. • Covariate-treatment interactions appear in error regressions.	• Prediction error increases with outcome magnitude. • Conditional on outcome level, baseline covariates do not explain residual error.

6 Empirical application

In this section, I conduct an empirical application to illustrate the identification, diagnosis, and debiasing components of this framework.⁹ I make use of data from the Digital Twin Mega-Study (Peng et al., 2025). In the study, researchers proposed 19 social science studies where both humans and their digital twins completed the same protocol.

I use data from the hiring algorithm substudy. In this study, participants were shown and asked to evaluate four hypothetical job postings. The treatment of interest is whether hiring algorithms or humans conduct the initial stage of screening. For each participant, two of the four job postings were randomly assigned to include hiring algorithms (while the rest of the job posting remained the same). For each job posting, participants reported eight outcomes, including likelihood of applying, likelihood of being hired, and beliefs about fairness.

In Figure 1, I reproduce the average treatment effects from the both the human sample and the initial digital twin sample. The results indicate that for some outcomes, like likelihood of applying, digital twins are able to recover the true treatment effect in the human sample. However, for the majority of outcomes, the digital twin estimates are biased.

⁹As shown later in this section, I observe only one potential outcome per digital twin in Peng et al. (2025). This means I cannot test the efficiency results derived earlier in the paper.

For the remainder of this section, I walk through the procedure for the outcome regarding perceived clarity of community (*Clear Communication*). Later in the section, I repeat the analysis across all eight outcomes.

In the full human sample, hiring algorithms decrease perceived clarity of communication by $\hat{\tau}_H = -0.265$. Meanwhile, the corresponding estimate in the digital twin sample is $\hat{\tau}_{DT} = -0.025$. Therefore, the digital twin sample substantially attenuates the negative effect observed in the human sample ($\hat{\tau}_{DT} - \hat{\tau}_H = 0.240$).

I next split my human sample into either a validation or holdout sample. I randomly assign 25 percent of participants to the validation sample (261 participants totaling 1,044 observations) and the remaining 75 percent of the holdout sample ((738 participants totaling 2,952 observations). I use the validation sample is used to diagnose the simulation DGP, test its identifying restriction, and estimate treatment-dependent simulation bias.

I first examine whether the observed data are consistent with observable restrictions implied by the candidate DGPs developed above.

Under the additive DGP, $e_{ij}(t) = \alpha_i + \beta_t + \xi_{ij}(t)$. Treatment may shift the level of simulation error through β_t , but it should not systematically change which observable types of individuals the simulator predicts relatively well or poorly.

I assess this restriction using the following regression:

$$e_{ij} = \alpha_i + \beta T_{ij} + \lambda_j + T_{ij} X_i' \delta + u_{ij}, \quad (13)$$

where X_i contains gender, logic performance, and an indicator for being 18–29 years old. Because these characteristics do not vary within individual, their main effects are absorbed by the individual fixed effects. The restriction imposed by the additive specification is therefore

$$H_0 : \delta = 0. \quad (14)$$

I fail to reject this joint restriction:

$$F(3, 260) = 0.78, \quad p = 0.507. \quad (15)$$

None of the individual treatment interactions are statistically distinguishable from zero. Thus, along the observed characteristics considered here, we find no evidence that treatment changes how individual characteristics map into simulation error. These results indicate we cannot rule out the additive DGP.

I next assess the pure outcome-scaled DGP, where $\widehat{Y}_{ij}(t) = Y_{ij}(t)(1 + \beta_t) + \xi_{ij}(t)$.

A key implication of this specification is that the systematic mapping between human and simulated outcomes is proportional. I therefore estimate the specification

$$\widehat{Y}_{ij} = a + \delta T_{ij} + bY_{ij} + \phi(T_{ij}Y_{ij}) + u_{ij}. \quad (16)$$

The pure outcome-scaled DGP imposes

$$H_0 : a = 0 \quad \text{and} \quad \delta = 0. \quad (17)$$

The data strongly reject these restrictions. The estimated intercept is 4.383 ($p < 0.001$), while the standalone treatment coefficient is -0.062 ($p = 0.726$). The joint test gives

$$F(2, 260) = 993.43, \quad p < 0.001. \quad (18)$$

I also find no evidence that the slope linking human and digital-twin outcomes differs across treatment conditions ($\widehat{\phi} = 0.025$, $p = 0.478$).

Thus, among the DGPs assessed, the data allow us to reject the outcome-scaled specification. Moreover, they provide no evidence that observed-characteristic treatment interactions are required relative to the more parsimonious additive specification. I therefore use the additive DGP as the leading specification for the subsequent identification.

I now test the identifying assumption in this case. Under an additive DGP, unadjusted digital twin experiments require TISA to recover unbiased estimates in the human sample. Since I jointly observe both human and digital twin outcomes, I can test this condition using:

$$e_{ij} = \alpha_i + \Delta\beta T_{ij} + \lambda_j + u_{ij}, \quad \Delta\beta \equiv \beta_1 - \beta_0. \quad (19)$$

For the ‘‘Clear Communication’’ outcome, I estimate $\widehat{\Delta\beta} = 0.321$, with a corresponding standard error

of $SE = 0.082$ and a p-value of $p < 0.001$. Therefore, I am able to reject TISA, as the digital twin treatment effect is 0.32 points too positive in the validation sample.

I then turn to my adjusted estimator, $\hat{\tau}_{DT}^{\text{adj}} = \hat{\tau}_{DT} - \widehat{\Delta\beta}_V$, where $\widehat{\Delta\beta}_V$ is estimated from the validation sample. I then evaluate the adjustment on the holdout sample. Using only digital-twin responses for the holdout individuals gives $\hat{\tau}_{DT,H} = -0.045$ with a standard error of $SE = 0.022$. Applying the validation-sample bias estimate gives $\hat{\tau}_{DT,H}^{\text{adj}} = -0.0448 - 0.3208 = -0.3655$, with a standard error of approximately 0.085.

I then reveal the human outcomes in the holdout sample. The corresponding human treatment effect is $\hat{\tau}_{H,H} = -0.256$, with a corresponding standard error of $SE = 0.046$. The absolute difference between the raw digital-twin and human estimates is 0.211. After the adjustment, this gap shrinks to 0.109, or approximately 48 percent. The adjusted estimate overshoots the held-out human treatment effect, but it is substantially closer to the human benchmark than the unadjusted digital-twin estimate. I display the corresponding treatment effects in Figure 2, which shows the confidence intervals overlap with the human sample.

I next repeat the same procedure across all eight experimental outcomes. Table 3 reports the DGP diagnostics, estimated validation-sample bias, test of TISA, and human, unadjusted digital-twin, and adjusted digital-twin treatment effects.

The DGP diagnostics reveal a consistent pattern across outcomes. For all eight outcomes, I fail to reject the restriction that treatment does not interact with the observed individual characteristics used. The associated joint-test p -values range from 0.162 to 0.980. Thus, along these observed dimensions, I find no evidence that treatment systematically changes which types of individuals the digital twin predicts relatively well or poorly.

In contrast, the restrictions implied by the pure outcome-scaled DGP are strongly rejected for every outcome ($p < 0.001$ throughout). Taken together, these diagnostics consistently leave the additive specification as the most parsimonious of the candidate DGPs considered here that is not rejected by the observable restrictions.

Although each outcome is closest to the additive case, the identifying assumption does not hold uniformly across outcomes. I reject TISA for five of the eight outcomes and fail to reject it for three: for Apply ($p = 0.463$), Hired ($p = 0.614$), or Timely Decision ($p = 0.212$). My adjustment works

particularly well for the five outcomes where I rejected TISA, where the mean absolute digital-twin–human differences falls from 0.27 to 0.12, a reduction of approximately 55 percent. For the three outcomes for which TISA is not rejected, adjustment does not improve the realized point estimate.

This empirical exercise highlights three features of the framework. First, DGP diagnosis and causal identification are distinct. The observable diagnostics consistently leave the additive DGP as the leading parsimonious specification across the eight outcomes, yet TISA is rejected for five of them. A simulator can therefore align with the observable restrictions of a particular DGP while failing the additional assumptions for identification.

Second, experimental validity is outcome-specific. A single digital-twin system can closely reproduce some human ATEs while substantially misestimating others. This finding cautions against validating a simulator on one outcome or task and then treating it as generally valid for experimentation.

Third, when human validation data are available, treatment-dependent simulation bias can be estimated and used to improve digital-twin causal estimates. For all five outcomes in which TISA is rejected, the adjustment moves the out-of-sample digital-twin estimate toward the corresponding human effect. At the same time, the remaining discrepancies emphasize that adjustment introduces its own source of sampling uncertainty: the correction must itself be estimated from a finite human sample.

7 Conclusion

Causal inference using digital twins 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, and human data allow researchers to select across different data-generating processes and their identification assumptions.

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 to discriminate across potential data-generating processes). Only by grounding digital-twin applications in a formal causal framework

can we determine whether AI can substitute for human participants in academic research and business practice, or whether they will still need us in the loop.

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, Ziyang, 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.
- Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan**, “Large Language Models: An Applied Econometric Framework,” 2025.
- Luo, Haoyan and Lucia Specia**, “From Understanding to Utilization: A Survey on Explainability for Large Language Models,” 2024.
- McAdams, Tom A., Fruhling V. Rijdsdijk, 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, “Digital Twins as Funhouse Mirrors: Five Key Distortions,” 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.

Tables

Table 3: Diagnosing and Correcting Digital-Twin Treatment Effects

Outcome	Diagnosis and Validation				Holdout Evaluation		
	p : Scaled DGP	Validation Bias	TISA p	Human ATE	Raw DT ATE	Adjusted DT ATE	Error Reduction
Apply	< .001	0.029 (0.039)	0.463	-0.032 (0.017)	-0.042 (0.018)	-0.070 (0.043)	-0.029
Hired	< .001	0.019 (0.038)	0.614	-0.018 (0.017)	-0.038 (0.019)	-0.057 (0.043)	-0.019
Clear Communication	< .001	0.321 (0.082)	< .001	-0.256 (0.046)	-0.045 (0.022)	-0.366 (0.085)	0.102
Fair Policies	< .001	0.345 (0.089)	< .001	-0.281 (0.047)	-0.060 (0.020)	-0.405 (0.091)	0.097
Timely Decision	< .001	0.098 (0.079)	0.212	-0.067 (0.041)	-0.058 (0.021)	-0.157 (0.081)	-0.081
Clear Expectations/Culture	< .001	0.266 (0.085)	0.002	-0.245 (0.043)	-0.050 (0.025)	-0.316 (0.089)	0.124
Identify Top Applicant	< .001	0.420 (0.089)	< .001	-0.302 (0.046)	-0.055 (0.023)	-0.475 (0.092)	0.074
High Social Interaction	< .001	0.591 (0.118)	< .001	-0.578 (0.061)	-0.110 (0.028)	-0.701 (0.121)	0.345

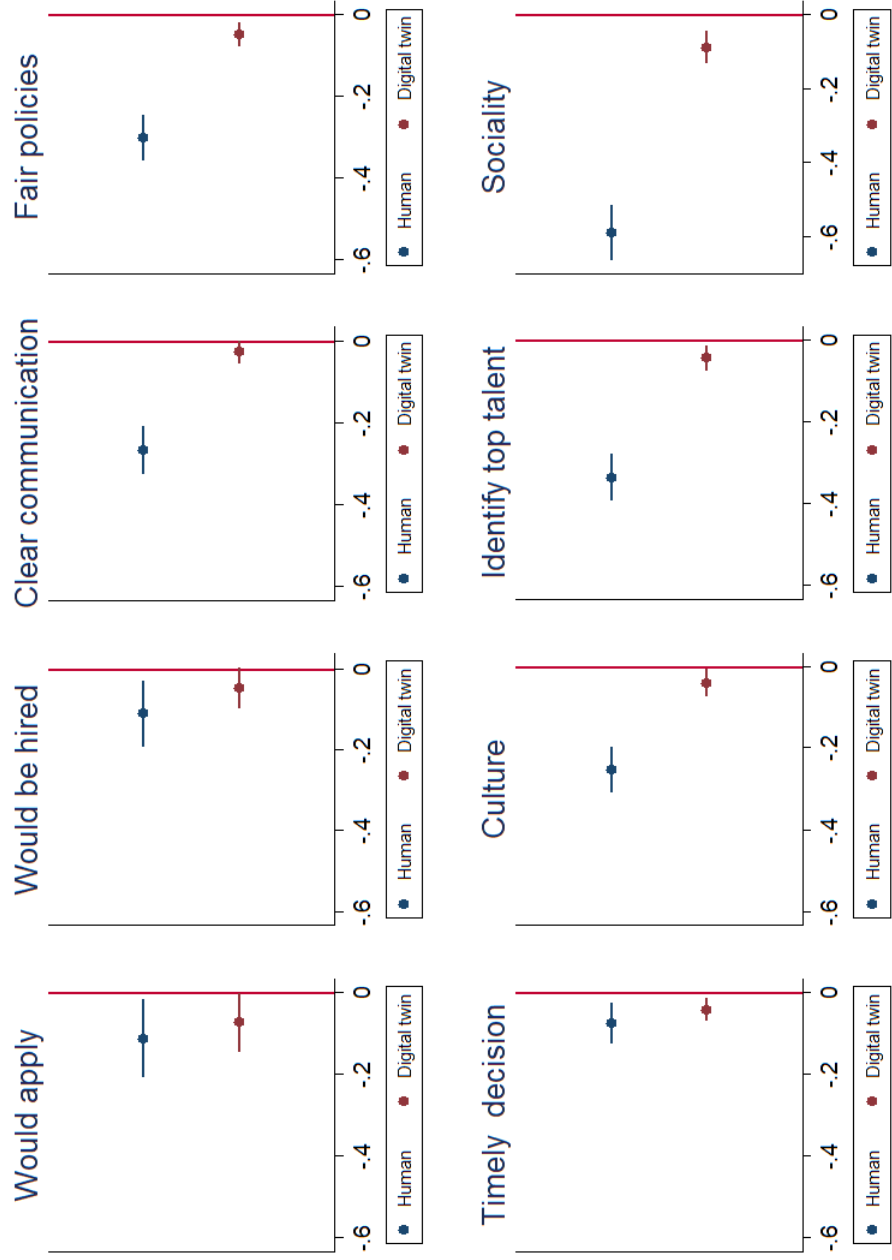
Notes: The table reports results from the 25% human validation sample and the 75% holdout sample. “Scaled DGP” reports the p -value from the joint test of the restrictions implied by the pure outcome-scaled DGP. Validation Bias is the estimated treatment-induced change in prediction error in the validation sample, $\hat{B}_V = E[e(1)] - E[e(0)]$. Under the additive DGP, TISA requires $B_V = 0$; the TISA column reports the p -value for this restriction. Human ATE and Raw DT ATE are estimated using only the holdout sample. The adjusted digital-twin estimate is $\hat{\tau}_{DT}^{corr} = \hat{\tau}_{DT} - \hat{B}_V$. Standard errors are reported in parentheses and are clustered by individual. The standard error of the adjusted DT estimate incorporates uncertainty in both the holdout DT estimate and the validation-sample bias estimate. Error Reduction is

$$|\hat{\tau}_{DT} - \hat{\tau}_H| - |\hat{\tau}_{DT}^{corr} - \hat{\tau}_H|,$$

so positive values indicate that adjustment moves the digital-twin estimate closer to the held-out human treatment effect.

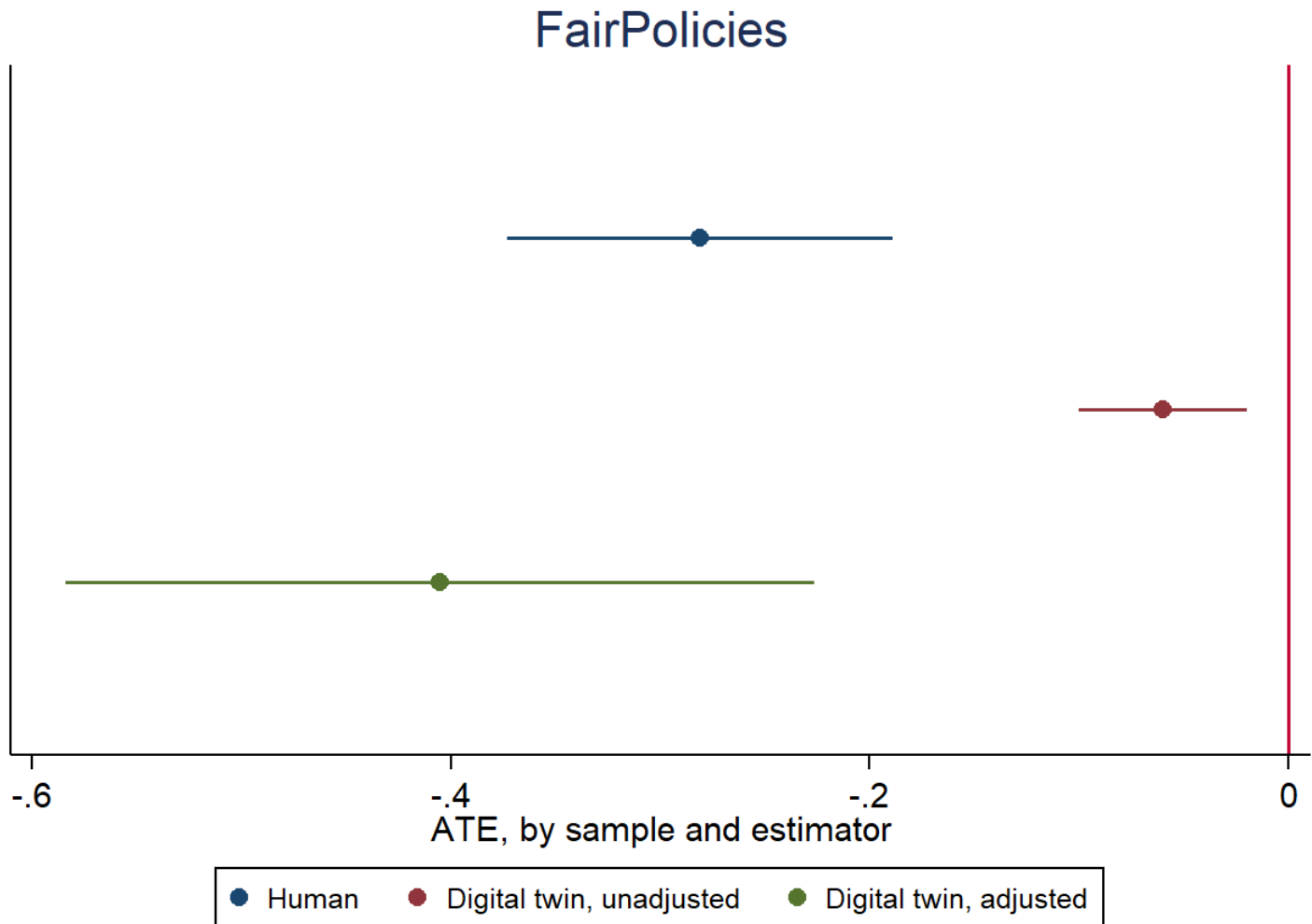
Figures

Figure 1: Comparison of ATEs, human sample versus digital twin sample



Notes: This figure compares the average treatment effects in the human sample and the digital twin sample.

Figure 2: Comparison of ATEs for DV = Clear Communication



Notes: This figure compares the average treatment effect on beliefs in clear communication.

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:

$$Y_i(t) = \theta_i + t * \tau_i + \epsilon_i(t) \quad (20)$$

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

The research then uses the simulated CATE estimator:

$$\hat{\tau}^{LLM}(x) = E_i[\hat{Y}_i(1) - \hat{Y}_i(0) \mid X_i = x]. \quad (21)$$

Substituting the simulation model from equation 2 and simplifying, we get:

$$\hat{\tau}^{LLM}(x) = \tau(x) + (\beta_1 - \beta_0) + E[\xi_i(1) - \xi_i(0) \mid X_i = x] \quad (22)$$

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.

B Inference under an interactive data-generating process

B.1 An interactive model of LLM-generated outcomes

This subsection considers a data-generating process in which individual-specific prediction bias interacts with treatment-specific prompting bias in how the LLM generates potential outcomes. Unlike the additive case, treatment prompts do not shift simulated outcomes uniformly. Instead, the magnitude of prompting bias depends on how accurately the LLM represents a given individual.

Let $Y_i(t)$ denote the human potential outcome for individual i under treatment $t \in \{0, 1\}$. As in the main text, the estimand of interest is the average treatment effect

$$\tau = E[Y_i(1) - Y_i(0)].$$

I assume constant treatment effects in the human population but relax this assumption later.

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

$$\hat{Y}_i(t) = Y_i(t) + \alpha_i \beta_t + \varepsilon_i(t), \quad (23)$$

where α_i captures individual-specific simulation bias, β_t captures treatment-specific prompting bias,

and $\varepsilon_i(t)$ is idiosyncratic simulation noise with mean zero conditional on i and t .

Compared to the additive benchmark, prompting bias here enters multiplicatively with individual-level simulation bias. As a result, treatment-specific distortions in simulated outcomes depend on which individuals the LLM represents poorly, rather than operating as a common shift across the population.

B.2 The simulated ATE and its decomposition

Consider the simulated difference-in-means estimator

$$\hat{\tau}_{\text{LLM}} = E[\hat{Y}_i(1) - \hat{Y}_i(0)].$$

Substituting equation (23) yields

$$\hat{Y}_i(1) - \hat{Y}_i(0) = (Y_i(1) - Y_i(0)) + \alpha_i(\beta_1 - \beta_0) + (\varepsilon_i(1) - \varepsilon_i(0)). \quad (24)$$

Taking expectations and using the mean-zero noise assumption,

$$E[\hat{\tau}_{\text{LLM}}] = \tau + E[\alpha_i(\beta_1 - \beta_0)]. \quad (25)$$

The second term captures bias arising from the interaction between individual-level simulation bias and treatment-specific prompting differences.

B.3 The Interaction-Invariant Simulation Assumption

Identification of the true human ATE requires that the interaction bias term in equation (25) equals zero:

$$E[\alpha_i(\beta_1 - \beta_0)] = 0. \quad (26)$$

I refer to condition (26) as the *Interaction-Invariant Simulation Assumption (IISA)*. IISA requires that either (i) TISA (as in the linear case, where $\beta_1 = \beta_2$), or that (ii) individual-level prediction bias

not consistently push the treatment effect in one direction or the other. The LLM can mispredict potential outcomes at the individual level, but these must cancel in expectation across the sample.

For comparison, TISA in the additive case requires only that prompting bias enter as a common shift. IISA is weaker in that it allows prompting bias to differ across treatment arms, but stronger in that it restricts how such differences may covary with individual simulation bias.

B.4 Using human data to test the interaction structure

Human outcome data are required for assessing the plausibility of IISA. These outcomes allow the researcher to observe participant-level prediction errors and thus construct empirical proxies for individual simulation bias.

Suppose human outcomes are observed for a validation sample, with each individual observed under their realized treatment assignment. The researcher can compute overall prediction error as

$$e_{it} \equiv \hat{Y}_i(t) - Y_i(t), \quad (27)$$

which captures the discrepancy between the LLM-generated outcome and the corresponding human outcome under treatment condition t .

I focus on prediction error in the control condition:

$$e_{i0} = \hat{Y}_i(0) - Y_i(0), \quad (28)$$

In the interactive data-generating process, individual simulation bias is represented by the unobserved parameter α_i , which influences how accurately the LLM represents individual i across prompts. Under this model, $e_{i0} = \alpha_i\beta_0 + \varepsilon_i(0)$ so baseline prediction error identifies across-participant variation in α_i only up to the common scaling factor β_0 and noise. As a result, e_{i0} is informative about the relative magnitude of individual simulation bias across participants but does not recover α_i itself. Combining this with equation (24), IISA implies that simulated treatment effects $\hat{Y}_i(1) - \hat{Y}_i(0)$ should not systematically covary with baseline prediction error.

Empirically, this can be tested by estimating:

$$\hat{Y}_i(1) - \hat{Y}_i(0) = \gamma + \delta e_{i0} + u_i, \quad (29)$$

where e_{i0} serves as a scaled proxy for individual simulation bias. Under IISA, the coefficient δ should be equal to zero up: a statistically significant relationship between simulated treatment effects and baseline prediction error indicates that individuals with larger individual simulation bias experience systematically different prompting effects, providing evidence against IISA.

B.5 Bias correction under an interactive data-generating process

Even if IISA fails, the researcher can turn to a bias-corrected estimator. Using a validation sample where she observes treated potential outcomes for some participants and control potential outcomes for others, the research can estimate $e_{i1} = \hat{Y}_i(1) - Y_i(1) = \alpha_i * \beta_1 + \epsilon_{i1}$ and $e_{i0} = \hat{Y}_i(0) - Y_i(0) = \alpha_i * \beta_0 + \epsilon_i(0)$. By mean-zero noise, we have $E[\epsilon_{i0} | T = 0] = E[\epsilon_{i1} | T = 1] = 0$. By random assignment, we have $E[\alpha_i | T = 1] = E[\alpha_i | T = 0] = E[\alpha_i]$. Thus subtracting prediction errors gives us:

$$E[e_{i1} | T = 1] - E[e_{i0} | T = 0] = E[\alpha_i(\beta_1 - \beta_0)].$$

She can thus subtract the difference in prediction errors across treated and control units ($e_{i1} - e_{i0}$) to return an unbiased estimate in the human sample.

C Inference under an outcome-scaled data-generating process

C.1 An outcome-scaled model of LLM-generated outcomes

This subsection considers a data-generating process in which treatment-specific prompting bias scales the level of the underlying human potential outcome in the LLM’s simulation. In contrast to the additive and interactive cases, simulation bias here is proportional to the magnitude of the outcome itself.

Let $Y_i(t)$ denote the human potential outcome for individual i under treatment $t \in \{0, 1\}$. As in the

main text, the estimand of interest is the average treatment effect

$$\tau = E[Y_i(1) - Y_i(0)].$$

I again begin with constant treatment effects in the human population.

The LLM generates simulated potential outcomes according to the outcome-scaled data-generating process

$$\hat{Y}_i(t) = Y_i(t) (1 + \beta_t) + \varepsilon_i(t), \quad (30)$$

where β_t captures treatment-specific scaling bias and $\varepsilon_i(t)$ is idiosyncratic simulation noise with mean zero conditional on i and t .

Under this model, treatment prompts do not interact with individual-specific representation error. Instead, treatment-specific bias operates by rescaling outcomes as a function of the level of the potential outcome, so across-individual variation in prediction error arises solely from heterogeneity in the underlying potential outcomes.

C.2 The simulated ATE and its decomposition

Consider the simulated difference-in-means estimator

$$\hat{\tau}_{\text{LLM}} = E[\hat{Y}_i(1) - \hat{Y}_i(0)].$$

Substituting equation (30) yields

$$\hat{Y}_i(1) - \hat{Y}_i(0) = (Y_i(1) - Y_i(0)) + \beta_1 Y_i(1) - \beta_0 Y_i(0) + (\varepsilon_i(1) - \varepsilon_i(0)). \quad (31)$$

Taking expectations and by the mean-zero noise assumption, we get:

$$E[\hat{\tau}_{\text{LLM}}] = \tau + E[\beta_1 Y_i(1) - \beta_0 Y_i(0)]. \quad (32)$$

Unlike the additive case, the bias term depends on the joint distribution of potential outcomes

and treatment-specific scaling parameters. Even under constant treatment effects, this term does not generally vanish without additional restrictions.

C.3 The Outcome-Scale-Invariant Simulation Assumption

Identification of the true ATE using the unadjusted simulated difference-in-means estimator requires that the scaling bias term in equation (32) vanish:

$$E[\beta_1 Y_i(1) - \beta_0 Y_i(0)] = 0. \quad (33)$$

I refer to condition (33) as the *Outcome-Scale-Invariant Simulation Assumption (OSISA)*. OSISA can be re-written as

$$\frac{E[Y_i(1)]}{E[Y_i(0)]} = \frac{\beta_0}{\beta_1}, \quad (34)$$

This means that the ratio of treated and control potential outcomes must be equal to the ratio of control and treated prompting bias. Although prediction error may increase with the level of a participant's potential outcome, treatment assignment must not change how outcomes are scaled in expectation. Under OSISA, participants with larger potential outcomes are not systematically more or less mispredicted under treatment than under control in a way that shifts the average simulated treatment effect.

C.4 Using human data to test OSISA

Human data is required for assessing the plausibility of OSISA.

Suppose human outcomes are observed for a validation sample, with each individual observed under their realized treatment assignment. The researcher can compute prediction error as

$$e_{it} \equiv \hat{Y}_i(t) - Y_i(t). \quad (35)$$

Under the outcome-scaled data-generating process in equation (30), prediction error satisfies:

$$e_{it} = \beta_t Y_i(t) + \varepsilon_i(t), \quad (36)$$

where β_t captures treatment-specific outcome scaling and $\varepsilon_i(t)$ is mean-zero noise. Paired human and simulated outcomes therefore allow the researcher to estimate the scaling parameters β_0 and β_1 from regressions of prediction error on observed outcomes within each treatment arm.

In addition, human outcome data identify the average potential outcomes $E[Y_i(0)]$ and $E[Y_i(1)]$. OSISA imposes the restriction

$$E[\beta_1 Y_i(1) - \beta_0 Y_i(0)] = 0,$$

which is a function only of these four identified quantities. As a result, OSISA is directly testable using a combination of human outcome data and paired human–LLM validation data. Empirically, the researcher can assess this restriction by estimating β_0 and β_1 , computing the implied scaling-adjusted means $\beta_0 E[Y_i(0)]$ and $\beta_1 E[Y_i(1)]$, and testing whether this difference is statistically distinguishable from zero.

C.5 Bias correction under an outcome-scaled data-generating process

Bias correction is possible under this data-generating process. The researcher can estimate treatment-specific scaling parameters β_t through human validation data like above (using the coefficient on the outcome on a regression of prediction error on outcome within a given treatment arm).

The bias-corrected simulated outcome is then

$$\tilde{Y}_i(t) = \frac{\hat{Y}_i(t)}{1 + \hat{\beta}_t}.$$

The corrected simulated ATE is obtained by computing the difference-in-means of $\tilde{Y}_i(t)$ across treatment conditions.