The Limits of Behavioral Simulation With AI*

Thomas Graeber Christopher Roth

November 2, 2026

Abstract

How well can AI simulate human behavior? We review over 100 studies to ask what the evidence means for credible use cases. First, how accurate are simulations where they have been checked against human data? Accuracy falls with specificity: simulations often get the direction and ranking of well-documented effects right but are far less reliable about effect sizes when simulating respondents, and about response distributions, subgroup differences, and individual choices. Second, does accuracy measured in one setting transfer to new settings, where simulation is most useful but cannot be checked? We show that it transfers to new settings of a familiar kind, but evidence for genuinely new contexts, such as a new rule or a later period, is scarce and, where it exists, negative. The evidence is most consistent with simulations drawing on a reduced-form account of what has been written about people rather than a structural model of behavior. Better models may improve accuracy where it can be checked, but that does not establish accuracy in the new settings one cares about. We discuss the promise of three possible approaches. Adding behavioral theory carries predictions beyond the training data, but only as far as the theory is accurate, complete and applicable. Additional training on behavior from settings just like the one of interest improves simulation only if they are highly similar in kind. Human data close to the context of interest can instead be used to directly measure and correct simulation error in application, which carries as far as the error structure stays similar.

*Leon Kriesmair, Veljko Petrovic, and Timo Stenz provided outstanding research assistance. Graeber: University of Zurich, Department of Economics, thomas.graeber@econ.uzh.ch; Roth: LMU Munich and CEPR, christopher.roth@econ.lmu.de. This research program and related results were made possible by the support of the NOMIS Foundation.

1 Introduction

Many decisions by firms, governments, and researchers turn on a single question: how will people behave? For example, a firm wants to know which of two advertising messages will generate more revenue, a government agency wants to know how many households will take up a benefit program, and a researcher wants to know which experiment to run next to identify a policy that maximizes take-up. For each of these examples, the ideal data would come from real humans in exactly those situations. However, these data do not exist yet, and collecting them, if at all possible, would be slow and expensive. Generative artificial intelligence offers a shortcut by allowing for flexible behavioral simulation at very low cost. Indeed, a growing number of companies offer such synthetic respondents as a service (Gui and Toubia, 2025).[1] The promise goes beyond costs alone: large language models allow researchers to simulate behavior across a large set of counterfactuals that would be too expensive ever to test with a human subject pool. Moreover, a language model trained on a large corpus of text and data might be able to restore and integrate the context that most experiments strip away, for any person, decision, and situation that can be described.

However, this promise rests on an assumption that is rarely made explicit. The reason to simulate is to learn about something new, such as a message not yet tested or a policy not yet introduced, for which no human data exist yet. By contrast, we can only judge how accurate a simulator is in situations where we already have human data. Its value, in turn, is contingent on the assumption that accuracy in familiar situations transfers to unfamiliar ones.

Suppose the product, the rules, or the period changes. The key question is then whether the relationship the model has learned between circumstances and behavior still holds after the change. We refer to this requirement as *invariance*. A model that satisfies it under a given change is *structural* with respect to that change. A *reduced-form* account, by contrast, summarizes regularities in its source data, without any guarantee that they hold under the change. Invariance is therefore relative to the change: the same model can be structural for one change and reduced-form for another (Haavelmo, 1944; Hurwicz, 1962). Whether a simulation is invariant depends on what it has learned. Behavior depends on who is choosing, what is at stake, and all the specifics of the situation. A model that has picked up a regularity without learning how it depends on these determinants will reproduce it only where they are combined as in its source data.[2] Accuracy where it can be checked does not reveal which case one is in: work on world models shows that accurate in-domain prediction does not imply that a model has

[1] See, e.g., Bort (2026), Gillin (2026), and Roytburg (2026).

[2] The sharpest case is a change in the rules themselves, because the regularity in the record was produced by people's response to the old rules, and no record of behavior under the new ones exists; this is the Lucas critique (Lucas, 1976).

learned the structure required for out-of-domain prediction (Vafa et al., 2024, 2025).

This review asks whether it is justified to use simulations to predict behavior in novel settings, where no human data exist yet. As a first step, we ask how accurate simulations are where human benchmarks are available. In a second step, we examine whether the accuracy established in one setting carries over to another setting. This second step involves checking the same simulation against human data in at least two settings and treating one of them as if it were new. How informative one check is for the other depends on how these settings differ from one another. When a new setting changes something all the tested settings had in common, like who is answering, what is at stake, the rules, or the time period, we say it differs in kind.

To address these questions with the existing evidence, we characterize each study by three objects of interest. First, the *estimand* is what the simulation is meant to predict, such as a take-up rate, the effect of a message, or one customer's choice. We characterize it along four dimensions: its *granularity*, its *outcome type*, its *population*, and its *context*. Second, the *estimator* is the full procedure that produces the estimate. For example, one can ask the LLM directly for a forecast or alternatively ask it to simulate individual responses and then aggregate them. Third, the *evaluation* compares the output with human data. These data often describe a different estimand from the one of interest, such as another population, outcome type, or context, and what an evaluation shows about the target depends on how the two differ.

Our first main finding is that accuracy declines as the estimand becomes more granular. This means that models have a high accuracy in recovering the direction and ranking of large, familiar average treatment effects. When it comes down to predicting magnitudes, distributions or differences across subgroups they become less reliable. The results in Ashokkumar et al. (2026) illustrate these performance differences. They show that predicted and observed effects correlate at 0.85, yet the predictions are 70 to 80 percent too large. Using LLMs as experts to make forecasts of treatment effects in economics from pre-registration documents instead of as simulated respondents, Vivalt (2026) shows that predictions end up within about 0.05 SD of the realized effect at the median and beat simple benchmarks, indicating that using the LLM as an expert may be less prone to magnitude inflation. Synthetic samples also understate the variance of responses and distort differences between groups (Bisbee et al., 2024; Grief-Albert et al., 2026). The pattern extends to the most granular level: predicting how a given individual will respond. Evidence shows that even when rich data on individuals is supplied, individual-level predictions fall well short of people's own test–retest consistency (Park et al., 2026; Toubia et al., 2025). Moreover, accuracy in predicting survey answers does not transfer to incentivized choices made by the same people (Park et al., 2026).

What do these findings on accuracy imply for settings where no human benchmark is available? This depends critically on how different the new setting is from familiar ones. Within a

familiar kind of setting, accuracy appears to carry over to some extent: in Ashokkumar et al. (2026), experiments unpublished at the training cutoff are predicted about as well as published experiments, with a correlation of 0.90. Although this rules out simple recall, the unpublished experiments come from the same pool of U.S. survey experiments and may closely resemble studies described in its training data. Moving across populations, the evidence is more conditional, with accuracy tracking coverage in the written record. In 74 countries, simulations reproduce differences by gender, age, and cognitive ability far more often than differences between countries, and tend to make larger errors in poorer and less digitally connected countries (Chen et al., 2026). Moving to the question of how accuracy extends across contexts, the body of evidence is scarce. Changed rules have not been tested, and the sole test across time is negative: in a field experiment, a correction learned from human data carries over between regions but not from the first intervention week to later weeks (Zhang et al., 2026). Thus, transfer across kinds remains largely untested, and the settings where simulation is most in demand are those least often checked.

Our preferred interpretation of the two findings is that the simulations draw on a reduced-form account of the record rather than a structural model of behavior. The two findings speak to this in different ways. First, whether a simulation is structural with respect to a change can only be tested by that specific change. The evidence from tests across changes in population or context is thin and, where it exists, negative. Second, the granularity finding does not speak to invariance directly.[3] But the pattern of errors it reveals, that it tends to get headline effects right while their size, spread, and group differences are off, is what one would expect from a model that has absorbed what has been written about behavior rather than the behavior itself, which is the argument we make precise in Section 3. Note that our interpretation should thus be taken as based on what the currently accumulated evidence supports, rather than a general claim that AI-based simulations have no structure.

We next ask whether the reduced-form pattern is what one should expect, given how a simulation is produced. The record a model learns from describes behavior more than it links people and circumstances to choices. Training then rewards matching that record and producing approved answers, not predicting what people do (Binz et al., 2026). A simulation therefore may resemble a well-read expert. It reproduces the literature's account of how people respond and tends to miss departures from it, which are often why an experiment is run in the first place (Vivalt, 2026). Human experts have the same problem. Their forecasts of experimental results are informative about averages but miss the surprises (DellaVigna and Pope, 2018; DellaVigna et al., 2019; Milkman et al., 2021), and they face the same transport problem once the setting

[3] A model can be invariant to a change but coarse. A representative-agent model, for example, may predict the mean response to a tax change correctly while saying nothing about how responses differ across people.

changes. What is different about simulation is not this limit but what follows from it. Near-zero marginal cost invites use exactly where validation is absent. Cheap repetition creates an illusion of precision. And because the model has read the research record and does not reveal what it has seen, it is hard to know what a given prediction is based on. We conjecture that better models alone will not change this pattern. Progress in coding and mathematics relies on reinforcement learning against answers a computer can check almost for free (Wei, 2025). Checking a behavioral prediction requires costly data on what people actually do.

These limits suggest three ways to use AI-based simulations in situations where they have not been checked. The *theory route* adds explicit behavioral structure where it is missing, meaning that predictions rest on some account of how people respond (e.g., Manning and Horton, 2025). Predictions formed this way can reach situations that differ from the training data but this approach has obvious boundaries. Behavioral theory is always just a partial account of the entire set of factors that determine behavior in a real-world context, and the more the unexplained factors matter the smaller the improvement will be. In particular, most behavior is notoriously context-dependent, meaning that it is strongly determined by features that are not about the person but about the context. However, behavioral science still lacks domain-general accounts of context effects. The two other approaches are reduced form in the sense that they do not recover structure but instead rely on proximity to human data, either before use or at use. The *training route* further trains or fine-tunes a model on observed behavior from settings that are close to the setting of interest (Binz et al., 2025). This works to the extent that the setting of interest is covered by the additional training data, i.e., it is from the same family of or "of the same kind", which a genuinely new application precludes. Finally, the *correction route* measures the simulation error against a human sample that is as close as possible to the setting of interest and then transfers the correction to the latter. This route does not need the simulation to be accurate at the setting of interest, only to get the change between the checked setting and the setting of interest right. That is still an invariance assumption. What differs is where it sits and when it applies. It is stated by the user, at the point of use, and can be stress-tested there. For practitioners, this means that one has to match the setting with human data to the specific claim and decision environment as closely as possible so that the distortion from the approximation stays minimal. The central open question is how far evidence about accuracy in one setting transfers to another.

Our starting point is the literature that uses LLMs to predict human behavior (Horton, 2023; Aher et al., 2023; Argyle et al., 2023; Manning et al., 2024; Park et al., 2023). We also draw on econometric, machine learning, and statistical work that treats model output as an imperfect measurement and uses human validation data to estimate or correct its error (Ludwig et al., 2026; Dell and Rambachan, 2026; Angelopoulos et al., 2023a). Several studies have also chal-

lenged specific methods and uses of behavioral simulation (Gui and Toubia, 2025; Bisbee et al., 2024; Hullman et al., 2026; Schröder et al., 2025; Chen et al., 2026). Relative to this work, we make two contributions. We distinguish accuracy in validated settings from evidence on whether this accuracy extends to new unfamiliar ones. We then systematically discuss three alternative ways of improving model accuracy in new unfamiliar settings.

We review 108 studies that compare model output with observed human reports, choices, or treatment effects, most of them targeting survey answers. They were identified through a documented search that screened 675 records up to August 31, 2026 (Appendix D). Several more studies have appeared after the search date and were added by hand, e.g. Vivalt (2026) and Chen et al. (2026). Because the large collection of studies uses different outcomes and measures, we do not pool their results.

Our paper proceeds as follows. Section 2 presents the framework and the evidence. Section 3 explains the pattern and why it is likely to persist. Section 4 derives credible uses and states the open question. Section 5 concludes.

2 Where Behavioral Prediction Works and Where It Fails

In this section, we first describe our taxonomy for evaluating AI-based simulations. We then use it to review the evidence on three questions: how accurately simulations reproduce human data in existing evaluations, whether this accuracy generalizes to new populations and contexts, and what inferences synthetic estimates and their standard errors support.

2.1 What Is Claimed, How It Is Produced, and How It Is Validated

2.1.1 The Estimand

The estimand is the target quantity one seeks to estimate, such as the share of eligible households that will take up a benefit, the effect of a message on donations, the choice one customer will make. We characterize the estimand along four dimensions, described in Table 1 below.

Granularity. Granularity is the level of detail a claim contains about a given population. For example, the sign of an average effect size in a given population is a coarse claim. More granular claims would be the size of an effect or the distribution among different subgroups.

Outcome type. Outcome type is the specific behavioral record that's being predicted, which ranges from reported beliefs or attitudes, to stated intentions, hypothetical choices, or actual behavior. A vast literature studies the nuanced relationship between different types of outcomes.

Table 1: Four dimensions of an estimand

Dimension	Question it answers	Examples
Granularity	How much detail?	An effect's sign or size; a response mean or distribution; population, subgroup, or individual predictions
Outcome type	What response is measured?	A survey answer; a stated intention; a hypothetical choice; an incentivized choice; a purchase
Population	Whose behavior?	German households; a retailer's customers; an experiment's participants
Context	Under what conditions?	Program rules; experimental incentives; prices; time period

Notes. Granularity admits a partial ordering from coarse to fine. The other dimensions distinguish settings of interest without ranking them. A comparison between two settings may change more than one dimension at once.

For instance, stated intentions correlate with actual purchasing behavior, but the two are far from interchangeable (e.g., LaPiere, 1934).

Population. The population here means the specifics of the group of people whose behavior is being predicted. This could be, for example, German households, the customers of one retailer, or the participants of one online panel.

Context. Context refers to the entire set of circumstances within which people act. The full context is, by definition, infinite-dimensional, because it includes each and every feature of the decision setting. This comprises not only the clearly relevant ones, such as the rules, incentives, available options, information, and time period, but also the minor details of the environment, such as what the decision-maker sees and hears. Because of this dimensionality, there is no decision setting in which literally all features of the context can be described fully in an estimand. Of course, some features are far more consequential for behavior than others, so the question is often which parts of the context are the relevant ones.

2.1.2 The Estimator

The estimator refers to the full procedure used in generating an estimate. It starts with the model, by which we mean a specific type of algorithm with a set of trained weights. The model receives a prompt that describes the simulation task and, sometimes, the people to be simulated. Additional decoding settings such as the temperature then determine how answers are drawn. Finally, the model's outputs are turned into the estimate, for example by extracting an

answer from each output and combining the answers across simulated respondents. Any of these choices can change how accurate the estimate is for a given estimand.

We distinguish two broad classes of prompts. A *direct* estimator asks the LLM for the target quantity itself: "What share of customers will choose A?" A *respondent* estimator, on the other hand, asks the LLM to simulate individual responses and then aggregates these simulated responses into an estimate.

Simulation noise is variation across repeated runs of a fixed procedure with the same supplied information. A given procedure run many times thus yields a distribution of estimates. We call μ^M the mean of that distribution, which is a fixed property of the model, prompt, and settings, and $\bar{\theta}^M$ the average across a finite number of actual runs. A larger number of runs brings $\bar{\theta}^M$ closer to μ^M but does nothing per se to close the gap between μ^M and the human quantity of interest, because every run comes from the same procedure. This means that repeated simulation pins down what the procedure says on average without establishing whether that is right.

2.1.3 The Evaluation: What Was Checked, and Where

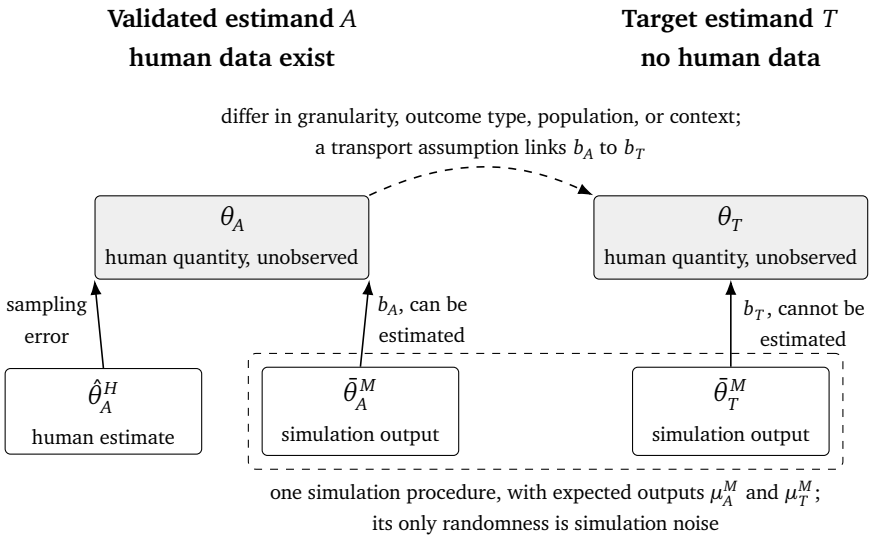
An evaluation compares the estimator's output with an estimate from human data. We call the quantity we ultimately want to predict the *target estimand*, θ_T. Because the very motivation for running AI-based simulations is that human data is not available for the situation of interest, θ_T, we often have to evaluate the accuracy of the procedure in a different situation, described by the *validated estimand*, θ_A. The human data provide an estimate in that validation setting, $\hat{\theta}_A^H$, of θ_A, subject to sampling error. Comparing the model's output with this estimate allows us to assess its error at θ_A; whether this also establishes accuracy at θ_T depends on how the two estimands differ.

Suppose an agency wants to use an AI-based simulation to estimate the share of eligible households nationwide that will take up a new benefit after launch. It does not, by definition, have access to the actual outcomes of that intervention before it happens. The only available human data may allow it to check the outputs of the simulator against a different quantity. Early data from a regional pilot, for example, make it possible to assess accuracy of its predictions of take-up in that very region without establishing accuracy nationwide. Alternatively, a survey of households would provide input on the predictions of stated intentions, but it does not in itself establish accuracy in predicting actual take-up. Now suppose instead that the agency observes take-up after launch in a random sample of eligible households drawn from across the country. Although the sample does not include all households, it allows the agency to estimate the nationwide take-up rate. In this last case, the validated and target estimands coincide. Here, the agency can use these observations to estimate the model's error at the target. In fact, com-

bining this sample with model predictions for additional households improves precision after correcting for the model's error.

These examples illustrate three cases. When the validated and target estimands coincide, the model's error at the target can be estimated directly, subject to sampling uncertainty. When they differ along at least one dimension of Table 1, error can be measured only at the validated estimand; Figure 1 illustrates this case. When no human data are available, the output remains an unchecked forecast.

Figure 1: The evaluation setup



Notes. Each estimand is a quantity about behavior in context. The two differ along at least one dimension of Table 1 and both are unobserved. Human data exist for A but not for T. The simulator's systematic error at the validated estimand, b_A, can be estimated by comparing the simulation output with the human estimate. Its systematic error at the target, b_T, cannot. Only an assumption carries what was learned about b_A over to b_T. Simulation noise around the simulation outputs can be reduced by more runs and is uninformative about either error.

Write $b_A = \mu_A^M - \theta_A$ for the model's systematic error at the validated estimand and $b_T = \mu_T^M - \theta_T$ for its error at the target. Human data at A allow us to estimate b_A by comparing the simulator's output with the human estimate, $\hat{b}_A = \bar{\theta}_A^M - \hat{\theta}_A^H$. Without human data at the target, we cannot estimate b_T directly. What the simulation reveals about the target therefore depends on an assumption that links the two errors. We call it a *transport assumption*, following work on the transportability of causal effects across populations (Pearl and Bareinboim, 2014). Economists will recognize this as the external validity problem for experiments. Site selection, heterogeneous effects across contexts and the limits of extrapolating from trials are all instances of it (Allcott, 2015; Vivalt, 2020; Meager, 2019; Deaton and Cartwright, 2018). The difference lies in what gets transported. In this case it is not a treatment effect but the simulation's error. If the two errors are assumed to be equal, the model's output at the target, corrected by $\hat{b}_A$, identifies the target. If they are assumed to differ by at most δ, the target is bounded by the

simulation. Without such an assumption, the simulator's output does not identify the target (see Chen et al., 2026). "Error" here generally refers to a suitable measure of discrepancy in the setting of interest. For quantities other than a mean or a treatment effect, such as rankings, distributions, or individual choices, it can be a distance between distributions, or a rate of correct predictions, and then a similar logic applies.

Transport assumption. Transport assumptions assume invariance and come in a strong and a weak form. The strong form means the simulator makes no systematic error at both estimands, so it does not need to be tested at T against human data. The simplest version of the weak form, by contrast, only says that the error stays the same, i.e. $b_T = b_A$. In that case, the simulation output at T can then be corrected by $\hat{b_A}$ and will be right however large that error is. The strong form is a special case of the weak form where this error equals zero. Both the strong and weak forms require the model to get the change from A to T right, yet only the strong form implies that the model also gets the level right. There are more general versions of the weak transport assumption. Dell and Rambachan (2026) allow the error to differ between settings as long as that difference is predictable from observables. This requires more data at A to estimate this relationship. A different transport assumption is exchangeability (Andrews et al., 2025). Here, the settings that are similar to the setting of interest are treated as draws from a population of settings. Comparing the model's errors across pairs of validated settings yields a forecast interval for how far its error at the setting of interest departs from its error in a validated setting. This gives a bound that holds with a stated probability, but only under the exchangeability assumption itself. The idea of exchangeability specifies what we mean in this review when referring to a new setting "of a familiar kind." The concept rules out systematic differences between the validated settings and the setting of interest that are predictable in advance, such as a time trend. As a result, its guarantee only holds for a setting that's drawn *at random* from the population that's used to construct it. If the setting of interest is believed to be different from the validation ones—which it often is in practical cases—the concept cannot be applied. This same assumption applies to training a model on "nearby" behavioral data (Section 4).

How demanding a transport assumption is depends on how the target estimand differs from the validated estimand and on what evidence connects the two. Take the case of two estimands with different resolution. As the estimand becomes finer, little carries over. Specifically, a sample that pins down a population mean potentially tells us nothing about subgroup differences, and aggregate summaries tell us nothing about how responses relate within a person. Instead, a finer estimand needs to be validated at its own resolution.[4] When the outcome type changes

[4]Conversely, a fine validated estimand validates a coarser but otherwise identical target estimand, as long as

between estimands, one needs evidence that accuracy on one outcome, say stated intentions, carries over to the other, say realized choices. Finally, if the population or context changes, the evidence needs to support that the error measured in the validated setting is the same for new people or new conditions. This is a demanding requirement: validation across several settings helps but in principle extensions to a new setting require a theory of context, i.e. an account of which features of a setting affect simulation error and how. Theories of context in behavioral research are incomplete. This is why invariance of a simulator to changes in context tends to be relatively hard to justify as the foundation for an AI-based behavioral simulation.

Uncertainty. An evaluation of the simulation approach can also assess whether the reported precision of an estimate is calibrated. Calibration means, for example, that the 95 percent interval around the estimate contains the human quantity in 95 percent of repeated samples. Checking this requires human data. An interval can only be calibrated if something links it to the human quantity, such as the sampling design, the treatment assignment, or a model of the estimation error. As discussed above, rerunning the simulation provides no such link because it reveals how much the procedure varies around its own mean rather than how far that mean lies from the human quantity. We give more details on valid measurement of uncertainty in Section 2.4.

2.2 Accuracy in Validated Settings

2.2.1 Granularity

With a structure for evaluating the evidence on AI-based simulation in hand, we now turn to using that framework by first reviewing findings on simulation accuracy where it has been checked, organized by the different dimensions of an estimand. First, how does the accuracy of predictions vary with the granularity of the estimand?[5]

Treatment effect rankings and magnitudes. Treatment effects simulated using synthetic respondents reproduce the ranking of effects well but overstate their size. Across 469 treatment effects from 70 nationally representative U.S. survey experiments, effects computed from GPT-

the fine-level validation concerns calibrated probabilities or distributions.

[5]One caveat applying throughout this subsection is that human benchmarks become noisier as the estimand becomes finer. Subgroup means, distributional moments, and small effects are estimated less precisely than population averages, which attenuates correlations and inflates measured errors even for a perfect predictor. Part of the decline we document can therefore be viewed as mechanical. We note where the underlying studies allow a correction for this, and otherwise interpret the finer-grained comparisons as upper bounds on the error.

4-simulated respondents correlate with observed effects at 0.85 (Ashokkumar et al., 2026).[6] The magnitude of simulated effects, however, is 70 to 80 percent too large. The two findings are compatible because a correlation ignores a common scale factor: overstating every effect by the same proportion leaves the ranking intact. Direct forecasts provide promising evidence on magnitude prediction. Vivalt (2026) evaluates forecasts on a screened sample of 673 effects from 33 papers, with model-specific subsets determined by reported knowledge cutoffs. On its eligible subset of 339 effects from 20 papers, GPT-6 Astra achieves a median absolute forecast error of 0.054 SD. Its mean squared error is 0.0218, less than half the 0.0463 obtained by predicting the leave-one-paper-out mean effect. Simulated respondents may thus be useful for screening which treatments look most promising overall, but for applications that depend on effect sizes, the most promising evidence so far comes from using LLMs as experts that make direct forecasts.

Response distributions. How well do AI-based simulations match the distribution of responses or behavior? Persona-based samples tend to reproduce survey means approximately while understating empirical response variation (Bisbee et al., 2024). Such excessive homogeneity also appears across repeated experimental responses, among simulated forecasters, and within identity groups (Park et al., 2024; Iadisernia and Camassa, 2025; Wang et al., 2025), even as differences between groups are often exaggerated (Wang et al., 2025). Synthetic responses also shift with answer order and miss how people react to changes in wording (Dominguez-Olmedo et al., 2024; Tjuatja et al., 2024).

The existing evidence also shows that distributional accuracy hinges on how the model is used, as expert forecaster or by simulating synthetic respondents. Holding both model and target fixed, asking for a population's distribution usually outperforms aggregating simulated responses (Meister et al., 2025). Relatedly, when LLMs state a distribution and are then asked to draw from it, they tend to return the same answer almost every time rather than reproducing that distribution through their draws (Jang et al., 2026). Two remedies help in part. One can request several answers with probabilities in the same prompt (Zhang et al., 2025), and fine-tuning on survey data reduces discrepancies by up to 46 percent (Suh et al., 2025). Overall, however, the findings suggest that respondent simulation fails to express distributional information that the model has. An LLM prompted to generate a distribution of answers through individual draws should therefore not be thought of as a genuine statistical simulator of that distribution.

[6] Among small effects, the correlation falls to 0.25, partly reflecting restricted variation and noise in the human estimates.

Subgroup patterns. The evidence on subgroup predictions is mixed. Prompting with specific demographics has been shown to recover some opinion patterns in U.S. surveys (Argyle et al., 2023). However, simulations capture certain groups far better than others and group-specific prompting only partly closes those gaps (Santurkar et al., 2023; Durmus et al., 2023). For example, (von der Heyde et al., 2024) use simulations on German vote-choice data and find that some partisan tendencies are well captured while population and subgroup vote shares are misestimated. In Ashokkumar et al. (2026), predictions are relatively accurate within gender, ethnic, and partisan groups. Yet, persona-based samples that match population means well may miss some subgroup means by 10 to 20 points on a 100-point scale. In the data of Bisbee et al. (2024), 48 percent of coefficients relating opinions to demographics differ significantly from human estimates.

Individual responses and digital twins. Is it possible to use rich information about an individual to build a digital twin that reliably represents that individual's preferences? To address this question, Park et al. (2026) build an agent for each of 1,052 participants from a two-hour qualitative interview. These agents predict held-out survey answers with 65.7 percent accuracy, compared with 58.1 percent when the prompt contains only four demographic attributes. We view this difference in prediction accuracy as fairly modest, especially given the costliness of collecting the qualitative interviews.

The study created widespread excitement about AI-based behavioral simulation on the basis of its headline accuracy of 83 percent. Yet, this accuracy number is produced by dividing accuracy by a person's own test-retest reliability: how often participants repeat their own answers after two weeks, which was 79.5 percent. A person's own test-retest agreement, however, is not necessarily the right upper bound for a simulation's accuracy. Suppose someone answers A to a binary question with probability 0.8 every time they are asked. Two independent answers then agree with probability 0.68, which is the test-retest score. A simulation only needs to predict once, so always predicting A, the person's modal answer, achieves accuracy 0.8, *above* the test-retest score of 0.68. More generally, under independent draws from a stable response distribution, the best fixed prediction has accuracy between R and $\sqrt{R}$, where R is retest agreement (Appendix A). At $R = 0.795$, this gives 79.5 to 89.2 percent. This range is a reference point under stated assumptions, not an estimate for these participants. Normalizing by retest agreement can therefore overstate how close a simulation is to the best attainable prediction. Normalizing the raw accuracy of 65.7 percent by the upper bound of 89.2 percent instead would give 73.6 percent. The "naive" benchmark with four demographics would then stand at 65.1 percent, so the two-hour interview buys less than the headline number suggests.

In a related study, Toubia et al. (2025) build digital twins from more than 500 answers per

person for 2,058 participants, withholding behavioral tasks for evaluation. The main specification achieves a 71.7 percent score, compared with 59.2 percent for random guessing and 81.7 percent for participants repeating the tasks two weeks later.[7] The twins thus score about twelve points above chance and about ten points below retest performance, a benchmark subject to the reservation above.

The simulation errors in these types of applications also reveal systematic departures from human behavior. The twins reproduce several classic experimental results but reverse the sunk-cost effect and choose the payoff-maximizing option too often in a probability-matching task. In a hypothetical vaccination decision, 4 percent refuse vaccination, compared with about 45 percent of real respondents. These patterns suggest a tendency toward normatively prescribed responses rather than observed behavior.

2.2.2 Outcome Type: Reports and Actions

Most studies evaluate predictions of reported attitudes, beliefs, or intentions. In our evidence map, 71 of 108 studies concern reports or aggregate quantities rather than actions. Survey answers are a legitimate target and, in some cases, the very thing a user wants to learn about or predict. They are not the same target as what people do when something is at stake (LaPiere, 1934), partly because of the intention–behavior gap (Sheeran and Webb, 2016) and social desirability bias (Bursztyn et al., 2025).

Few of the studies we reviewed permit testing the difference in the same sample. Park et al. (2026) is one exception, finding that interviews that improve survey prediction accuracy from 58 to 66 percent do not detectably improve predictions of participants' choices in five incentivized economic games (Park et al., 2026). While additional personal information helps with one outcome type, it does not help with the other. This highlights that accuracy for self-reports cannot substitute for validation against the actions a simulation is meant to predict.

2.2.3 Same Estimand, Different Estimators

For a fixed human target, changing how the model is used can substantially change the estimate. Validation therefore applies to the full procedure.

Direct forecasts and simulated respondents. Asking an LLM to describe a response distribution can yield a closer match to human data than sampling responses from it (Meister et al., 2025). Direct forecasts and simulated respondents can also produce sharply different

[7]The score is not raw accuracy: it rewards exact categorical matches and closeness on numerical answers; it is not the percentage of answers predicted exactly.

effect sizes. Across 12 effects, GPT-4's direct forecasts average 13.2 times the effects generated by simulated respondents, or 5.3 times after excluding the largest discrepancy (Manning et al., 2024). Without a human benchmark, this comparison cannot establish which procedure is more accurate, but it shows that the two are not interchangeable. The two best human-benchmarked magnitude results point in a similar direction. Effect sizes from tallying simulated respondents are inflated (Ashokkumar et al., 2026), whereas direct forecasts land closer to realized effects (Vivalt, 2026). The two studies are not a controlled comparison, however, because they differ in the field of application, the outcome type, the error metric and the models used. Moreover, the result in Manning et al. (2024) demonstrates that direct forecasts can also be the larger of the two. No study has yet compared *both* estimators against their corresponding human magnitudes. Until that happens, the advantage of direct forecasts for effect sizes remains suggestive.

Richer personas. More information about a person does not reliably improve predictions of their behavior. Adding qualitative interviews improves predictions of survey answers but not choices in games (Park et al., 2026). Across hundreds of experiments, adding a participant's age, nationality, education, and questionnaire answers leaves predictions of their choices virtually unchanged (Binz et al., 2026). Prima facie, it might come as a bit of a surprise that richer descriptions do not lead to greater behavioral accuracy, because we often have the feeling that when we "know a person", we are somewhat able to predict what they would do in a situation. We conjecture that part of the reason is that the domain-specific predictive power of what a person says in an interview or does in other decisions is in fact quite low. This is an old finding. Correlations between a person's behavior in different situations are typically modest (Mischel, 1968; Ross and Nisbett, 1991), and the same holds for economic preferences measured in different domains and for behavioral biases measured in different tasks (Einav et al., 2012; Dean and Ortoleva, 2019; Stango and Zinman, 2023; Chapman et al., 2023). There is just not that much signal. Even if the AI-based simulation managed to fully extract it, the gain in prediction would be small. We view this as a general challenge to the approach of AI personas, which in recent years have often been promised to stand in for real people in a way that can be used for any question one might ask of them. While tempting, the evidence collectively suggests this is unlikely to happen.

Behavioral data and theory. Training on observed behavior and incorporating explicit theory can improve predictions. A model fine-tuned on 10.7 million human choices from 160 experiments predicts choices by new participants and behavior in new tasks (Binz et al., 2025). Fine-tuning on survey data improves predicted response distributions for new questions and groups (Suh et al., 2025). Theory-based agents assign higher probability to observed play in

strategic games than do cognitive hierarchy and Nash equilibrium benchmarks (Manning and Horton, 2025).

2.2.4 Predicting How Uncertain People Are

Often we do not only care about what people do, but also about how confident they are, which is inextricably linked to their actions. Two people can make the same choice while differing substantially in how sure they are that it is best. This uncertainty predicts how strongly people respond to fundamental parameters of decision problems (Enke and Graeber, 2023; Enke et al., 2024). Existing evidence suggests two sources of predictive information. Features of the decision problem predict average stated uncertainty (Enke and Shubatt, 2023; Zhu et al., 2025), while a person's recent confidence ratings predict her confidence across tasks and domains (Mei et al., 2023).

Can LLMs predict human uncertainty from problem descriptions and observed confidence histories? Existing evidence suggests that LLMs predict human confidence patterns only weakly (Bombassei De Bona et al., 2024; Huff and Ulakci, 2025; Cash et al., 2025).

2.3 Does Accuracy Generalize Across Settings?

Section 2.2 shows how accurate simulations are where human data can check them. We next ask whether that accuracy carries over to unchecked settings, particularly those that differ in kind: settings that change a feature all validated settings share, such as the population, the stakes, the rules, or the time period. Answering this requires checking the same estimator against human data in two settings that differ along such a feature. This evidence, as of now, is still scarce. Of the 108 studies, 71 primarily target reports rather than choices, and few validate an estimator in one setting and then test it in another. We organize the available evidence into three categories: new settings of a familiar kind, different populations, and different contexts.

2.3.1 New Settings of a Familiar Kind

AI-based simulation methods can predict behavior in some new tasks and settings. A correlation of 0.90 for experiments unpublished at the training cutoff provides evidence against simple recall of effects that are covered by the training data (Ashokkumar et al., 2026). Vivalt (2026) evaluates forecasts of results released after models' reported knowledge cutoffs. On its eligible sample of 339 effects from 20 papers, GPT-6 Astra achieves a median absolute forecast error of 0.054 SD and less than half the mean squared error of the leave-one-paper-out mean benchmark. A model fine-tuned on laboratory choices predicts behavior in new, similar laboratory

tasks better than without fine-tuning (Binz et al., 2025), while agents built around a theory of strategic reasoning predict play in new games from a prespecified family (Manning and Horton, 2025). Models given news through a specified date track professional, household, and consumer expectations across quarters (Bybee, 2023; Graham et al., 2026; Anesti et al., 2026; Wu et al., 2025; Iadisernia and Camassa, 2025; Huang et al., 2026). In a field experiment on delivery options at an online retailer, a correction estimated from human data in 10 percent of regions also works in the remaining regions (Zhang et al., 2026). While these transfers to new settings are meaningful, they remain close to those used for validation.

2.3.2 Across Populations

Accuracy strongly varies across populations. Models capture some groups' opinions better than others (Section 2.2.1), and a model that captures some German partisan tendencies misstates vote shares (von der Heyde et al., 2024). In a test spanning 74 countries, Chen et al. (2026) use 12 model configurations to simulate the preferences of 67,201 Global Preference Survey respondents. The simulations explain just 1.3 percent of individual-level variance and 4.2 percent of variance in country averages. They reproduce differences by gender, age, and cognitive ability in 78 percent of comparisons, but differences between countries ranked high and low on standard cultural dimensions in only 48 percent. Removing country information from prompts barely changes accuracy. Models also make larger errors in poorer, less digitally connected countries. This pattern is consistent with uneven coverage of populations in training data. Human data from the target population can help measure and correct these errors.

2.3.3 Across Contexts

The evidence above illustrates why invariance is inherently context-dependent. Accuracy across tasks or populations does not establish accuracy across periods, institutions, or policy rules.

Not only is it hard to extrapolate from findings about contextual changes, but such evidence is also scarce. The regional experiment provides the one test across time: a correction learned from the first week of data reproduced that week's effect of about −40 basis points but not the −60 basis-point effect over the full two months, as behavior changed over time (Zhang et al., 2026). The correction transferred more successfully across regions than across time, but we note two things about this test. First, the simulated customers were not told anything about time. Their prompt was the same in week one and in week eight, describing just the person, the product, and the delivery options. The simulation could not have tracked a change over time directly because no change was presented to it. This means the finding does not show what a model would have done had it been informed. Second, in fact nobody could have informed it to

begin with. Zhang et al. (2026) note that the drift could have come from customers adapting to the new options, from new customers entering, or from holidays, and that this was not known while the experiment was running. In a way, this application nicely illustrates the practical problem that in a new context, the user often does not know which features have changed, and a prompt can only describe what the user knows.

2.4 What is Valid Inference From Synthetic Data?

When we measure or predict human behavior, a standard error is supposed to tell us how far our estimate may be from the human quantity we care about, e.g. the actual take-up rate. Standard errors derived from human data do precisely this. Put simply, under random sampling, a 95 percent interval around the sample mean contains the true rate in 95 percent of samples. Standard errors computed from synthetic data, however, do not have this property. Valid inference from simulated data requires something that links the estimate to the human quantity, which could be a sampling design, randomization, or a justified model of the estimation error. In this section, we argue that the simulation alone provides none of these.

Several numbers that come out of simulations can "look like standard errors" and are sometimes reported or treated as such. Each of these constructs measures something different, and none is a standard error in the sense relevant for statistical inference. The spread across repeated runs measures simulation noise, i.e. how much the procedure varies around its own mean. The model's stated confidence is just another output of the same procedure. It claims to capture how likely the simulated answer is to be right, but whether it is or not is an empirical question that can only be answered by checking against human outcomes. The spread across simulated respondents is a prediction of how much people differ, which is itself a quantity to be validated, not an uncertainty about the mean. The spread across prompts, models and settings measures how sensitive the result is to the user's choices. None of these, on its own, tells us how far the estimate is from the human quantity.

2.4.1 What the Quantities Measure

Repeated draws and token probabilities. Repeated runs estimate the distribution of a fixed procedure's answers. Where available, token probabilities can estimate this distribution more efficiently than tallying sampled answers (Sun and Zhang, 2025). Either approach characterizes the model's output rather than its error relative to people.

Let μ^M denote the procedure's expected output and θ the human quantity. The error of an

average over repeated runs has two parts:

$$\bar{\theta}^M - \theta = \underbrace{\left(\bar{\theta}^M - \mu^M\right)}_{\text{simulation noise}} + \underbrace{\left(\mu^M - \theta\right)}_{\text{systematic error } b} . \tag{1}$$

More independent draws reduce simulation noise but leave systematic error unchanged. This means that the interval shrinks around μ^M rather than θ. As a result the interval's coverage of the human quantity falls toward zero as draws accumulate whenever $b \neq 0$. Consistent with this logic, Huang et al. (2025) find that adding simulated respondents narrows reported intervals while reducing coverage of human quantities. Cui et al. (2025) likewise report that many effects that are statistically insignificant in human data become significant in simulations of 156 experiments.

Spread across respondents and specifications. The variation among simulated respondents is a prediction of human heterogeneity, and its accuracy must be validated against human data (Section 2.2.1). Variation across prompts, models, and settings is a measure of sensitivity to specification choices. By construction, neither kind of variation yields a valid confidence interval without further assumptions, and averaging across models does not necessarily remove the errors they share (Kim et al., 2025).

2.4.2 The Calibration of a Model's Stated Confidence

Calibration of stated confidence means that, for example, when the procedure claims to be 70 percent sure, it is right about 70 percent of the time. AI-based simulations can be asked to state such a confidence, and a model's token probabilities can similarly be read as a measure of its confidence. It is tempting to take these numbers at face value. They are only useful to the extent that they are calibrated, however, and this has to be checked for each type of application. The evidence so far comes mostly from questions with checkable answers rather than from predictions of human behavior.

Most evidence comes from multiple-choice questions with objectively correct answers. The two measures, stated confidence and token probabilities, do not need not agree and, in fact, they typically don't. Stated confidence usually lies between 80 and 100 percent, stays high even when the answer is wrong (Xiong et al., 2024; Leng et al., 2024), and depends on how the model is prompted (Tian et al., 2023; Zhou et al., 2023). A model's token probabilities tend to be fairly well calibrated before post-training and less well after it (Kadavath et al., 2022; OpenAI, 2023). For post-trained models, stated confidence has better calibration than token probabilities but tends to be too high (Tian et al., 2023; Xiong et al., 2024). These results,

again, are about answering questions with checkable answers, not simulating human behavior. There is far less evidence for human targets, and it points to various forms of miscalibration. When predicting socioeconomic outcomes for individuals described in text, base models hedge too much and instruction-tuned models too little. Asking directly for a numerical probability helps some models but only on some tasks (Cruz et al., 2024).

For a population mean or treatment effect, the relevant question is instead whether reported intervals contain the target quantity at the stated rate. For direct forecasts of average effects, such coverage has, by now, been measured just once. Vivalt (2026) finds that the 10th to 90th percentile intervals that models attach to their forecasts of economics treatment effects contain the realized effect at about the stated rate. Wider intervals correctly mark the effects that turn out harder to predict. This result shows calibration against human outcomes for one target and procedure as required by the argument. It does not necessarily carry to other targets or to settings of a different kind. Nor does calibration alone imply that predictions distinguish among people: a predictor that assigns the population take-up rate to everyone is calibrated in that population without identifying who takes up the program. Reported uncertainty therefore needs validation for the target and procedure actually used. Extending that validation to a new setting requires additional justification.

2.4.3 Human Data Restore Valid Inference

One constructive response to the lack of standard errors from AI-based simulation is to directly build human outcome data into the design. When those data come from precisely the target situation that's being simulated, they can be used to estimate and correct the model's error. If they can only be obtained for a similar but different setting, the resulting interval is valid only under a transport assumption (Dell and Rambachan, 2026).

Human data from the setting of interest. The logic of the first case is straightforward. A small human "calibration sample" is used to measure how far off the model's predictions are on average. The model is then used to predict the outcome for a much larger set of people, and the measured error is subtracted from the average of these predictions. For a mean,

$$\hat{\theta} = \frac{1}{N}\sum_{j=1}^{N} f(X_j) + \frac{1}{n}\sum_{i=1}^{n}\big[Y_i - f(X_i)\big], \tag{2}$$

where $f(X)$ is the model's prediction for a unit with characteristics X. The first sum runs over a large sample without human outcomes, and the second over a smaller sample in which the human outcome Y is observed. The second term estimates $-b$, the systematic error in Equa-

tion (1). Prediction-powered inference and design-based supervised learning extend this same approach to regression coefficients and other targets (Angelopoulos et al., 2023a; Egami et al., 2024; Ludwig et al., 2026; Broska et al., 2025). We note two properties of this approach. First, validity comes from the human sample itself. Provided both samples are drawn from the target population acting in precisely the situation of interest, the resulting corrected estimate of the simulation is unbiased however poorly the model predicts prior to that correction. Second, precision comes from the model. Once N is large, the estimator's variance is approximately $\text{Var}[Y - f(X)]/n$. A model whose errors vary little across units leaves little for the human sample to improve. In that case, a poor model yields no gain and, in the form of Equation (2), can even do worse than the human sample alone.[8] Precision is therefore capped by n, and more model output does not raise this cap. In simulations, 125 to 250 human observations deliver intervals with the intended coverage while the model improves precision (Ludwig et al., 2026). In survey simulation, correcting against a human subsample can remove the bias that prompting alone leaves (Krsteski et al., 2025). Two literatures describe and use this estimation approach in opposite directions. In statistics, people start from the human sample and use the model's predictions to make it more precise. For simulation, users start from the simulation and use the human sample to correct it. The basic formula is the same and in both cases the validity comes from the human sample.

Human data from other settings. In the second case, the human data do not come from the setting of interest itself, so that the correction does not equal the simple difference in Equation (2). As a result, the guarantee for removing simulation error no longer comes from the sampling design alone. Leveraging human data from "nearby" situations has not been widely done for behavioral simulation yet, but we illustrate the approach with the two studies we're aware of. Zhang et al. (2026) simulate customers in each region of their field experiment, then fit a correction to observed outcomes in 10 percent of the regions. In the remaining regions, the corrected simulation yields an effect of -65 ± 10 basis points (95 percent interval), against -60 ± 8 from the full experiment and -57 ± 27 from the human data in the 10 percent of regions alone. In this application, the model narrows the interval substantially. While the regions used for fitting were randomly drawn, the correction itself is a fitted model for that subset of regions. Its validity rests on that model being right for the regions it was not fitted on. The bootstrap behind the interval captures how much the fitted correction varies across samples, but not whether the model is misspecified to begin with. It therefore does not guarantee coverage, i.e. that a 95 percent interval contains the true effect in 95 percent of cases. Consistent with this potential concern, the same correction does not carry over from the first week to later

[8] Variants that tune the weight on the model's predictions rule this out (Angelopoulos et al., 2023b).

ones (Section 2.3.3).

Huang et al. (2025) instead use human data on benchmark questions to specify the number of simulated respondents in such a way that the resulting intervals from the simulated sample contain the human quantity at the intended rate on average across questions. The chosen number is the size of a human sample that is "as noisy as the model is wrong", or, to put it differently, it captures how many respondents the model is "worth." The guarantee by construction holds on average over questions like the benchmark questions, but not for each new one. Moving to new questions therefore requires the assumption that they are like the benchmark questions, and the number has to be re-estimated for each model.

Implications for simulation design and reporting. In short, uncertainty must be established for the claim being made. For a population mean, the standard error must reflect both the human sample and the estimated correction. For one person's choice, the relevant object is a calibrated probability (Section 2.4.2). If many prompts or configurations were tried, the analysis should say so (Baumann et al., 2025). We believe that a precision gain is best reported in terms of the number of human respondents saved a la Huang et al. (2025), i.e. how many fewer observations reach a given precision. The quantities produced by an AI-based simulation approach cannot certify validity in and of themselves, but they can direct human data. Any disagreement across draws or models can flag where human observations are likely to be most informative. In Gligorić et al. (2025), stated confidence measures have been used to allocate a fixed budget of human labels so that a given precision is reached with fewer of them.

Note that one property of simulation outputs can be checked without any human data, which is its internal consistency. Often forecasts produce quantities that should be logically related and consistent with a single probability distribution, such as the share of households that apply for a benefit, the share that apply and are approved, and the share that apply and are rejected. Andrews and Sarkar (2026) measure departures from this requirement by the profit an arbitrageur could guarantee by betting against the stated probabilities. This profit equals the largest gap between the forecasts and the nearest coherent set of probabilities. Because the human quantities are themselves coherent, it is a lower bound on the largest error among the related forecasts, and it can be computed at the setting of interest. This check can reveal obvious error, which is not the same as establishing accuracy. A forecaster that is confident and wrong can be perfectly coherent.

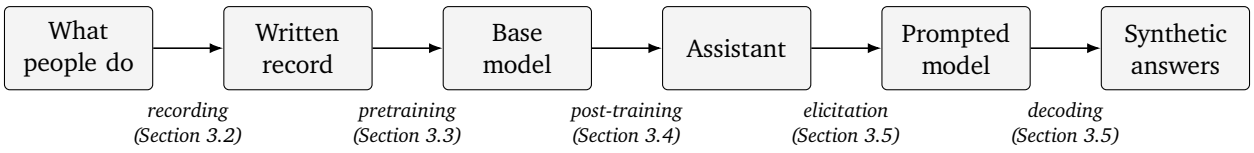
3 Why Behavioral Prediction Fails Where It Does

The preceding section shows that human data reveal systematic patterns in the errors of AI-based simulations, affecting how we should think about whether a given procedure can be trusted when applied to a new question in a new setting. This section asks why the overall picture of specific types of errors emerging from Section 2 arises. We trace a possible explanation through the stages that turn records of human behavior into synthetic answers. We end by asking why these losses might persist as models improve.

3.1 From What People Do to Synthetic Answers

Figure 2 traces the path from human behavior to a synthetic answer in five stages. First, recording turns what people do into a selective written record which is the foundation for large language models. Second, pretraining teaches the model the regularities in that record. Third, post-training tunes the model toward answers that raters or automatic checks approve. Fourth, elicitation frames the task through the prompt and persona. Finally, decoding samples the synthetic answers. At each stage, information needed to reproduce or predict human decisions may be missing, lost, or distorted.

Figure 2: From what people do to synthetic answers



3.2 The Record

Pretraining corpora are so large that the answer to almost any question about behavior can seem to be somewhere in them. By *the record* we mean the data a model is pretrained on, as the developer assembled it. For current models this record overwhelmingly consists of text, including code, transcripts and tables rendered as text, with a smaller share of images, audio and video. We use "written record" as shorthand for this. Data that enter later, in post-training or in fine-tuning on behavior, are treated as separate stages below.

What gets recorded. There are several sources that constitute the foundation for the record. It comprises (i) descriptions, such as accounts, summaries of studies, and people saying what they would do, (ii) direct traces, such as reviews, forum posts, and logs (Horton, 2023), and

(iii) tables written out as text, which are abundant (Eggert et al., 2023). Records that link a person's characteristics and circumstances to that person's choices, such as survey microdata, experimental datasets, and administrative files, tend to be restricted, and public ones often come in formats that text pipelines do not ingest.

What gets collected. Which kinds of data get collected? Most pretraining text comes from crawls of the public web, supplemented by code, books, encyclopedias, academic papers, licensed data, and text written by other models (Brown et al., 2020; OpenAI, 2023; Team OLMo, 2025). Llama 3's roughly 15 trillion tokens, for example, are about half general knowledge, a quarter mathematics and reasoning, 17 percent code, and 8 percent non-English text (Grattafiori et al., 2024). The composition of text is necessarily a reflection of the goals that developers have when creating these models and thus not a representative sample of what is written.

What survives curation. The collected text is then deduplicated, filtered for quality, and reweighted by source. Some developers do this in-house (Brown et al., 2020; Grattafiori et al., 2024) but many models are instead trained on public corpora whose creators have already made these choices (Penedo et al., 2024). To maintain high quality, classifiers are trained to identify educational text (Brown et al., 2020; Penedo et al., 2024). For open corpora, these steps can be inspected and their output audited (Elazar et al., 2024). For the closed models these steps mostly remain a black box: developers disclose little about their training data, and disclosure has declined over time (OpenAI, 2023; Wan et al., 2025). We know of no audit of any corpus for its behavioral content.

What this implies for behavior. The above filters suggest that the record contains more statements about what people tend to do than about who specifically did something, under what circumstances, and why. Moreover, curation favors text that describes behavior in the words of those who analyze it, and publication favors notable effects (DellaVigna and Linos, 2022).

3.3 Pretraining

Pretraining is the initial stage of training where the model learns to predict each next token (a word or part of a word) in the record from the text that precedes it. The result is a *base model* that continues text but is not yet tailored to answer questions or follow instructions.

The training objective at this stage, cross-entropy, is minimized when the model's probabilities equal the frequencies of continuations in the record (Gneiting and Raftery, 2007). If

descriptions of a situation are followed by option A 40 percent of the time, a model at this optimum assigns A a probability of 0.4. A pretrained probability therefore estimates how often an answer follows a description of that kind in the record, not how often people give it.

Real models fall short of this optimum in an uneven way. Compressing an enormous record into a model preserves frequently recurring patterns better than rare ones (Kandpal et al., 2023), so information can be lost even where the record contains it. Where the record does link personal characteristics and circumstances to behavior, any particular combination is extremely rare. This sparsity is a matter of arithmetic. Suppose we describe a person by ten characteristics with just five levels each and a situation by ten features with five levels each, an extreme simplification of the actual variation in humans and the situations we face. That already gives 5^{20}, or about 95 trillion, distinct cells. For comparison, Llama 3 was trained on about 15 trillion tokens, so this unrealistically stripped-down model of behavior has more cells than there are tokens in the whole training corpus. Real situations have far more than ten relevant features (Section 2.1). The model may retain how people of a broad category tend to behave in situations of a given kind, while losing what distinguishes a particular person in a particular situation. It can still produce a prediction for that person in that situation, but the record gives it nothing to base it on. As a result, the prediction is necessarily filled in from the coarser patterns it did retain. Additional persona details would then add little, even where they matter for the behavior of actual people. This could contribute to why simulations forecast broad patterns better than individual responses.

Whether these losses actually occur at the pretraining stage can be checked by looking at base models, which have not yet been post-trained. Across three model families and hundreds of experiments, adding participants' demographics and questionnaire answers to the prompt barely improves predictions of their choices, for base and post-trained models alike (Binz et al., 2026). Base models also compress opinion differences between demographic groups to about 70 percent of their observed size (Grief-Albert et al., 2026). Yet when sampled as respondents, they come closer than their post-trained versions to the spread of human answers (Grief-Albert et al., 2026; Jang et al., 2026), as a frequency-matching objective predicts. The available evidence does not allow us to disentangle whether the relevant relationships are absent from the record, present but not learned, or learned but not elicited by the prompt.

3.4 Post-training

Post-training tunes models toward answers that raters perceive to be more helpful, accurate, or safe (Ouyang et al., 2022; Bai et al., 2022a; DeepSeek-AI, 2025). Directly comparing the behavioral prediction accuracy of the same model before and after post-training reveals that

post-trained models assign lower probability to the choices people made than their base models do (Binz et al., 2026). As a result, post-training decreases how well models predict human responses to question wording (Tjuatja et al., 2024).

One first consequence of post-training is less varied answers. Because raters favor typical answers, rewards favor what the model already finds likely, and post-training further narrows toward it (Zhang et al., 2025). Output diversity falls through successive post-training steps, both in general text generation (Zhang et al., 2025; Kirk et al., 2024) and in simulated populations (Murthy et al., 2025). While post-training tends to produce less simulated variety within groups, it tends to exaggerate opinion differences between demographic groups (Grief-Albert et al., 2026).

The second consequence is a bias towards the textbook answer when accuracy is rewarded. Reasoning training, which rewards answers that an automatic check marks as correct, should push furthest in this direction. Consistent with this, reasoning-trained models depart further from people than instruction-tuned ones (Binz et al., 2026), and reasoning training impairs predictions of human disagreement, whereas prompting a model to reason step by step does not (Ni et al., 2026).

That said, post-training does not universally move predictions away from people. Binz et al. (2026) find that fine-tuning on an additional database of human choices can improve specific behavioral predictions. Because this changes both the overall training corpus ingested by the resulting model and the objective alike, it cannot tell us which of the two is responsible. It does suggest, however, that the problems documented above come from the particular data and objectives used in standard post-training.

3.5 Elicitation and Decoding

In the final two steps, the model is deployed into a particular estimator. The first step is elicitation. The prompt, including any persona, frames the task, and the rule by which answers are tallied turns the outputs into an estimate. The person and the situation reach the model only as a description. Choosing which features to describe is an implicit theory of context, and unmentioned features are filled in by association by the simulation procedure (Gui and Toubia, 2025).

The second step is decoding. The model's probabilities need to be turned into sampled answers. A key lever here is the temperature, which determines how widely answers will vary. Lower temperatures entail lower model variability, while higher temperatures imply that less likely answers appear more often. Truncation discards low-probability options before an answer is drawn (Holtzman et al., 2020). One might hope that temperature and truncation can

be used to capture the heterogeneity in human responses, but there is no reason to expect that they do, since they rescale the model's probabilities rather than add information about people. Consistent with this, GPT-4o gives the same answer in at least 98 percent of calls, and changing decoding settings does not fix this (Jang et al., 2026). Moreover, synthetic populations generated by varying temperature and prompts remain less diverse than human ones (Murthy et al., 2025).

3.6 Why the Gap May Persist: The Cost of Verification

Will better models simulate people more accurately? To improve performance, a model needs feedback on whether its own answers are right. Language models have improved fastest where this feedback is cheap and automatic: a test suite checks a program in seconds, and a reference solution checks the answer to a mathematics problem. Post-training can then reward correct answers in a loop that runs without people and can be repeated millions of times (Lambert et al., 2024; DeepSeek-AI, 2025). Wei (2025) conjectures that the ease of training a model on a task rises with the ease of verifying its answers.

How easy is it to verify behavioral predictions? The only way to verify a behavioral simulation's predictions is to observe the relevant people in the relevant setting with sufficient knowledge of the contextual factors, i.e. it comes down to field experimental data. This check is slow and expensive, because it requires recruiting and incentivizing people and sometimes observing them for weeks (Zhang et al., 2026). To draw reliable inference from such experimental data, large datasets are needed to ensure sufficient statistical power. Moreover, for specific prediction targets where simulations are most useful, such as a policy not yet introduced, it is not possible to even collect the human data.

This has an implication that we're stating as a conjecture. Gains in general model capabilities concentrate where verification is cheap and need not improve behavioral prediction. Progress in behavioral prediction will instead be driven by the supply of high-quality human data that link people, situations and choices. There are two objections to this view worth mentioning. First, supervised training on recorded behavior does not need a verifier at training time. While true, the bottleneck then moves to collecting that behavior in the first place, which is the same constraint under a different name. Second, large platforms run thousands of experiments a year, meaning that the supply of checks is large and growing in some domains. Our conjecture predicts that progress will come first in exactly these domains and slowest for offline behavior, where data collection is costly. The evidence so far is thin. In one model family, the gap between post-trained models and their base versions in fitting human choices has widened across generations (Binz et al., 2026). Our conjecture on the role of the cost of verfication is testable over

time. If correct, the reduced-form pattern belongs to the general use case of behavioral prediction and not to the models of 2026. We predict that it persists where checks stay expensive and fade wherever they become cheap.

4 Credible Uses

Any use of a simulation where it has not been validated requires an invariance assumption. The three routes we discuss differ in where they place this assumption and in how it can be checked. In the *theory route* it comes from behavioral theory. The modeled mechanism is assumed to govern behavior in the setting of interest. The *training route* places it in additional behavioral training data. The setting of interest is assumed to belong to the family of settings that the additional behavioral data cover, so that accuracy measured on that family carries over. The *correction route* places it in the error of the simulation method. Here, the invariance effectivley amounts to assuming that the simulation gets the change from a nearby validated setting to the setting of interest right, which implies that the error measured in the former can be accounted for in the latter. Both the training and correction routes are reduced form in nature. They do not supply behavioral structure. Table 2 summarizes the three routes. Note that they are not mutually exclusive. Even after adding theory and fine-tuning, a simulation will likely have systematic error that can be measured and corrected.

Restricting the claim. Sometimes, none of the three routes may provide a viable path to a credible simulation. In those cases, the baseline approach should be to restrict the claim to only those types of things that existing validation establishes. In particular, the average effect of a familiar intervention is far more reliably forecast than individual responses. For such cases a direct forecast with a predictive interval is now the best-supported use (Vivalt, 2026). When the structural, training and correction routes are unavailable, simulations may also still be useful in suggesting hypotheses, survey questions and candidate treatments (Horton, 2023; Manning et al., 2024), which complements and guides human testing rather than replacing it.

Compared with what. Whatever claim is made, the usefulness of a simulation should be judged against the alternatives a practitioner actually has. These are a small human sample, an expert forecast or a statistical predictor, each at its own cost. All of them face the same transport problem. Expert forecasts of experimental results are informative about averages but miss surprises (DellaVigna and Pope, 2018; Milkman et al., 2021). Flexible predictors trained on behavior transfer less well than theory once the new setting differs from the training data (Andrews et al., 2025). Placing simulations on the same scoreboard would show whether the

Table 2: Three routes to unvalidated settings

Route	Where the invariance assumption sits	Data required	Characteristic failure	Best existing evidence
Theory	The named mechanism governs behavior in the setting of interest	A theory that applies with corresponding parameters	Theory incomplete, missing relevant forces. Parameters shift with context	Manning and Horton (2025): play in new games from a prespecified family
Training	The setting of interest belongs to the family the added data cover	Behavioral training data from nearby settings of the same kind	The new setting lies outside the family	Binz et al. (2025), Suh et al. (2025): new participants, tasks, and questions within the covered domain
Correction	The simulation gets the change from the validated setting to the setting of interest right	One human sample from a nearby setting	Behavior changes in ways the simulation does not track	Zhang et al. (2026): transfer across regions, not across weeks

Notes. A human sample from the setting of interest itself requires no transport assumption; it is simply design-based variance reduction. It is the estimator of Section 2.4.3.

reduced-form pattern is specific to LLMs or common to any predictor without a theory of context. At the individual level, Gao et al. (2026) compare an off-the-shelf LLM to machine learning trained on human data (see Mullainathan and Spiess, 2017). They measure how much human data the algorithm needs to match the LLM. The LLM turns out to be useful for some outcomes, e.g. homeownership, and not for others, e.g. smoking. For aggregate quantities, text-based indicators have long been used to track expectations and economic activity (Baker et al., 2016; Shapiro et al., 2022; O'Connor et al., 2010; Ashwin et al., 2024). If a simpler or cheaper predictor performs as well as synthetic personas, the personas have no demonstrated benefit.

4.1 What Better Models Can Change

We distinguish between what better models imply for improving predictions and establishing under which circumstances they can be trusted. In terms of accuracy where it is checked, a larger training record could improve simulations in situations with better coverage. However, the issue that the space of possible situations is infinite-dimensional and thus impossible to cover even with a larger record remains. Moreover, it is not clear why a significant growth in the behavioral record of real human behavior should be expected, relative to what current LLMs are trained on. The effects of improvements of general capabilities in AI on the quality

of behavioral simulation are contested (e.g., Binz et al., 2026), not least because LLMs are not targeting behavioral simulation, as discussed in Section 3. As discussed throughout this paper, however, higher accuracy in checked situations does not in and of itself provide the license to expect similar accuracy in new settings. The requirements for establishing the validity of simulations in unchecked settings, i.e. recovered structure or demonstrated stability, remain unchanged.

In principle, one might speculate that novel variants of AI models, such as world models, do not form just a reduced form account of the behavioral training record in the way we argue current models do and instead recover deeper structure. In that scenario, off-the-shelf behavioral simulation might turn out to be altogether more promising, but we currently see no evidence for it.

4.2 Route One: Add Behavioral Theory

The vast theoretical literature on decision-making offers the substrate for endowing agents with specific structural properties that AI simulations may otherwise lack. The idea here is to add concrete instructions to the prompt that initiates a persona in a way that controls a specific aspect of their behavior in line with some theory. The evidence using this type of approach is still scant. The agents in Manning and Horton (2025) are told that they are human with all the corresponding cognitive biases and heuristics and given an additional statement that serves to control their level of strategic reasoning, such as being a level-0 thinker, or varying between a level-1 and a level-2 thinker. Theory suggests a specific type of behavioral heterogeneity, level-k thinking in this case, and the structure supplied to the agent tells it which level it is supposed to represent.

We see three main limitations to the improvements that can be achieved with this approach. First, and most importantly, every behavioral theory is incomplete in that it does not provide the full mapping from person and situation to behavior. It names a few specific forces but leaves the rest out. Fudenberg et al. (2022) quantify this as the *completeness* of a theory, i.e. the share of predictable variation it captures. For standard theories of risk and time preferences, a substantial share remains unexplained. To provide an example, a simulation might specify the depth of strategic reasoning but leaves out any number of other well-documented mechanisms that plausibly apply in a context, such as loss aversion or social comparisons.

Second, decision-making research robustly shows that many different kinds of context features strongly affect behavior, often dominating the effect of person-specific traits and leading to apparent instability of behavioral patterns. This means that characterizing an agent with a specific behavioral trait such as loss aversion or the depth of strategic reasoning will fail to

capture the systematic variation in real behavior that comes from how these different characteristics play out differently in different contexts. In fact, theorizing about the role of context effects tends to be less well developed, creating a real gap in what adding behavioral theory can achieve. The same instability undermines whatever the prompt pins down: a type distribution or a parameter such as the depth of strategic reasoning must be assumed constant across settings, yet such parameters shift with framing, reference points, and salience, so the heterogeneity the prompt imposes may be wrong in the setting being simulated.

Third, the theory route mentions a specific force in a prompt but does not and cannot specify how precisely this force plays out in the setting where the simulation is run. What a level-2 thinker does in a game and how exactly it reasons about payoffs and others is left to the reduced form. Put differently, adding theory does not replace the reduced form but constrains it to a degree. We believe that this is not a limitation specific to Manning and Horton (2025) but common to any theory one could prompt with.

Summarizing, the theory route only improves predictions to the extent that (i) the named force is the one that matters here and (ii) the agent's filling-in of how the theory affects behavior is correct.

4.3 Route Two: Train on Behavior from Nearby Settings

The idea behind the training route is that because every situation is characterized by infinitely many features and the original training corpus cannot possibly cover an infinite-dimensional space, it might help to add available behavioral data that is "close" to the setting of interest. Whatever form these data take, they enter before the simulation runs: they can be incorporated into the model weights through fine-tuning (Binz et al., 2025; Suh et al., 2025), retrieved at run time, or included in the prompt as examples of what people did in a related situation. Either way the additional data is supposed to improve the model's mapping from the input to the simulated behavior in the context of interest. The logic here is that because behavior depends on combinations of very many features, the original training record cannot be expected to cover any given setting of interest, and so more targeted behavioral training adds points to a space that cannot be filled. The evidence for the training route suggests that it improves predictions for new participants, tasks, and questions within the narrow domain the additional data cover (Binz et al., 2025; Suh et al., 2025).

We make two remarks on this approach. First, the training route does not add structure but it adds records where the record was previously thin. What models "learn" this way is still contested (Bowers et al., 2026; Schröder et al., 2025).[9] Second, this also implies that only data

[9]The working assumption for this approach is *exchangeability* as discussed in Section 2.1.3. The setting of

from nearby situations that share most features of the target context will help.

4.4 Route Three: Measure and Correct the Error at the Point of Use

Unlike the first two routes, the correction route does not seek to improve the simulator itself by improving the accuracy of the mapping from inputs to behavior. Instead, it takes the simulation output as it is, measures how wrong it is on a human sample, and then accounts for the error to correct the simulation in the setting of interest. Two cases need to be kept apart here. In some situations, the user has a human sample that comes from the setting of interest itself. Note that in this case, there is no transport problem to begin with. The sample identifies the error directly and the simulation only adds precision beyond what the sample itself establishes. This is the design-based estimator of Section 2.4.3, and it is the best-supported use of simulation in the evidence we review—but, of course, it describes the practically perhaps less relevant situation where some data in the precise setting of interest is available. Usually it isn't, or data is only available for some of the counterfactual states of interest. Survey statisticians refer to this case as the regression estimator from two-phase sampling. In the second case, the human sample comes from a nearby setting. This is the correction route proper, and the rest of this section is about this latter case.

There is a fundamental difference between what the human data is used for here versus in the case of the training route. In the case of training, one has to assume that the training data effectively cover the setting of interest in that the relevant combination of features somehow appears in the data. When we use human data for correction, all that matters is that the structure of the error is stable between the two settings. The error does not literally have to be a constant amount in both but is allowed to differ in ways that are predictable from observables. Put simply, the simulation does not have to be right, but it needs to be wrong in a stable and predictable way. This means it has to get the change between the two settings right.

When does this assumption hold? The correction is valid whenever the simulation gets the change across nearby settings right, irrespective of the levels it predicts. The experiment reported in Zhang et al. (2026) illustrated both cases. Across regions, the correction learned on a tenth of the regions worked for the remaining regions. Across time, people's responses changed substantially while the simulated responses did not, so the correction learned on the first week did not help for the prediction two months later. How much human data the correction needs depends on the structure of the error. Take the case where respondent-based effects are inflated by a roughly common factor, as in Ashokkumar et al. (2026). In that case, a small human sample pinning down one scale parameter corrects many effects at once. If the error instead varies

interest must be of the family the additional training data come from, which is an assumption. One cannot validate what exactly the added training does to the simulation output.

from effect to effect, each effect needs its own sample. Any application thus has to state clearly which features are assumed to leave the error structure stable. It is an open question whether the error systematically moves more with some than other classes of features, e.g. more with changes in setting than changes in people.

Note that unlike in the case of the training route, the check and assumption that's underlying the correction are made by the user, at the point of use, and separately for each application. This makes the correction route highly transparent in principle.

4.5 Lessons for Practitioners

This review argues that for any practitioner who considers using AI-based behavioral simulations, the question is not whether a given simulation is "good" in general but to what extent it can be trusted for the question at hand. The answer to this pivotal question hinges on where human data is available or can be collected, and the three routes provide a decision rule.

Sometimes some human data is available in precisely the situation of interest. Even with a small sample, such data can be used to measure and correct for the simulation's error. In this case no transport assumption is needed. The simulation augments the human sample by adding precision to what the human data alone would allow the user to learn (Section 2.4.3).

In other cases, the best available human data are from a nearby setting, i.e. one where the population or the decision might be slightly different. This calls for the correction route, where the nearby data is used to measure the local simulation error which can then be accounted for in the setting of interest. Here one needs to either assume that the error is locally stable or predictable from certain observable features.

Sometimes no new human data on the target situation or a nearby one are available and the best available data is from earlier studies or logs of past behavior. Such data can be used to locally train and improve simulations for situations within the same family of settings. A new product, rule or a later period is, by definition, outside that family and accuracy is thus not established through the additional training.

When none of these options is available, the claim should be restricted to what the decision at hand needs. The three decisions from the introduction need different things. Choosing between two messages requires a ranking, which simulations tend to get right when the effects differ substantially. Simulations might thus be worthwhile to use to discard weak ideas and to prune treatment arms. Forecasting take-up for a budget relies on a level forecast. Respondent-based simulations tend to get levels wrong, and direct forecasts get them right only for familiar interventions. Finally, deciding which experiment to run next requires knowing where the simulation is most uncertain, because that is where a human sample is most informative. Simulations

should not be used to select between similar options or to calibrate decisions that depend on the magnitude of an effect or on the segmentation of a population.

Note that the structure route is rarely a feasible option in practice. It requires an established theory of behavior for the situation of interest and the expertise to encode and calibrate it. Existing evidence for this route comes from narrow families of abstract games. Unlike the stylized games studied in many experiments that are stripped of real-life context, most practical simulation settings are full of idiosyncratic context that no single theory or set of theories should be expected, ex ante, to capture comprehensively.

There are a few general rules that follow from the findings of this review. First, the error bars for a synthetic sample are a feature of the simulation approach and should not be mistaken for their equivalent in a human sample. Put differently, simulating more respondents can never substitute for a human sample. Second, more sophisticated simulations with richer personas etc. have not been demonstrated yet to earn their cost in predictive accuracy. Real human responses remain the gold standard for validating and correcting AI-based simulations. Third, any validation only applies to the model and procedure that was actually used, meaning that any model update requires new validation. Finally, it helps to know what an unchecked simulation is comparable to. Our reading of the evidence is that it resembles a well-read expert forecast. Whenever one suspects that an actual A/B test could yield a surprising finding, a simulation is likely to miss it, just as an expert would.

4.6 The Open Question: How Far Does a Human Sample Travel?

Many attractive applications concern decisions for which human outcomes are not yet available. Can observations from another population, an earlier period, or a related decision support useful predictions? The question is how far evidence about accuracy in one setting carries over to another, across populations and contexts. It takes a different form for the two routes that rely on human data. The correction route needs to know how far the structure of a simulation's error stays stable; the training route needs to know which settings belong to the family over which accuracy was established. The evidence reviewed here offers little systematic guidance on either.

Measuring distance. Either question presupposes a measure of distance: in behavior for the training route, and in the simulation's error for the correction route, which can stay stable even when behavior changes, as long as the model tracks the change. A convenient proxy is the similarity between descriptions of two settings, measured for example with text embeddings. It can mislead in both directions: a vignette that changes one incentive is close in text and may

be far in behavior, while the same choice described in other words is far in text and may be close in behavior. Moreover, embeddings come from models trained on the same record as the simulator, so they cannot flag a change the record does not reflect, which is where a reduced-form simulation is most likely to fail. For changes in population, a better measure is sometimes available: differences in observed characteristics can be measured directly and the error allowed to vary with them (Pearl and Bareinboim, 2014; Dell and Rambachan, 2026). For changes in context, no comparable list of characteristics exists.

Designs. Three practices are feasible without such a theory. First, studies can report which coordinates in Table 1 differ between validation and target. Second, they can fix a model and correction, specify changes in population, timing, or incentives before observing the target outcomes, and hold out settings by those features rather than at random—a country, all incentivized tasks, a later period. Prediction error, the error remaining after correction, and interval coverage can then be measured along each named dimension. Andrews et al. (2025) provide a template: apply the procedure to every pair of training and target settings among those validated, and report quantiles of the resulting errors, which give the forecast intervals of Section 2.1.3. Defining settings by a single feature, as they do when each setting is one lottery, gives such a test along that feature. Third, where transfer remains uncertain, they can report how conclusions change under plausible departures from the transport assumption, for example by bounding how much the error may change (Section 2.1.3; Chen et al., 2026). A conventional confidence interval cannot account for an unspecified failure of that assumption. Designs that vary many features of a setting at once, rather than one at a time, would show where a correction continues to hold and begin to supply the missing theory of context (Almaatouq et al., 2024). The same data would also test our interpretation: under a reduced-form account, accuracy should track how familiar a setting's description is to the model, not how much behavior actually changed.

5 Conclusion

In this review, we have put together evidence to assess how good LLMs are at predicting human behavior. The evidence shows that accuracy falls with specificity and, for the changes tested so far, travels within familiar kinds but not beyond them. The key insight of this review is that for those applications that matter the most — novel unfamiliar problems where no human data can be easily collected — the key question concerns whether accuracy carries over across settings.

Three routes are most promising to increase accuracy of predictions in novel prediction problems. Adding behavioral theory works in as far as the theory is right, comprehensively

captures relevant factors in the setting of interest, and the agent correctly translates a named theoretical mechanism into actual decisions. Supplying additional behavioral data from settings that share many features with the situation of interest carries as far as the new setting can reasonably be considered to be part of the family covered by the additional data. Measuring the simulation error against a human sample and correcting for it in a new situation will be valid to the extent that the error structure is stable across the two situations. How far that is, across populations and across contexts, remains an open question.

References

Ackley, David H., Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. "A Learning Algorithm for Boltzmann Machines." *Cognitive Science* 9 (1): 147–169.

Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies." In *Proceedings of the 40th International Conference on Machine Learning*.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117–1165. 10.1093/qje/qjv015.

Almaatouq, Abdullah, Thomas L. Griffiths, Jordan W. Suchow, Mark E. Whiting, James Evans, and Duncan J. Watts. 2024. "Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences." *Behavioral and Brain Sciences* 47 e33. 10.1017/S0140525X22002874.

Andrews, Isaiah, Drew Fudenberg, Lihua Lei, Annie Liang, and Chaofeng Wu. 2025. "The Transfer Performance of Economic Models." Working Paper2202.04796, arXiv, https://arxiv.org/abs/2202.04796.

Andrews, Isaiah, and Suproteem Sarkar. 2026. "Dutch Books for Language Models." arXiv:2609.02797.

Anesti, Nikoleta, Edward Hill, and Andreas Joseph. 2026. "Inflation Attitudes of Large Language Models." Staff Working Paper 1190, Bank of England.

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023a. "Prediction-Powered Inference." *Science* 382 669–674.

Angelopoulos, Anastasios N., John C. Duchi, and Tijana Zrnic. 2023b. "PPI++: Efficient Prediction-Powered Inference." *arXiv preprint arXiv:2311.01453*.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–351. 10.1017/pan.2023.2.

Ashokkumar, Ashwini, Luke Hewitt, Isaias Ghezae, and Robb Willer. 2026. "Large Language Models Can Predict the Results of Social Science Experiments." *Nature* 656 (8126): 115–122. 10.1038/s41586-026-10742-x.

Ashwin, Julian, Eleni Kalamara, and Lorena Saiz. 2024. "Nowcasting Euro Area GDP with News Sentiment: A Tale of Two Crises." *Journal of Applied Econometrics* 39 (5): 887–905. 10.1002/jae.3057.

Bai, Yuntao et al. 2022a. "Constitutional AI: Harmlessness from AI Feedback." *arXiv preprint arXiv:2212.08073*.

Bai, Yuntao et al. 2022b. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." *arXiv preprint arXiv:2204.05862*.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring Economic Policy Uncertainty." *Quarterly Journal of Economics* 131 (4): 1593–1636.

Baumann, Joachim, Paul Röttger et al. 2025. "Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation." *arXiv preprint arXiv:2509.08825*.

Binz, Marcel et al. 2025. "A Foundation Model to Predict and Capture Human Cognition." *Nature* 644 (8078): 1002–1009. 10.1038/s41586-025-09215-4.

Binz, Marcel, Elif Akata, Eric Schulz et al. 2026. "Post-Training Makes Large Language Models Less Human-Like." *arXiv preprint arXiv:2605.07632*.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis*.

Bombassei De Bona, Francesco, Gabriele Dominici, Tim Miller, Marc Langheinrich, and Martin Gjoreski. 2024. "Evaluating Explanations Through LLMs: Beyond Traditional User Studies." *arXiv preprint arXiv:2410.17781*.

Bort, Julie. 2026. "Synthetic-User Startup Simile Raises $200M at $2B Valuation 5 Months after $100M Series A." *TechCrunch*, July 30, https://techcrunch.com/2026/07/30/synthetic-user-startup-simile-raises-200m-at-2b-valuation-5-months-after-100m-series-a/.

Bowers, Jeffrey S., Guillermo Puebla, Sushrut Thorat, Konstantinos Tsetsos, and Casimir Ludwig. 2026. "On the Misuse of LLMs as Models of Mind: A Case Study of Centaur." *PsyArXiv preprint*.

Broska, David, Michael Howes, and Austin van Loon. 2025. "The Mixed Subjects Design: Combining Experimental with Correlational Methods to Predict Human Behavior with Large Language Models." *Sociological Methods and Research*.

Brown, Tom B. et al. 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, Volume 33.

Bursztyn, Leonardo, Ingar K. Haaland, Nicolas Röver, and Christopher Roth. 2025. "The Social Desirability Atlas." NBER Working Paper33920, National Bureau of Economic Research. 10.3386/w33920.

Bybee, Leland. 2023. "Surveying Generative AI's Economic Expectations." *arXiv preprint arXiv:2305.02823*.

Cash, Trent, Daniel M. Oppenheimer, Sara Christie, and Mira Devgan. 2025. "Quantifying Uncert-AI-nty: Testing the Accuracy of LLMs' Confidence Judgments." *Memory & Cognition*.

Casper, Stephen et al. 2023. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback." *Transactions on Machine Learning Research*.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2023. "Econographics." *Journal of Political Economy Microeconomics* 1 (1): 115–161.

Chen, Fadong, Qijun Wang, Lei Wang, and Sudeep Bhatia. 2026. "Simulated Responses from Large Language Models Are Weak Proxies for Human Economic Preferences across 74 Countries." *Nature Human Behaviour*. 10.1038/s41562-026-02611-4, Published online October 12, 2026.

Chen, Xiaohong, Ashesh Rambachan, and Elie Tamer. 2026. "Partial Identification from LLM Prompts." *Working paper*.

Christiano, Paul F. et al. 2017. "Deep Reinforcement Learning from Human Preferences." In *Advances in Neural Information Processing Systems*, Volume 30.

Cruz, André F., Moritz Hardt, and Celestine Mendler-Dünner. 2024. "Evaluating Language Models as Risk Scores." *arXiv preprint arXiv:2407.14614*.

Cui, Ganqu et al. 2025. "The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models." *arXiv preprint arXiv:2505.22617*.

Cui, Ziyan, Ning Li, and Huaikang Zhou. 2025. "A Large-Scale Replication of Scenario-Based Experiments in Psychology and Management Using Large Language Models." *Nature Computational Science* 5 (8): 627–634. 10.1038/s43588-025-00840-7.

Dean, Mark, and Pietro Ortoleva. 2019. "The Empirical Relationship between Nonstandard Economic Behaviors." *Proceedings of the National Academy of Sciences* 116 (33): 16262–16267.

Deaton, Angus, and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210 2–21.

DeepSeek-AI. 2025. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." *arXiv preprint arXiv:2501.12948*.

Dell, Melissa, and Ashesh Rambachan. 2026. "The Measurement Revolution? Credible Measurement and Inference in the Age of AI." *arXiv preprint arXiv:2608.23524*, NBER Methods Lecture.

DellaVigna, Stefano, and Elizabeth Linos. 2022. "RCTs to Scale: Comprehensive Evidence From Two Nudge Units." *Econometrica* 90 (1): 81–116.

DellaVigna, Stefano, and Devin Pope. 2018. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126 (6): 2410–2456.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. "Predict Science to Improve Science." *Science* 366 (6464): 428–429.

Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. "Questioning the Survey Responses of Large Language Models." In *Advances in Neural Information Processing Systems*.

Durmus, Esin, Karina Nguyen, Thomas I. Liao et al. 2023. "Towards Measuring the Representation of Subjective Global Opinions in Language Models." *arXiv preprint arXiv:2306.16388*.

Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2024. "Using Large Language Model Annotations for Valid Downstream Statistical Inference." In *Advances in Neural Information Processing Systems*.

Eggert, Gus, Kevin Huo, Mike Biven, and Justin Waugh. 2023. "TabLib: A Dataset of 627M Tables with Context."

Einav, Liran, Amy Finkelstein, Iuliana Pascu, and Mark R. Cullen. 2012. "How General Are Risk Preferences? Choices under Uncertainty in Different Domains." *American Economic Review* 102 (6): 2606–2638.

Elazar, Yanai, Akshita Bhagia, Ian Magnusson et al. 2024. "What's In My Big Data?" In *International Conference on Learning Representations*.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive Uncertainty." *Quarterly Journal of Economics* 138 (4): 2021–2067.

Enke, Benjamin, Thomas Graeber, Ryan Oprea, and Jeffrey Yang. 2024. "Behavioral Attenuation." NBER Working Paper32973, National Bureau of Economic Research, Forthcoming, *Journal of Political Economy*.

Enke, Benjamin, and Cassidy Shubatt. 2023. "Quantifying Lottery Choice Complexity." NBER Working Paper31677, National Bureau of Economic Research.

Fan, Angela, Mike Lewis, and Yann Dauphin. 2018. "Hierarchical Neural Story Generation." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 889–898.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy* 130 (4): 956–990.

Gao, Wayne, Sukjin Han, and Annie Liang. 2026. "How Well Do LLMs Predict Human Behavior? A Measure of their Pretrained Knowledge." *arXiv preprint arXiv:2601.12343*.

Gillin, Paul. 2026. "Qualtrics Adds AI-Powered Synthetic Data and Research Tools to Speed Customer Insights." *SiliconANGLE*, March 18, `https://siliconangle.com/2026/03/18/qualtrics-adds-ai-powered-synthetic-data-research-tools-speed-customer-insights/`.

Gligorić, Kristina, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. 2025. "Can Unconfident LLM Annotations Be Used for Confident Conclusions?" In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

Gneiting, Tilmann, and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–378.

Graham, John R., Campbell R. Harvey, and Manish Jha. 2026. "CFOs Meet LLMs." *arXiv preprint arXiv:2606.13812*.

Grattafiori, Aaron, Abhimanyu Dubey et al. 2024. "The Llama 3 Herd of Models." *arXiv preprint arXiv:2407.21783*.

Grief-Albert, Bo, Jiao, and Anderson. 2026. "Emulate or Estimate? The Divergent Strengths of Base and Post-Trained Language Models for Opinion Simulation." *arXiv preprint arXiv:2608.03044*.

Gui, George, and Olivier Toubia. 2025. "The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective." *arXiv preprint arXiv:2312.15524*.

Haavelmo, Trygve. 1944. "The Probability Approach in Econometrics." *Econometrica* 12 iii–115, Supplement.

Hagendorff, Thilo, Sarah Fabi, and Michal Kosinski. 2023. "Human-Like Intuitive Behavior and Reasoning Biases Emerged in Large Language Models but Disappeared in ChatGPT." *Nature Computational Science*.

He, Horace, and Thinking Machines Lab. 2025. "Defeating Nondeterminism in LLM Inference." Thinking Machines Lab, Connectionism, September 2025.

von der Heyde, Leah, Anna-Carolina Haensch, and Alexander Wenz. 2024. "Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion." *arXiv preprint arXiv:2407.08563*.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. "The Curious Case of Neural Text Degeneration." In *International Conference on Learning Representations*.

Horton, John J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?", NBER Working Paper31122, National Bureau of Economic Research.

Huang, Chengpiao, Yuhang Wu, and Kaizheng Wang. 2025. "How Many Human Survey Respondents Is a Large Language Model Worth? An Uncertainty Quantification Perspective." *arXiv preprint arXiv:2502.17773*.

Huang, Yixu, Yunlu Yin, Jiayu Lin et al. 2026. "Uncovering Salience-Driven Dynamics in Consumer Confidence with Generative Social Simulation." *arXiv preprint arXiv:2606.30395*.

Huff, Markus, and Elanur Ulakci. 2025. "Judgments of Learning Distinguish Humans from Large Language Models in Predicting Memory." *Scientific Reports* 15.

Hullman, Jessica, David Broska, Huaman Sun, and Aaron Shaw. 2026. "This Human Study Did Not Involve Human Subjects: Validating LLM Simulations as Behavioral Evidence." *arXiv preprint arXiv:2602.15785*.

Hurwicz, Leonid. 1962. "On the Structural Form of Interdependent Systems." In *Logic, Methodology and Philosophy of Science*, edited by Ernest Nagel, Patrick Suppes, and Alfred Tarski, 232–239, Stanford: Stanford University Press.

Iadisernia, Giulia, and Carolina Camassa. 2025. "Prompting for Policy: Forecasting Macroeconomic Scenarios with Synthetic LLM Personas." *arXiv preprint arXiv:2511.02458*.

Jang, Chaemin, Dongman Lee, and Jihee Kim. 2026. "Instruction-Tuned Language Models Cannot Sample from Distributions They Can Describe." *arXiv preprint arXiv:2607.25292*.

Kadavath, Saurav et al. 2022. "Language Models (Mostly) Know What They Know." *arXiv preprint arXiv:2207.05221*.

Kandpal, Nikhil, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. "Large Language Models Struggle to Learn Long-Tail Knowledge." In *Proceedings of the 40th International Conference on Machine Learning*, 15696–15707.

Kim, Elliot, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. "Correlated Errors in Large Language Models." In *Proceedings of the International Conference on Machine Learning*, arXiv:2506.07962.

Kirk, Robert et al. 2024. "Understanding the Effects of RLHF on LLM Generalisation and Diversity." In *International Conference on Learning Representations*.

Krsteski, Stefan, Daniele Russo, Serina Chang, Robert West, and Kristina Gligorić. 2025. "Valid Survey Simulations with Limited Human Data: The Roles of Prompting, Fine-Tuning, and Rectification." *arXiv preprint arXiv:2510.11408*.

Lambert, Nathan et al. 2024. "Tülu 3: Pushing Frontiers in Open Language Model Post-Training." *arXiv preprint arXiv:2411.15124*.

LaPiere, Richard T. 1934. "Attitudes vs. Actions." *Social Forces* 13 (2): 230–237.

Lee, Katherine et al. 2022. "Deduplicating Training Data Makes Language Models Better." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Leng, Jixuan, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2024. "Taming Overconfidence in LLMs: Reward Calibration in RLHF." *arXiv preprint arXiv:2410.09724*.

Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, Volume 1. of Carnegie-Rochester Conference Series on Public Policy 19–46.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2026. "Large Language Models: An Applied Econometric Framework." *Annual Review of Economics* 18 (1): 283–316. 10.1146/annurev-economics-120925-105620.

Manning, Benjamin S., and John J. Horton. 2025. "General Social Agents." *arXiv preprint arXiv:2508.17407*.

Manning, Benjamin S., Kehang Zhu, and John J. Horton. 2024. "Automated Social Science: Language Models as Scientist and Subjects." NBER Working Paper32381, National Bureau of Economic Research.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11 (1): 57–91.

Mei, Ning, Dobromir Rahnev, and David Soto. 2023. "Using Serial Dependence to Predict Confidence Across Observers and Cognitive Domains." *Psychonomic Bulletin & Review*.

Meister, Nicole, Carlos Guestrin, and Tatsunori Hashimoto. 2025. "Benchmarking Distributional Alignment of Large Language Models." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

Milkman, Katherine L. et al. 2021. "Megastudies Improve the Impact of Applied Behavioural Science." *Nature* 600 (7889): 478–483. 10.1038/s41586-021-04128-4.

Mischel, Walter. 1968. *Personality and Assessment*. New York: Wiley.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Murthy, Sonia K., Tomer Ullman, and Jennifer Hu. 2025. "One Fish, Two Fish, but Not the Whole Sea: Alignment Reduces Language Models' Conceptual Diversity." In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

Ni, Jingwei, Yu Fan, Vilém Zouhar et al. 2026. "Can Reasoning Help Large Language Models Capture Human Annotator Disagreement?" In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 36–54.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 4. (1): 122–129. 10.1609/icwsm.v4i1.14031.

OpenAI. 2023. "GPT-4 Technical Report." *arXiv preprint arXiv:2303.08774*.

Ouyang, Long et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." In *Advances in Neural Information Processing Systems*, Volume 35.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Park, Joon Sung, Carolyn Q. Zou, Jonne Kamphorst et al. 2026. "LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals." *arXiv preprint arXiv:2411.10109*, June 2026 manuscript version.

Park, Peter S., Philipp Schoenegger, and Chongyang Zhu. 2024. "Diminished Diversity-of-Thought in a Standard Large Language Model." *Behavior Research Methods*.

Pearl, Judea, and Elias Bareinboim. 2014. "External Validity: From Do-Calculus to Transportability Across Populations." *Statistical Science* 29 (4): 579–595.

Penedo, Guilherme, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov et al. 2024. "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale." In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.

Pezeshkpour, Pouya, and Estevam Hruschka. 2024. "Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions." In *Findings of the Association for Computational Linguistics: NAACL*.

Project Apertus. 2025. "Apertus: Democratizing Open and Compliant LLMs for Global Language Environments." *arXiv preprint arXiv:2509.14233*.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model." In *Advances in Neural Information Processing Systems*, Volume 36.

Ross, Lee, and Richard E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. New York: McGraw-Hill.

Roytburg, Eva. 2026. "A 21-Year-Old Cofounder's Sales Pitch to Clients Begs Them to Question the Company's Results: 'Do Not Trust Us. Do Not Trust Our Model'." *Fortune*, June 17, https://fortune.com/2026/06/17/aaru-cofounder-ned-koh-ai-startup-sales-pitch-do-not-trust-us/.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. "Whose Opinions Do Language Models Reflect?" In *Proceedings of the 40th International Conference on Machine Learning*, Volume 202. 29971–30004.

Schröder, Sarah, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. "Large Language Models Do Not Simulate Human Psychology." *arXiv preprint arXiv:2508.06950*.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting." In *International Conference on Learning Representations*.

Shapiro, Adam Hale, Moritz Sudhof, and Daniel J. Wilson. 2022. "Measuring News Sentiment." *Journal of Econometrics* 228 (2): 221–243. 10.1016/j.jeconom.2020.07.053.

Sheeran, Paschal, and Thomas L. Webb. 2016. "The Intention–Behavior Gap." *Social and Personality Psychology Compass* 10 (9): 503–518. 10.1111/spc3.12265.

Stanford CRFM. 2024. "Foundation Model Transparency Index, May 2024: OpenAI GPT-4 Report." https://crfm.stanford.edu/fmti/May-2024/company-reports/OpenAI_GPT-4.html.

Stango, Victor, and Jonathan Zinman. 2023. "We Are All Behavioural, More, or Less: A Taxonomy of Consumer Decision-Making." *Review of Economic Studies* 90 (3): 1470–1498.

Suh, Joseph, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. "Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 21147–21170. 10.18653/v1/2025.acl-long.1028.

Sun, Hongshen, and Juanjuan Zhang. 2025. "From Model Choice to Model Belief: Establishing a New Measure for LLM-Based Research." *SSRN Working Paper No. 5399546*.

Team OLMo. 2025. "2 OLMo 2 Furious." *arXiv preprint arXiv:2501.00656*.

Tian, Katherine, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. "Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Tjuatja, Lindia, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. "Do LLMs Exhibit Human-Like Response Biases? A Case Study in Survey Design." *Transactions of the Association for Computational Linguistics*.

Toubia, Olivier, George Z. Gui, Tianyi Peng, Daniel Merlau, Ang Li, and Haozhe Chen. 2025. "Twin-2K-500: A Data Set for Building Digital Twins of Over 2,000 People Based on Their Answers to Over 500 Questions." *Marketing Science*.

Vafa, Keyon, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. 2025. "What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models." In *Proceedings of the International Conference on Machine Learning*, arXiv:2507.06952.

Vafa, Keyon, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. 2024. "Evaluating the World Model Implicit in a Generative Model." In *Advances in Neural Information Processing Systems*, arXiv:2406.03689.

Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18 (6): 3045–3089. 10.1093/jeea/jvaa019.

Vivalt, Eva. 2026. "Can Artificial Intelligence Forecast Treatment Effects in Economics?", Working paper, University of Toronto, October 28, 2026.

Wan, Alexander, Kevin Klyman, Sayash Kapoor, Nestor Maslej, Shayne Longpre, Betty Xiong, Percy Liang, and Rishi Bommasani. 2025. "The 2025 Foundation Model Transparency Index." *arXiv preprint arXiv:2512.10169*.

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2025. "Large Language Models That Replace Human Participants Can Misportray and Flatten Identity Groups." *Nature Machine Intelligence*.

Wang, Qian, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. 2025. "Assessing Judging Bias in Large Reasoning Models: An Empirical Study." *arXiv preprint arXiv:2504.09946*.

Wang, Xiao. 2026. "More Thinking, More Bias: Length-Driven Position Bias in Reasoning Models." *arXiv preprint arXiv:2605.06672*.

Wang, Xinpeng, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "'My Answer is C': First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models." In *Findings of the Association for Computational Linguistics: ACL 2024*, 7407–7416.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." In *International Conference on Learning Representations*.

Wei, Jason et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In *Advances in Neural Information Processing Systems*, Volume 35.

Wei, Jason. 2025. "Asymmetry of Verification and Verifier's Law." Blog post, July 15, https://www.jasonwei.net/blog/asymmetry-of-verification-and-verifiers-law.

Wu, Jing Cynthia, Jin Xi, and Shihan Xie. 2025. "LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification." NBER Working Paper34308, National Bureau of Economic Research. 10.3386/w34308.

Xiong, Miao, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs." In *International Conference on Learning Representations*.

Zhang, Bohan, Jiaxuan Li, Ali Hortaçsu, Xiaoyang Ye, Victor Chernozhukov, Anqi Ni, and Edward W. Huang. 2026. "Agentic Economic Modeling." *arXiv preprint arXiv:2510.25743*.

Zhang, Jiayi, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. 2025. "Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity." *arXiv preprint arXiv:2510.01171*.

Zheng, Chujie, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. "Large Language Models Are Not Robust Multiple Choice Selectors." In *International Conference on Learning Representations*.

Zhou, Kaitlyn, Dan Jurafsky, and Tatsunori Hashimoto. 2023. "Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5506–5524.

Zhu, Jian-Qiao, Joshua C. Peterson, Benjamin Enke, and Thomas L. Griffiths. 2025. "Capturing the Complexity of Human Strategic Decision-Making with Machine Learning." *Nature Human Behaviour*.

Ziegler, Daniel M. et al. 2019. "Fine-Tuning Language Models from Human Preferences." *arXiv preprint arXiv:1909.08593*.