LLMs Do Not Emulate Populations

Brice Green[a] and Greg Leo[b]

[a]Massachusetts Institute of Technology
[b]Loyola Marymount University

August 26, 2026

Abstract

Increasingly, researchers are using large language models (LLMs) to generate synthetic survey responses and behavioral data as a substitute for human samples. The justification for this approach is that appropriate prompt conditioning can recover latent patterns that LLMs learn about humans from training data. Existing evaluations of this technique typically compare LLM output to direct measurements of human behavior. However, these evaluations do not reflect the goal of this approach, which is often to make inferences about unobserved scenarios and populations. By contrast, we devise tests that can be run on LLM output alone and are based solely on structural properties implied by the assumptions underlying the technique. We show that LLMs produce data that is incoherent with respect to the statistical structure of real populations, which requires that the mean of any population lies within the range of the means of its constituent subpopulations. LLMs violate this condition at rates similar to purely random predictions. Further, we show that model outputs are subjective responses rather than objective predictions about a common (even biased) human benchmark population. In our tests the chosen model explains *far* more of the variation in output than can be explained by the choice of emulated demographic group. We conclude that LLM output is unreliable for inference about human populations.

1 Introduction

Large language models (LLMs) are trained on an enormous quantity of human data, which includes a substantial amount of information about how people respond to survey questions or play games, which may imbue LLMs with a generalized model of human behavior. With this justification, a growing research program in the social sciences uses LLMs to generate synthetic survey responses and behavioral data.

This technique, often called silicon sampling, asks LLMs to emulate human respondents with "personas," or prompts that include demographic or other relevant identity information. The LLM's responses are used to construct synthetic populations for surveys, behavioral experiments, and other analyses. The data sampled from many LLM personas are then aggregated to make inferences about populations those personas compose.

In this paper, we ask two questions. First, do LLMs coherently emulate populations through this technique? Second, to the extent that LLMs do emulate populations, do they do so because they have learned how humans act?

One approach to evaluating silicon sampling is to compare LLM responses with observed human data. However, agreement with a particular dataset does not imply that the model will generalize to new populations, demographic subpopulations, or hypothetical scenarios. A model may match an observed population while simultaneously implying other statistical relationships that no real population would satisfy.

By contrast, we develop tests that rely solely on output from LLMs. These tests are derived purely from structural properties that underlie the assumptions inherent in silicon sampling: that LLMs can coherently emulate populations, and that the population they emulate is some common objective benchmark (ideally real humans).

Do LLMs emulate populations? If they do, they must respect the statistical structure of populations, which are mixtures of their constituent subpopulations. By the law of iterated expectation, the parent population mean must lie between the subpopulation means. We refer to violations of this as **incoherence**. For example, predicting that the average height of a man is 5 feet 9 inches, the average height of a woman is 5 feet 4 inches, and the average height of an adult is 5 feet 2 inches is incoherent because no weighted average of the man and woman means could produce the adult mean.

Do LLMs emulate how humans act? If they do, variation in responses must be driven by the humans being emulated, not the models themselves. When models fail to do this, we say they are **subjective**. For example, if one model generates responses twice as risk averse as another model, across all demographic groups, then their responses are not simulations from a common model of human behavior.

These criteria follow directly from common assumptions in the literature. For example, a popular approach in the synthetic survey literature is to construct "digital twins" (Park et al., 2024) that represent responses of specific demographic groups, which can then be weighted to infer something about a target population. If LLMs are *incoherent*, the demographic hierarchy (e.g., woman vs. 56-year-old woman) is internally inconsistent and a researcher does not know which level to trust. Researchers also use LLM responses to predict what humans will do in unseen experiments (Ashokkumar et al., 2026). If LLMs are *subjective* then their predictions might not be representative of human behavior in unseen circumstances.

Across multiple behavioral games and opinion survey ques-

tions, we find that LLMs frequently violate these criteria. These failures occur across contemporary and older models, across a wide range of model sizes, across both proprietary and open-weight models, and across 'thinking' and 'non-thinking' models. Our findings suggest that matching isolated survey or experimental results should not be taken as sufficient evidence that LLMs can be used to emulate real human populations.

These failures are frequent and large in magnitude. Around half of all marginal distributions generated by LLMs are incoherent. Further, LLMs are subjective across all games and surveys by a large margin. The model matters substantially *more* than the demographic prompt across all tasks.

2 Literature

To our knowledge, Aher et al. (2023) were the first to suggest that LLMs can substitute for human subjects, and Argyle et al. (2023) were the first to propose the use of LLMs as substitutes for human subpopulations in polling.

In both academic research and industry there are attempts to construct "digital twins" of individual people (Park et al., 2024; Toubia et al., 2025), using these in place of human responses in polling and market research. Several prominent papers posit that LLMs can be used to cheaply explore the hypothesis space. Because measurement is close to free, we can use LLMs to extrapolate to new situations without human subjects experiments (Henning and Camerer, 2026; Anthis et al., 2025; Ashokkumar et al., 2026; Xie et al., 2025). The argument for this is straightforward: large language models are trained on large quantities of human-generated data, and so in their training process learn a representation of human behavior accessible via prompting (Grossmann et al., 2023; Horton et al., 2023). In practice, researchers have found empirical correlations between LLM and human responses in several domains (Dillion et al., 2023; Manning and Horton, 2026).

Unsurprisingly, such a bold claim has generated controversy. There is also a growing literature documenting ways in which LLMs fail at these tasks. Several papers document that LLMs are sensitive to seemingly inconsequential changes in prompts, like the order of the allowed responses (Dominguez-Olmedo et al., 2024; Tjuatja et al., 2024). Others find that LLM output is often correlated with human subject samples, but biased in ways which make it an inadequate substitute (Pataranutaporn et al., 2025; Boelaert et al., 2025; Santurkar et al., 2023). Further, LLM output has substantially lower variance than that of real populations (Bisbee et al., 2024), and prompted response distributions are internally inconsistent when aggregated using real survey weights (Li et al., 2025).

Finally, there is a literature that analyzes the required conditions for inferences from LLM simulations to be valid. Ludwig et al. (2024) show that inference about human subjects using LLM output is equivalent to a non-classical measurement error adjustment problem. There are a number of proposed approaches that attempt to repair these potentially biased inferences and validate LLM experiments; see Hullman et al. (2026) for a thorough discussion.

Existing approaches to test the validity of silicon sampling rely on population weights (Li et al., 2025) or human benchmark data (Xie et al., 2026). We contribute to this literature by deriving a set of tests for the assumptions inherent in this technique that can be performed with LLM output alone, without reference to any real data. These tests show that LLMs are unsuitable substitutes for human subjects.

3 Theory

Populations are inherently hierarchical, a composite of various groups, themselves composed of further nested subpopulations. Throughout this section we consider a (parent) population that can be decomposed into K mutually exclusive subpopulations indexed by i. Let Y_p be a random variable representing responses of individuals in a population for some task, and let Y_i be a random variable representing responses for individuals from subpopulation i for the same task. Each subpopulation i represents a proportion w_i (with $w_i \geq 0$ and $\sum_{i=1}^{K} w_i = 1$) of the total population.[1]

Similarly, let $\hat{Y}_{p,m}$ represent model m's simulated responses to the same task for the same population, and let $\hat{Y}_{i,m}$ be a random variable representing the predictions of model m conditioned on a "persona" representing subpopulation i for the same task.

3.1 Incoherence

Because a parent population is a mixture of its subpopulations, by the law of iterated expectation, the mean of the parent population must be the weighted sum of the means of the subpopulations:

$$\mathbb{E}[Y_p] = \sum_{i=1}^{K} w_i \, \mathbb{E}[Y_i]. \tag{1}$$

Since the weights w_i sum to one, the population mean is a convex combination of the subpopulation means, and so the mean of the parent population must be within the extremes of the subpopulation means:

$$\min\left\{\mathbb{E}[Y_i]\right\} \leq \mathbb{E}[Y_p] \leq \max\left\{\mathbb{E}[Y_i]\right\}. \tag{2}$$

If an LLM, m, can coherently emulate populations, then the predictions it makes directly about a parent population should be consistent with *some* mixture (using weights $\tilde{w}_i$) of its subpopulation predictions. That is $\mathbb{E}[\hat{Y}_{p,m}] = \sum_{i=1}^{K} \tilde{w}_i \mathbb{E}[\hat{Y}_{i,m}]$ for some weights $\tilde{w}_i$.

Expression (2) requires that the mean of the parent population lies between the means of the most extreme subpopulations. We say an LLM is **incoherent** if its predictions violate this:

$$\mathbb{E}\left[\hat{Y}_{p,m}\right] < \min\left\{\mathbb{E}\left[\hat{Y}_{i,m}\right]\right\} \text{ or } \mathbb{E}\left[\hat{Y}_{p,m}\right] > \max\left\{\mathbb{E}\left[\hat{Y}_{i,m}\right]\right\}. \tag{3}$$

[1] We note that any population can be decomposed into many such families of subpopulations. For example, the population of adults can be decomposed into subpopulations based on age, based on sex, based on age/sex combinations, or based on other criteria. These constraints apply to any population and any decomposition.

This structural restriction must hold recursively. Since each subpopulation is itself composed of lower-order subpopulations, the mean for each subpopulation must lie between the extremes of its lower-order subpopulations and so on.

Note that we do not insist that the weights $\tilde{w}_i$ correspond to the true population weights w_i. In this sense, the condition for coherence in (3) is weaker than what would hold in real data. Our condition tests, within each task, whether there exists any set of weights that can justify the relationship between the parent population mean and the subpopulation means. For a real population, there is a fixed set of weights that can justify this relationship across *all* tasks.

3.2 Subjectivity

A common argument for using LLMs to predict human behavior is that they are trained on a large corpus of human-generated data, and have thus learned a common set of patterns that reproduce human responses on average. This view is captured well in Horton et al. (2023):

> The human-generated text path provides LLMs with true aspects of the social world "in action:" how people communicate, make decisions, interact with each other, and perceive the world. We might be skeptical that a language model could, say, represent physics we have yet to theorize, but it surely has internalized relevant knowledge about the social world that has never been written down in academic work.

If LLMs learn to emulate populations by reproducing human behavior observed in the training data, then on average predictions should not depend on which model we ask. If each model $m \in \mathcal{M}$ has learned a (possibly biased) representation of human behavior for people in group i,

$$\mathbb{E}\left[\hat{Y}_{i,m}\right] = \mathbb{E}\left[Y_i\right] + \Delta_i = \mathbb{E}\left[\hat{Y}_{i,m'}\right] \ \forall\, m, m' \in \mathcal{M}, \quad (4)$$

where Δ_i is bias, which may depend on the task and population/subpopulation being emulated. This is a model of LLM behavior considered widely in the literature (Ludwig et al., 2024). We can reject this assumption if the outer terms in (4), which only depend on LLM output, are not equal. We say LLM outputs are **subjective** if there exist $m, m' \in M$ such that

$$\mathbb{E}[\hat{Y}_{i,m}] \neq \mathbb{E}[\hat{Y}_{i,m'}]. \quad (5)$$

4 Methods

4.1 Personas

In silicon sampling, a subpopulation is represented by a persona: a text description of the demographic group the LLM is being asked to emulate, placed in an LLM's system prompt. A system prompt can be understood as instructions for the LLM on *how* to behave as opposed to the user prompt, which is used to convey the task.

Our test for incoherence relies on the nested structure of populations. To generate useful data, we use a hierarchical set of personas which are nested in two dimensions: age and sex. We use adults living in the United States as our baseline parent population, which is then subdivided into subpopulations by age, sex, and age/sex combinations. We use four age brackets (18-29, 30-49, 50-64, and 65+) and two sex categories (man, woman). The text for all of our personas is provided in Appendix Section A.1.

This hierarchical structure ensures that every one of our most specific personas (for example, a woman aged 18-29) is a member of multiple parent populations (in this case, the group of all women, the group of all 18-29 year-olds, and the baseline adult population). This nesting allows us to test coherence along the 9 marginals depicted as rectangles in Figure 1.

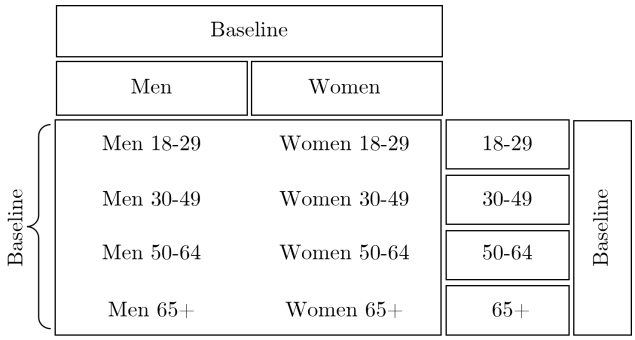


Figure 1: The hierarchical structure of personas. Each of the nine rectangles represents a marginal for which we can test incoherence. For example, the eight sex/age personas must be coherent with respect to the baseline. The two sex personas must be coherent with respect to the baseline. The four age personas must be coherent with respect to baseline. For each of the four age groups, the two sex/age personas with that age must be coherent with respect to the relevant age-only persona. For each of the two sexes, the four sex/age personas with that sex must be coherent with respect to the relevant sex-only persona.

In common practice, for example with digital twins, personas are significantly more detailed than those we use in this research. However, in our setting more detailed prompts add more marginal distributions, which makes it harder for models to pass a coherence test. In this sense, our choice of simple but richly nested personas is a conservative one.

4.2 Tasks

We selected a set of representative tasks from behavioral economics and opinion polling. The behavioral economics tasks include the dictator game, the prisoner's dilemma, the beauty contest game (guessing two-thirds of the average), and a risk elicitation task. The opinion polling tasks include questions on presidential approval, abortion legality, and the impact of AI, adapted from recent Pew Research Center surveys. We turned each task into a user prompt that instructs the LLM to respond with a single number. See Appendix Section A.1 for details.

4.3 Models and Sampling

We chose models to cover a range in terms of generation, size, and provenance under the criterion that we could find inference providers that return token distributions, which greatly

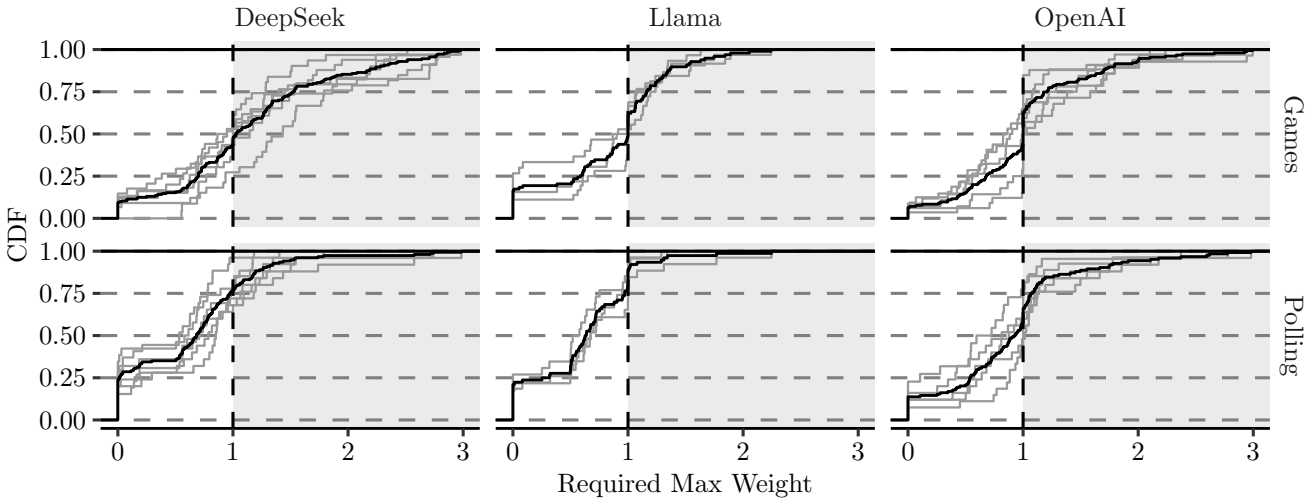


Figure 2: Required maximum weight on a single subpopulation implied by the parent mean. Weights above 1 imply incoherence. Grey lines show individual model CDFs of the required maximum weight, while the black line is pooled by model family. Grey area indicates region of incoherence. The x-axis is truncated from the right at 3 for legibility. Details on the weight construction are in Appendix Section A.2.

improves our ability to precisely measure the means predicted by each model.

To represent the model frontier at the time of our data collection, we include several models from OpenAI's GPT 5.4 family. To get a range of sizes, we include the standard model and its smaller mini and nano variants. Because GPT models have been widely used in this literature in the past few years, we include two previous-generation frontier models in GPT 3.5 Turbo and GPT 4o.

To represent frontier open-weight models we include several latest-generation DeepSeek models (V4). Within the DeepSeek family, we include the Flash and Pro size variants with max, high, and no reasoning variants of each.

For historical reasons, we also include the Llama family of models. While these are no longer frontier models, they have been used in the open-weight silicon sampling literature (Li et al., 2025; Xie et al., 2025). They are also small enough for us to run locally, which allows us direct access to the distribution of token responses.

For consistency, we used a temperature of 1 for all models and variants. A temperature of 1 corresponds to the "raw" predictions of the model, unadjusted by temperature, which is what typical API endpoints return when asked for a model's token distribution. This allows us to estimate an expectation even when thinking is on, since sampling with a temperature of 1 corresponds to the expected response (the mean token), whereas, e.g., a temperature of 0 would correspond to the most likely response (the modal token).

For models without chain of thought, the distributions of tokens are very consistent; the variation in output comes from changes in floating-point operations between batches (Yuan et al., 2026). For chain-of-thought models the predictive distribution usually degenerates to a single token with positive weight. There is still substantial variance in the responses, but this variance is not characterized well by the log probability distribution, which is conditioned on the thinking trace.

We use repeated sampling to precisely measure the mean prediction of thinking models. To achieve uniform power across models we use a cutoff rule that is unrelated to our hypothesis test, sampling until the plug-in standard errors fall below two percent of the range of the allowed response distribution.

5 Results

5.1 Incoherence

Recall that a model is incoherent for a task and marginal if the predicted parent population mean is outside the range of its constituent subpopulation means. For example, if the average 18-29 year old's approval rating for Donald Trump is higher than the average 18-29 year old man's approval rating and the average 18-29 year old woman's approval rating, then this distribution is incoherent. Otherwise, it is coherent.

Given the non-determinism of LLM output, predicted parent and subpopulation means are random variables. Although we sampled each model/task/persona combination aggressively to reduce sampling error, it is possible that a model appears incoherent by chance.

To assess incoherence, we test the null hypothesis that the parent mean is between the highest and lowest subpopulation means; a rejection implies incoherence by (3).

Denote by $\bar{Y}_p$ the sample mean of the LLM output for the parent group, and $\bar{Y}_i$ the sample mean of the LLM output for subpopulation i. Let $\bar{i}$ be the subpopulation associated with the largest mean, and $\underline{i}$ be the subpopulation associated with the smallest. We can only reject in cases where $\bar{Y}_p > \bar{Y}_{\bar{i}}$ *or* $\bar{Y}_p < \bar{Y}_{\underline{i}}$. This gives us two Z-values depending on which case holds:

$$Z_{\text{above}} = \frac{\bar{Y}_p - \bar{Y}_{\bar{i}}}{\sqrt{\widehat{\text{se}}_p^2 + \widehat{\text{se}}_{\bar{i}}^2}}, \qquad Z_{\text{below}} = \frac{\bar{Y}_{\underline{i}} - \bar{Y}_p}{\sqrt{\widehat{\text{se}}_p^2 + \widehat{\text{se}}_{\underline{i}}^2}}. \tag{6}$$

In Table 1, the "Rate (5%)" column shows the rate at which this test rejects coherence at the 5% level, while "Rate" shows the rate of incoherence implied by the point estimate.

On average, models are incoherent on 43% of their marginals. 33% of the time we can reject coherence at the 5% level. For comparison, a discrete uniform random number generator would be incoherent for about 53% of marginals.[2]

The lower rate of incoherence for GPT 5.4 is notable. The main driver is a tendency to produce distributions where all marginals are *trivially coherent* because it responded the same way for all personas. It uniformly responded with 0 in the beauty contest game and 5 in the dictator game, and responded 2 for the question on legalizing abortion for all but one persona.[3] Interestingly, GPT 5.4 was the only model that regularly responded in this manner.

Of the tasks where GPT 5.4 did not predict a constant mean, its overall rate of incoherence was 36%, and we can reject coherence at the 5% level for 31% of its marginals. These numbers are far closer to the rest of the field. Thus, GPT 5.4 is not more coherent than other models because it is faithfully emulating true data, but rather because it sometimes completely ignored the persona conditioning, a form of extreme subjectivity.

These cases demonstrate that incoherence is a weak, coarse test of the "unnaturalness" of LLM predictions. In addition to this trivial coherence, models can also be coherent but make predictions which can only be justified by extreme subpopulation weights. For example, predictions that the average height of a man is 5 feet 9 inches, the average height of a woman is 5 feet 4 and the average height of an adult is 5 feet 8.5 inches is a coherent set of predictions. However, these means can only be justified if men make up 90% of the adult population. The predictions are nearly incoherent.

To provide a more nuanced picture of incoherence, we develop a continuous statistic that nests incoherence, while also allowing us to measure trivial coherence and near-incoherence. For each model, task, and marginal, we calculate the minimum weight that could be placed on each subpopulation mean to justify the parent population mean. We then take the maximum over these to find the most constrained subpopulation weight. We call this the **maximin weight**.

When there are only two subpopulations, as in the example above, the calculation of the maximin weight is straightforward. In the example, the only weights on the male and female subpopulations that justify the adult mean are 0.9 (men) and 0.1 (women). Since the maximum of these is 0.9, that is the maximin weight.

For an incoherent marginal, the maximin weight is greater than 1. Suppose in the example above that the adult mean was predicted to be 5 feet 10 inches. This is incoherent since it exceeds the largest subpopulation mean (men). In this case, the only weights that justify the mean are 1.2 (for men) and -0.2 (for women). The maximum of these is 1.2. When there are more than two subpopulations, the calculations are slightly more complicated but can be reduced to a rather simple formula that we derive in Appendix Section A.2.

Model	**Rate**	**Rate (5%)**
DeepSeek		
DeepSeek Flash	44%	40%
DeepSeek Flash High	52%	29%
DeepSeek Flash Max	56%	30%
DeepSeek Pro	54%	48%
DeepSeek Pro High	37%	27%
DeepSeek Pro Max	38%	22%
Llama		
Llama 3.1 8b	35%	35%
Llama 3.1 70b	32%	32%
Llama 3.3 70b	41%	41%
OpenAI		
GPT 3.5	43%	37%
GPT 4o	40%	32%
GPT 5.4 nano	54%	37%
GPT 5.4 mini	59%	37%
GPT 5.4*	21%	17%

Table 1: Rates of statistical incoherence by model, as defined in (3). Rate reports the percent of marginals that are incoherent, and Rate (5%) reports the rate at which we can reject coherence at the 5% level for each task. Many of the GPT 5.4 marginals are *trivially* coherent because the model predicts the same mean across all personas, deflating the reported incoherence relative to other models.

Figure 2 shows the CDFs of the maximin weights for our data. Each column represents a model family. Grey lines correspond to models within the family (the models are described in Section 4.3), and the black line is the pooled CDF across all models within the family. The first row corresponds to the four games we ask the LLMs to play: the prisoner's dilemma, the beauty contest, the dictator game, and a risk aversion elicitation. The second row corresponds to opinion polling questions taken from Pew: AI change, legalization of abortion, and Trump's approval rating. Details on persona and task prompts are in Appendix Section A.1.

In the case where subpopulations tie, the weights are ambiguous, so we first limit to unique responses. This tie-breaking convention generates a noticeable jump at 1 which can be seen for many models in Figure 2, and which often comes from a nearly degenerate distribution that reports constants across all subpopulations. As discussed earlier, this is especially common in GPT 5.4's responses, which can be seen in the particularly large spike at 1 in the CDF of one of the OpenAI models.

Models tend to be more coherent on polling-related questions, but we still see failures in more than a quarter of responses from DeepSeek and Llama models, and in half of OpenAI's responses. Otherwise, we do not observe systematic differences in the distribution of maximin weights across or within model families.

[2]There are $n+1$ locations the parent group could fall in the ordering of these means. Two are incoherent. Thus, the probability of incoherence for randomly generated means for a marginal with n subpopulations is $2/(n+1)$.

[3]In the case of the legalizing abortion question, GPT 5.4 produced an answer of 2 for all personas except men 65 and older, for whom it predicted a mean response of 2.08. GPT 5.4's responses were uniformly constant across all personas in the dictator game and beauty contest.

5.2 Subjectivity

Recall that our definition of subjectivity in (5) is that models learn systematically different population/subpopulation means. In contrast, if models learn a common (but possibly biased) model of human behavior which is latent in the training data, then we should only observe noise around a common persona mean, without systematic model effects.

With this logic in mind, we use the following procedure to test the formal subjectivity criterion in (5) while taking into account the sampling error of our model predictions.

For a given task, define $\mathbb{E}[\hat{Y}] = \mu$ as the grand mean across models and persona prompts, $\mu_i = \mathbb{E}[\hat{Y}_i]$ as the mean across models for a given persona prompt, $\mu_m = \mathbb{E}[\hat{Y}_m]$ as the mean across persona prompts for a given model, and $\mu_{i,m} = \mathbb{E}[\hat{Y}_{i,m}]$ as the mean of each persona prompt / model combination.

We can write the difference in cell-specific means from the grand mean in terms of its components:

$$\mu_{i,m} - \mu = \underbrace{(\mu_i - \mu)}_{\text{prompt}} + \underbrace{(\mu_m - \mu)}_{\text{model}} + \underbrace{(\mu_{i,m} - \mu_i - \mu_m + \mu)}_{\text{residual}} \quad (7)$$

Note that under the null hypothesis, $\mu_{i,m} = \mu_i \ \forall i, m$ and $\mu_m - \mu = 0 \ \forall m$. Taking the variance of both sides and dividing by $\text{Var}(\hat{Y})$, we get

$$R^2_{i,m} = R^2_i + R^2_m + R^2_{resid} \quad (8)$$

because the components are orthogonal by construction. This allows us to read off two clean tests of the hypothesis: if $R^2_{i,m} > R^2_i$ or if $R^2_m > 0$ we can reject the sharp null of mean equality in favor of subjectivity as defined in (5).

This procedure also tells us the degree to which a model is subjective, or the degree to which predictions are driven by the prompt or the model. R^2 is commonly interpreted as the percentage of the variance explained by the conditional means, which gives each term an effect size interpretation. We report per-task results in Table 2.[4]

Table 2 gives a firm answer: the model, and not the prompt, drives most of the variation in average responses. In every case, models explain more of the variation than prompts, typically by enormous magnitudes. Across tasks, the model explains at *minimum* around three times as much of the variation in responses than the persona prompt. Thus we can not only conclude that LLMs are subjective in a statistical sense, but also that the magnitude of subjectivity is substantial.

6 Conclusion

In this paper, we have documented a set of systematic failures of LLMs to emulate populations. First, LLMs are incoherent, generating data with impossible statistical relationships about half of the time, a rate similar to purely random predictions. Second, LLMs are subjective, generating data that cannot be understood as arising from a common representation of human behavior. Instead, the data is highly dependent on the chosen model, and presumably on details like model architecture, training pipeline, fine-tuning, alignment, and parameters like thinking effort.

Task	R^2_i	$R^2_i + R^2_m$	$R^2_{i,m}$
Games:			
Beauty Contest	0.00	0.72	0.74
Dictator	0.01	0.35	0.39
Prisoner's Dilemma	0.01	0.61	0.68
Risk Aversion	0.05	0.44	0.51
Polling:			
AI Change	0.08	0.42	0.49
Legalize Abortion	0.14	0.40	0.48
Trump Approval	0.10	0.38	0.48

Table 2: R^2_i includes only persona averages, $R^2_i + R^2_m$ includes persona and model average effects, and $R^2_{i,m}$ includes crossed averages.

Some may argue that the extent of incoherence and subjectivity that we document is due to our failure to use a set of prompts that let LLMs achieve their full potential. A different set of prompts (Manning and Horton, 2026) or appropriate fine-tuning (Xie et al., 2025) may lead to better results. It could also be that our prompts were simply not comprehensive enough. For example, some American adults do not identify as men or women. While this is possible, we believe our results offer a meaningful critique of the argument that LLMs are capable of emulating real human populations.

We emphasize that no real data would fail these tests. In fact, coherence is weaker than the structure inherent in real data. While our notion of coherence asks whether, for a specific marginal, there are weights that can justify the parent population mean given the set of subpopulation means *per task*, for a real dataset, the same weights would work *across all tasks*.

In light of these results, we believe researchers should be skeptical of using LLM output in place of human subjects.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR.

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Brynjolfsson, E., Evans, J., and Bernstein, M. S. (2025). Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

[4] We estimate this with a linear regression, weighting each model equally within a task / persona cell and weighting tokens by their probability. This regression was run per-task with standard errors clustered at the sample level, to account for the covariance in the sample-level logits. Formal test statistics on the model restriction are large to the point they are not useful (e.g. with F-statistics on the order of 10^{13}), and in cases of "collapse" not estimable as the problem becomes singular.

Ashokkumar, A., Hewitt, L., Ghezae, I., and Willer, R. (2026). Large language models can predict the results of social science experiments. *Nature*, pages 1–8.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. (2024). Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Boelaert, J., Coavoux, S., Ollion, É., Petev, I., and Präg, P. (2025). Machine bias. how do generative language models answer opinion polls? *Sociological Methods & Research*, 54(3):1156–1196.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can ai language models replace human participants? *Trends in cognitive sciences*, 27(7):597–600.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. (2024). Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., and Cunningham, W. A. (2023). Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Henning, T. and Camerer, C. F. (2026). Synthetic ai agents for experimental social science. In *Proceedings of AI for Accelerated Research Symposium*, volume 3, pages 77–84.

Horton, J. J., Filippas, A., and Manning, B. S. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Hullman, J., Broska, D., Sun, H., and Shaw, A. (2026). This human study did not involve human subjects: Validating llm simulations as behavioral evidence. *arXiv preprint arXiv:2602.15785*.

Li, D., Li, L., and Qiu, H. S. (2025). Chatgpt is not a man but das man: Representativeness and structural consistency of silicon samples generated by large language models. *arXiv preprint arXiv:2507.02919*.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2024). Large language models: An applied econometric framework. *Annual Review of Economics*, 18.

Manning, B. S. and Horton, J. J. (2026). General social agents. Technical report, National Bureau of Economic Research.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52.

Pataranutaporn, P., Powdthavee, N., Archiwaranguprok, C., and Maes, P. (2025). Simulating human well-being with large language models: Systematic validation and misestimation across 64,000 individuals from 64 countries. *Proceedings of the National Academy of Sciences*, 122(48):e2519394122.

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

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., and Neubig, G. (2024). Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., and Chen, H. (2025). Database report: 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*, 44(6):1446–1455.

Xie, Y., Li, Z., Wang, X., Pan, Y., Liu, Q., Cui, X., Lo, K.-Y., Gao, R., Zhang, X., Huang, J., et al. (2025). Be. fm: Open foundation models for human behavior. *arXiv preprint arXiv:2505.23058*.

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

Yuan, J., Li, H., Ding, X., Xie, W., Li, Y.-J., Zhao, W., Wan, K., Shi, J., Hu, X., and Liu, Z. (2026). Understanding and mitigating numerical sources of nondeterminism in llm inference. *Advances in Neural Information Processing Systems*, 38:169819–169851.

A Appendix

A.1 Prompts

In this section, we provide additional details on the prompts used in our research. We have two sets of prompts: personas and tasks. Personas are provided as system prompts, and tasks are provided as user prompts.

A.1.1 Persona Prompt Wording

Baseline: You are an adult living in the United States.
18-29: You are an adult living in the United States. You are between 18 and 29 years old.
30-49: You are an adult living in the United States. You are between 30 and 49 years old.
50-64: You are an adult living in the United States. You are between 50 and 64 years old.
65+: You are an adult living in the United States. You are 65 years old or older.
Male: You are an adult living in the United States. You are a man.
Female: You are an adult living in the United States. You are a woman.
Male, 18-29: You are an adult living in the United States, you are between 18 and 29 years old, and you are a man.
Male, 30-49: You are an adult living in the United States, you are between 30 and 49 years old, and you are a man.
Male, 50-64: You are an adult living in the United States,

you are between 50 and 64 years old, and you are a man.
Male, 65+: You are an adult living in the United States, you are 65 years old or older, and you are a man.
Female, 18-29: You are an adult living in the United States, you are between 18 and 29 years old, and you are a woman.
Female, 30-49: You are an adult living in the United States, you are between 30 and 49 years old, and you are a woman.
Female, 50-64: You are an adult living in the United States, you are between 50 and 64 years old, and you are a woman.
Female, 65+: You are an adult living in the United States, you are 65 years old or older, and you are a woman.

A.1.2 Task Prompt Wording

Prisoner's Dilemma: You are playing a game with a partner. You choose whether to cooperate or not. If you both choose to cooperate, you both receive $5. If you choose to cooperate and your partner chooses not to, you receive $0 and your partner receives $10. If you choose not to cooperate and your partner chooses to cooperate, you receive $10 and your partner receives $0. If you both choose to not cooperate, you both receive $2. Will you choose to cooperate or not? Respond with just a number 0 or 1. Respond with 1 for 'cooperate' and 0 for 'not cooperate'.
Beauty Contest (Guess 2/3 Average): You are playing a game with a large group of people. Everyone must choose a number between 0 and 10 (inclusive). The person whose number is closest to 2/3 of the average of all numbers chosen wins $10. What number do you choose? Respond with just a number between 0 and 10 (inclusive).
Risk Aversion: You receive $10. You can keep as much of this money as you want, or you can choose to invest some or all of it into a risky option. The money you invest has a 50% chance of being lost, and a 50% chance of being multiplied by 2.5 and returned to you. You keep any money you do not invest. How much of your $10 will you choose to invest? Respond with just a number between 0 and 10 (inclusive).
Dictator Game: You are playing a game where you receive $10. You must choose how much of this money to split with a partner who has $0. How much will you keep for yourself? Respond with just a number between 0 and 10 (inclusive).
AI Change: You are responding to an opinion survey. Do you think that the impact of artificial intelligence (AI) on the U.S. over the next 20 years will be... Please respond with one of the following options: 1 = Very positive, 2 = Somewhat positive, 3 = Equally positive and negative, 4 = Somewhat negative, 5 = Very negative. Respond with just a number between 1 and 5 (inclusive).
Abortion: You are responding to an opinion survey. Do you think abortion should be... Please respond with one of the following options: 1 = Legal in All Cases, 2 = Legal in Most Cases, 3 = Illegal in Most Cases, 4 = Illegal in All Cases. Respond with just a number between 1 and 4 (inclusive).
Trump Approval: You are responding to an opinion survey. Do you approve or disapprove of the way Donald Trump is handling his job as President? Please respond with one of the following options: 1 = Strongly Approve, 2 = Somewhat Approve, 3 = Somewhat Disapprove, 4 = Strongly Disapprove. Respond with just a number between 1 and 4 (inclusive).

A.2 Calculating Maximin Weights

In this section, we derive a closed-form expression to calculate maximin weights.

Let $\bar{Y}_p$ be the parent mean, and let $\bar{Y}_i$ be the subpopulation mean for subpopulation $i \in \{1, ..., K\}$. Let W be the set of weight vectors $w = (w_1, w_2, ..., w_K)$ which sum to one and justify the population mean:

$$W \equiv \left\{ w \middle| \sum_{i=1}^{K} w_i \bar{Y}_i = \bar{Y}_p, \sum_{i=1}^{K} w_i = 1 \right\}. \tag{9}$$

We do not restrict weights to be positive. When the data is incoherent such that $\bar{Y}_p$ is outside the convex hull of the subpopulation means, negative weights are required to justify the mean and solve the condition in (9). However, if there are three or more subpopulations, the possibility of arbitrary negative weights implies that, for any subpopulation, there exists a weight vector where $w_i = 0$.

For this reason, we restrict negative mass to be as small as possible. For coherent means, this implies the weight vectors considered in the calculation of the maximin weight are restricted to actual distributions (non-negative weights). When the means are incoherent, no distribution of the subpopulations can justify the parent mean. Instead, what we consider in the calculation of the maximin weight is the set of weight vectors that justify the parent mean and are as close as possible (in terms of negative mass) to a true probability distribution. The amount of negative mass acts as a continuous measure of how far away the data is from coherence.

To formalize this, we construct the set S of weight vectors in W that have the smallest negative weight. Let l be the smallest negative mass of all weight vectors in W. That is:

$$l \equiv \min_{w \in W} \left\{ \sum_{i=1}^{K} \left| \min\{w_i, 0\} \right| \right\}. \tag{10}$$

Using this, we can define S as,

$$S \equiv \left\{ w | w \in W, \sum_{i=1}^{K} \left| \min\{w_i, 0\} \right| = l \right\}. \tag{11}$$

We can now define the maximin weight. Let $\underline{w}_i$ be the smallest w_i in the set S. This is the minimum weight needed on subpopulation i in any set of weights that justify the parent population's mean:

$$\underline{w}_i \equiv \min\{w_i | w \in S\}. \tag{12}$$

The subpopulation weight that is *most constrained* by the data is then the maximum over these minimum weights for each subpopulation:

$$w_{\text{maximin}} = \max_{i \in \{1,...,K\}} \underline{w}_i. \tag{13}$$

When there are two identical subpopulation means, there always exist weight vectors where either of them is zero. This is the case even if, in every weight vector in S, one of these weights must be strictly positive. This only happens with identical values, i.e. when a model "collapses" and reports a

constant response across subpopulations. This never happens in real data: samples from two different subpopulations will always have different means.

In order to make the statistic well defined, we take $\bar{Y}_i$ to represent the *unique* subpopulation means, and interpret weights in the context of *distinct* subpopulations.[5]

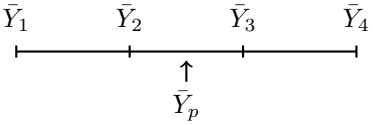
Furthermore, we assume without loss of generality that the unique subpopulation means are ordered:

$$\bar{Y}_1 < \bar{Y}_2 < ... < \bar{Y}_{K-1} < \bar{Y}_K. \tag{14}$$

When we have only two subpopulations, S is a singleton, so the calculation of w_{maximin} is straightforward. However, we show below that the following closed form expression is equivalent to w_{maximin} for any number of subpopulations.

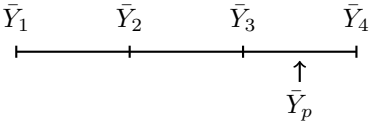
$$w_{\text{maximin}} = \begin{cases} \frac{\bar{Y}_K - \bar{Y}_p}{\bar{Y}_K - \bar{Y}_1} & \bar{Y}_p < \bar{Y}_1 \\ \frac{\bar{Y}_2 - \bar{Y}_p}{\bar{Y}_2 - \bar{Y}_1} & \bar{Y}_1 \leq \bar{Y}_p < \bar{Y}_2 \\ 0 & \bar{Y}_2 \leq \bar{Y}_p \leq \bar{Y}_{K-1} \\ \frac{\bar{Y}_p - \bar{Y}_{K-1}}{\bar{Y}_K - \bar{Y}_{K-1}} & \bar{Y}_{K-1} < \bar{Y}_p \leq \bar{Y}_K \\ \frac{\bar{Y}_p - \bar{Y}_1}{\bar{Y}_K - \bar{Y}_1} & \bar{Y}_p > \bar{Y}_K \end{cases} \tag{15}$$

To see how this expression works, consider the case with four unique subpopulation means. We start with the scenario that $\bar{Y}_p$ is between the middle two unique subpopulation means. This is the third case in (15).



In this case we can find weights where $\bar{Y}_p$ is either a weighted average of $\bar{Y}_2$ and $\bar{Y}_3$, or a weighted average of $\bar{Y}_1$ and $\bar{Y}_4$. While weights must sum to one, no single point is *required* to have strictly positive weight and so $w_{\text{maximin}} = 0$.

Now suppose $\bar{Y}_p$ is adjacent to one of the extremes. Here, we take the case of $\bar{Y}_3 < \bar{Y}_p < \bar{Y}_4$, but the logic also applies to $\bar{Y}_1 < \bar{Y}_p < \bar{Y}_2$. These correspond to the fourth and second case of (15) respectively.

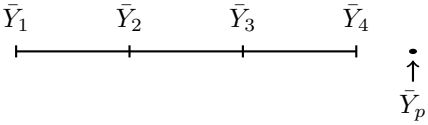


Here, the *only* way to justify $\bar{Y}_p$ is to put strictly positive weight on $\bar{Y}_4$. There are many averages that do this, but the one which requires the *smallest* weight on w_4 is the unique weighted average of $\bar{Y}_3$ and $\bar{Y}_4$ equal to $\bar{Y}_p$. The unique weight on subpopulation 4 in this case is given by the following formula:

$$w_4 = \frac{\bar{Y}_p - \bar{Y}_3}{\bar{Y}_4 - \bar{Y}_3}. \tag{16}$$

Now consider the case where $\bar{Y}_p$ is outside of all of these points. We consider $\bar{Y}_p > \bar{Y}_4$ here, but the logic also applies to $\bar{Y}_p < \bar{Y}_1$. These are the first and fifth cases of (15) and are the incoherent possibilities.

[5]Note that this assumption is only used for calculating maximin weights. For example, Table 1 reports incoherence without relying on this unique value construction.



In this case, S is a singleton. The weight vector with the smallest negative mass requires that weight $\frac{\bar{Y}_p - \bar{Y}_1}{\bar{Y}_4 - \bar{Y}_1}$ be placed on $\bar{Y}_4$, the balancing negative weight goes to the furthest-away element of the set, $\bar{Y}_1$, and the other weights must be zero.

Now we prove that (15) is the minimax weight in S.

Proposition 1. *Let S, $\bar{Y}_i$ and $\bar{Y}_p$ be defined as above. Then the expression* (15) *is the maximin weight.*

Proof. We assume without loss of generality, as above, that $\bar{Y}_i$ are sorted by size with $\bar{Y}_1$ the smallest. Suppose that there exist two elements $\bar{Y}_1 < \bar{Y}_2 \leq \bar{Y}_p$ and two elements such that $\bar{Y}_p \leq \bar{Y}_{K-1} < \bar{Y}_K$. This is the third case of (15).

Recall that in our construction we work with *unique* responses, so exact equality of two means is impossible. Then the maximin weight is immediately zero, as there exist weights such that $w_1\bar{Y}_1 + w_K\bar{Y}_K = \bar{Y}_p$ and $w_2\bar{Y}_2 + w_{K-1}\bar{Y}_{K-1} = \bar{Y}_p$.

Next, consider the fourth case of (15) where $\bar{Y}_K > \bar{Y}_p > \bar{Y}_{K-1}$. This case is coherent, so all weight vectors in S have zero negative mass. Because of this, all elements of S must have positive weight on $\bar{Y}_K$, as any other construction requires negative weights.

The weights w justify the parent mean if:

$$\bar{Y}_p - \sum_{i=1}^{K-1} \bar{Y}_i w_i = w_K \bar{Y}_K. \tag{17}$$

Thus, w_K is minimized by making $\sum_{i=1}^{K-1} \bar{Y}_i w_i$ as close as possible to $\bar{Y}_p$, putting all the weight on $\bar{Y}_{K-1}$. Since w_K is the only weight required to be positive, this is all we need to consider. The case from below, case two from (15), is analogous.

Finally, consider the case five from (15) where $\bar{Y}_p > \bar{Y}_K$. S is a singleton since it requires minimizing the total negative mass. Because weights sum to one, this is equivalent to minimizing total positive mass. Splitting our problem into positive and negative mass,

$$\begin{aligned} \bar{Y}_p &= \sum_{\{i|w_i>0\}} w_i\bar{Y}_i + \sum_{\{i|w_i<0\}} w_i\bar{Y}_i \\ &\leq \bar{Y}_K \sum_{\{i|w_i>0\}} w_i + \bar{Y}_1 \sum_{\{i|w_i<0\}} w_i. \end{aligned} \tag{18}$$

Because weights sum to one, $\sum_{\{i|w_i<0\}} w_i = 1 - \sum_{\{i|w_i>0\}} w_i$. Substituting and re-arranging, we have

$$\sum_{\{i|w_i>0\}} w_i \geq \frac{\bar{Y}_p - \bar{Y}_1}{\bar{Y}_K - \bar{Y}_1}, \tag{19}$$

which is minimized at equality. The case from below, case one in (15), is exactly analogous. □