HALLUCINATED HETEROGENEITY: LLM SOCIAL SIMULATIONS EXAGGERATE WHO RESPONDS TO INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) social simulations are increasingly used as a fast, cheap substitute for human experiments, and are typically validated on whether they recover average treatment effects. But targeting and personalization depend on a different question: *who* responds? We introduce **H-Bench** (Heterogeneity Bench), a benchmark of 138 randomized experiments spanning economics, politics, health, and other domains, which uses causal inference to test whether simulations recover treatment effect heterogeneity, from subgroup differences to the value of targeting decisions, against randomized human outcomes. Across nine LLMs, we find **Hallucinated Heterogeneity**: simulations track average effects far better than they track who responds, and they exaggerate how much people differ. Claude Sonnet 5's average effects correlate with human estimates at $r = 0.74$, but its subgroup differences in treatment effects correlate at only $r = 0.14$ ($p < 0.001$ for the difference), a gap that persists after correcting for human sampling noise (0.80 vs. 0.29, $p = 0.04$). Eight of nine models overstate heterogeneity by 1.4–$2.7\times$, and LLM-based targeting performs near random assignment, even after calibration on human data. We explain why this failure arises and goes undetected. Simulated responses are often far from human ones, but the errors are similar with and without treatment and cancel in the average effect, so standard validation passes. And the heterogeneity that models produce follows familiar divides such as partisanship, where models exaggerate it without identifying which groups respond more. Together, these results show that LLM social simulations must be validated at the level at which they will be used before they can support targeting.

1 INTRODUCTION

Many decisions in social science, policy, and industry follow a common workflow. Before deploying a new message, product, or policy, decision-makers run a randomized experiment to learn whether it works and who will respond. The second question often matters most in practice: which people a public health message will persuade, which customers a new product will retain, or which voters a policy proposal will win over. Human experiments, however, are slow and costly, and must be rerun for every new message, product, or policy.

Large language model (LLM) social simulations offer a fast, cheap alternative. Given a description of a person, an LLM can generate how that person would respond with and without an intervention, making it possible to run an experiment on a synthetic population (Argyle et al., 2023; Aher et al., 2023; Horton et al., 2023; Park et al., 2024; Piao et al., 2025). These simulations can now scale to populations of billions of personas (Li et al., 2026). Across dozens of published experiments, they have also shown an ability to predict whether interventions are effective on average (Ashokkumar et al., 2026). Encouraged by these results, commercial platforms increasingly advertise simulations that go beyond estimating average effects to predict who will respond to a new product, price, or message.

However, predicting whether an intervention works is not the same as predicting who responds. Consider a video urging action against Russian election interference (Tappin et al., 2023). In a randomized experiment, the video increased support among conservatives by 0.35 standard deviations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

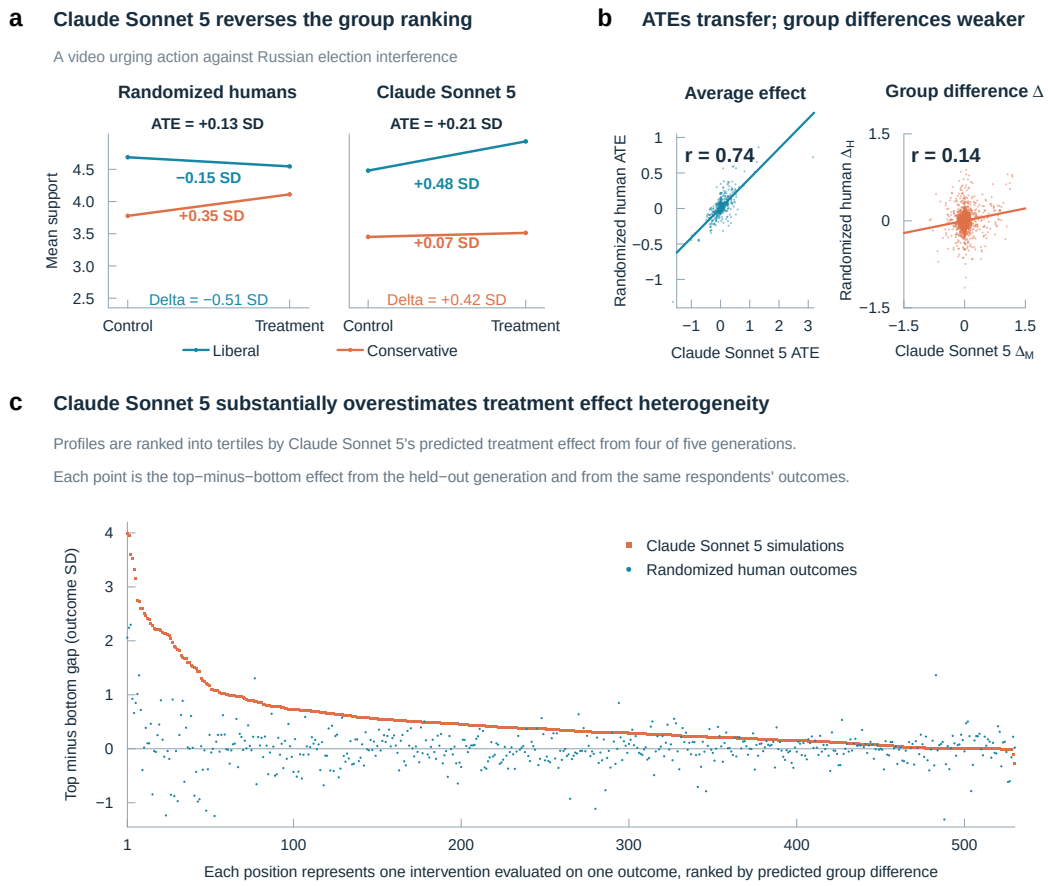

Figure 1: **Agreement on average effects extends only weakly to heterogeneity.** **(a)** Example in which Claude Sonnet 5 reverses which political group responds more strongly. **(b)** Claude Sonnet 5 estimates of average treatment effects and differences in treatment effects between prespecified groups, compared with randomized human estimates. **(c)** The difference in average treatment effects between the top and bottom tertiles of Claude Sonnet 5's predicted CATE is substantially larger in Claude Sonnet 5 simulations than when the same groups are evaluated using randomized human outcomes. Tertiles and both gaps are cross-fitted over the five generations.

(SD) and slightly decreased it among liberals (-0.15 SD). As shown in Figure 1a, Claude Sonnet 5 reverses this pattern: it predicts a $+0.48$ SD effect among liberals but only $+0.07$ SD among conservatives. A campaign that relied on this simulation would focus on liberals, the very group the video pushed in the opposite direction.

This example highlights a general problem: a simulation can predict average effects well while misjudging who responds. Because errors for different people and groups can cancel out when averaged, an accurate average treatment effect (ATE) provides little evidence that the simulated differences between people are correct. Existing evaluations compare simulated and human ATEs or response distributions (Ashokkumar et al., 2026; Kolluri et al., 2025; Hu et al., 2025), or compare opinions across groups without any intervention (Santurkar et al., 2023; Chen et al., 2026), but rarely test whether simulated differences in how groups *respond to an intervention* match randomized experiments. Such tests are hard to build: they require individual-level experimental data with respondent characteristics.

To address this gap, we introduce **H-Bench** (Heterogeneity Bench), a benchmark of 138 randomized human experiments spanning economics, politics, health, immigration, education, and moral decision-making, with 530 treatment–outcome contrasts and 2,906 subgroup comparisons. For ev-

ery respondent observed in the human data, we simulate responses with and without the intervention, then test whether the simulation recovers average effects, differences between groups, and the value of targeting decisions. Evaluating nine LLMs, we find **Hallucinated Heterogeneity**: simulations track average effects far better than they track who responds, and they exaggerate how much people differ. Claude Sonnet 5's average effects correlate with human ones ($r = 0.74$), but its subgroup differences in treatment effects correlate only weakly ($r = 0.14$), and eight of nine models overstate how much effects vary, by 1.4–2.7$\times$ (Figure 1; Appendix Table 3). Used for targeting, LLM rankings perform near random assignment, even after calibration on real data. Standard validation misses this failure because simulation errors cancel out in the average effect, and the heterogeneity that models do produce follows familiar group divides, especially partisan ones, where models exaggerate it without identifying which groups respond more.

To summarize, our contributions are:

- **H-Bench.** A causal framework and benchmark of 138 randomized experiments for testing whether LLM simulations recover who responds, not just whether interventions work on average (Section 2).
- **Hallucinated Heterogeneity.** In all nine LLMs, subgroup differences are recovered less well than average effects, and unevenly; eight of nine overstate heterogeneity by 1.4–2.7$\times$, leaving targeting little better than random (Sections 3–4).
- **Why it arises and goes undetected.** Errors cancel between treatment and control, so a simulator can pass average-effect validation while failing at targeting; its heterogeneity follows familiar group divides but misjudges how groups respond (Section 5).

2 A Causal Framework and Benchmark for Evaluating LLM Social Simulations

We first define what an ideal simulator would recover and derive the quantities we evaluate (Section 2.1), then describe H-Bench, the benchmark on which we evaluate them (Section 2.2).

2.1 Causal Framework

Human experiments and matched simulations. We use the potential-outcomes framework (Ding, 2024). Let X_i denote respondent i's pretreatment characteristics, $Z_i \in \{0, 1\}$ randomized assignment to control or treatment, and $Y_i^H(z)$ the human potential outcome under condition z. A human experiment reveals only the assigned outcome, $Y_i = Y_i^H(Z_i)$. A simulation instead receives the same profile X_i under *both* conditions and returns $Y_i^M(0)$ and $Y_i^M(1)$ (Appendix Figure 5b). These are predictions, not observations of the missing potential outcome.

What an ideal simulator recovers. An ideal simulator reproduces the average human response for every profile in both arms:

$$\mathbb{E}[Y_i^M(z) \mid X_i = x] = \mathbb{E}[Y_i^H(z) \mid X_i = x], \qquad z \in \{0, 1\}. \tag{1}$$

It need not predict any individual's response, only the average among respondents who share a profile. For each source $s \in \{H, M\}$, define the conditional average treatment effect (CATE),

$$\tau^s(x) = \mathbb{E}[Y_i^s(1) - Y_i^s(0) \mid X_i = x]. \tag{2}$$

Equation 1 implies $\tau^M(x) = \tau^H(x)$ for every profile. An ideal simulator would therefore recover the average effect, the effect in every subgroup, and the correct ranking of whom to treat.

What we evaluate. Equation 1 cannot be tested directly: each human is observed in only one arm, and most profiles are shared by too few respondents to estimate their mean response. We instead evaluate six of its consequences, each identified by randomization at the level of groups rather than individuals, and organize them by the question they answer (Table 1):

- **Who responds** asks whether agreement on the average effect, the standard target in prior validations, extends to heterogeneity. The baseline gap is descriptive: it asks whether

groups differ without the intervention. The treatment interaction is causal: it asks whether the intervention moves them differently. Contrasting the two separates knowing how groups differ from knowing how they respond. The ranked-group gap asks the same question for the model's own predicted responders rather than for prespecified groups.

- **Whom to treat** evaluates the decision that targeting and personalization actually make. It is scored only on human outcomes, so it measures the value of the model's ranking directly.
- **Why it arises and goes undetected** compares simulated and human outcome levels in each arm. A model that misses responses by similar amounts in both arms can still recover the average effect, so accurate averages need not reflect accurate responses. We also break down baseline gaps and interactions by the characteristic defining the groups, to see what drives the heterogeneity that models produce.

Table 1 illustrates every quantity with the running example from Section 1: a video urging action against Russian election interference, with support for action as the outcome (Tappin et al., 2023).

Table 1: Validation targets implied by the ideal simulator, illustrated with the running example.

Target	Estimand	In the running example
Who responds		Section 3
Average effect τ	$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$	Does the video raise support on average?
Baseline gap $B(0)$	$\mathbb{E}[Y(0) \mid G{=}1] - \mathbb{E}[Y(0) \mid G{=}0]$	Do liberals and conservatives differ without it?
Interaction Δ	$\mathbb{E}[\delta \mid G{=}1] - \mathbb{E}[\delta \mid G{=}0]$	Which group does the video move more?
Ranked-group gap	$\mathbb{E}[\delta \mid T{=}3] - \mathbb{E}[\delta \mid T{=}1]$	Do predicted top responders respond more?
Whom to treat		Section 4
Targeting value	$\mathbb{E}[(\pi_q(X) - q)\,\delta]$	Does treating the model's top picks beat random?
Why it arises and goes undetected		Section 5
Outcome levels	$\mathbb{E}[Y(0)],\ \mathbb{E}[Y(1)]$	Average support with and without the video

Notes. $Y(1), Y(0)$: responses with and without the intervention; $\delta = Y(1) - Y(0)$. $G \in \{0, 1\}$: two prespecified groups (e.g., liberals vs. conservatives). $T \in \{1, 2, 3\}$: tertile of model-predicted CATE. $\pi_q(X) \in \{0, 1\}$: treats the fraction q with the largest model-predicted CATE.

2.2 H-BENCH

Experiments and simulation. H-Bench combines 138 randomized experiments from five public sources (Ashokkumar et al., 2026; Tappin et al., 2023; Coppock, 2023; Klein et al., 2018; Zhou et al., 2026a) spanning six substantive areas (Appendix Figure 5a). We keep each experiment's original treatment wording, outcome question, response scale, and pretreatment respondent attributes. Each treatment–control pair evaluated on one outcome forms a *contrast*, giving 530 contrasts. Splitting respondents within contrasts by political affiliation, ideology, gender, race, age, or education gives 2,906 *subgroup comparisons*, each with at least 10 human respondents per group-by-arm cell. For each of nine LLMs (Appendix A), we prompt with every distinct respondent profile observed in the human data, under every arm, five times each, asking for one response on the original scale. We average the five responses and weight profiles by the number of human respondents they represent, so both sources share the same profile distribution.

Estimation and comparison. Apart from targeting value, the quantities in Table 1 are means or differences of means, computed identically from human and simulated data in human outcome SD units. The model's CATE for a profile is the difference between its mean simulated responses under treatment and control; ranked-group gaps split profiles into tertiles of this score within each contrast. We summarize agreement across the benchmark by Pearson correlation for τ, $B(0)$, Δ, and ranked-group gaps, by RMSE for outcome levels and average effects, and by the ratio of model to human root-mean-square magnitude for ranked-group gaps. Baseline gaps use the subset of subgroup comparisons with arm-level estimates for all nine models (Appendix B). Targeting value uses a held-out design described in Section 4.

3 LLM SOCIAL SIMULATIONS SHOW HALLUCINATED HETEROGENEITY

We test whether simulations that recover average effects also recover who responds. We first compare model and human estimates for prespecified respondent groups (Section 3.1), then ask whether the respondents a model predicts will respond most actually do (Section 3.2). Both comparisons account for noise. Human subgroup estimates are imprecise, so we also report correlations corrected for human sampling noise (Appendix C.3). Simulations are random, so we cross-fit ranked groups and their gaps over the five generations (Appendix C.2).

3.1 LLM SIMULATIONS PREDICT AVERAGE TREATMENT EFFECTS BETTER THAN WHICH GROUPS RESPOND MORE

Figure 1b contrasts the two targets for Claude Sonnet 5. Across the 530 treatment–outcome contrasts, its average treatment effects correlate with randomized human estimates at $r = 0.74$. Across the 2,906 subgroup comparisons drawn from the same experiments, its treatment interactions Δ correlate at only $r = 0.14$.

This gap holds for all nine models (Appendix Figure 6a and Table 2). Correlations with randomized human estimates range from 0.32 to 0.74 for average effects and are similarly high for baseline gaps ($r = 0.66$ for GPT-5.6 Luna and $r = 0.72$ for Claude Sonnet 5), but only from -0.01 to 0.23 for treatment interactions. Models thus capture how groups differ without the intervention far better than they capture how it moves them differently.

Human sampling noise, three quarters of the variance of human Δ estimates but one seventh for average effects, explains only part of this gap. After correction, interaction recovery is weaker for every model but not uniformly negligible, from about zero (GPT-5.6 Luna, Qwen3 4B) to 0.47 (GPT-OSS 20B); the gap excludes zero for four models, including Claude Sonnet 5 (0.80 vs. 0.29, $p = 0.04$; Appendix C.3).

3.2 LLM SIMULATIONS OVERSTATE HETEROGENEITY BY 1.4–2.7×

Prespecified groups cover only a few characteristics. Targeting instead ranks respondents by the model's predicted CATE $\tau^M(x)$, so we also test that ranking directly. Within each contrast, we split profiles into tertiles of $\tau^M(x)$ and compare the average effect in the top and bottom tertiles, estimated both from the simulation and from randomized human outcomes for the same respondents. Because the tertiles are defined without human outcomes, the human estimate is an unbiased test of whether the model's ranking separates stronger from weaker responders.

Figure 1c shows the result for Claude Sonnet 5. The simulated top-minus-bottom gap is large, exceeding one outcome SD in 61 of 530 contrasts, whereas the human gap for the same respondents scatters around zero. Across contrasts, the two correlate at only $r = 0.26$, and the simulated gaps are $2.2\times$ larger in root-mean-square magnitude (Appendix Table 3). Eight of nine models show significant inflation, from $1.4\times$ to $2.7\times$, including the three whose rankings partly track human gaps ($r = 0.39$–0.52), which add no targeting value (Appendix C.2); for GPT-5.6 Luna, the ratio is $1.6\times$ and $r = -0.02$ (Appendix Figure 7). We call this pattern *hallucinated heterogeneity*.

4 LLM SOCIAL SIMULATIONS PROVIDE LIMITED VALUE FOR TARGETING

Section 3.2 showed that a model's top-ranked respondents respond little more than its bottom-ranked ones in human data. Targeting relies on exactly this ranking, so we now measure what hallucinated heterogeneity costs when it is used to decide whom to treat.

Setup. Targeting can only add value where treatment effects genuinely vary, so we focus on the 10 experiments with the strongest statistically significant evidence of predictable heterogeneity in a discovery half of the human data (Appendix D.1). This selection favors targeting: if model rankings add little value here, they are unlikely to add more where heterogeneity is weaker. Within these experiments, we compare each LLM policy, which ranks respondents by $\tau^M(x)$ without using any human outcomes, with a human benchmark policy trained on the discovery half. Every policy ranks the same respondents in the untouched evaluation half (Figure 2a). For a treatment budget q, let

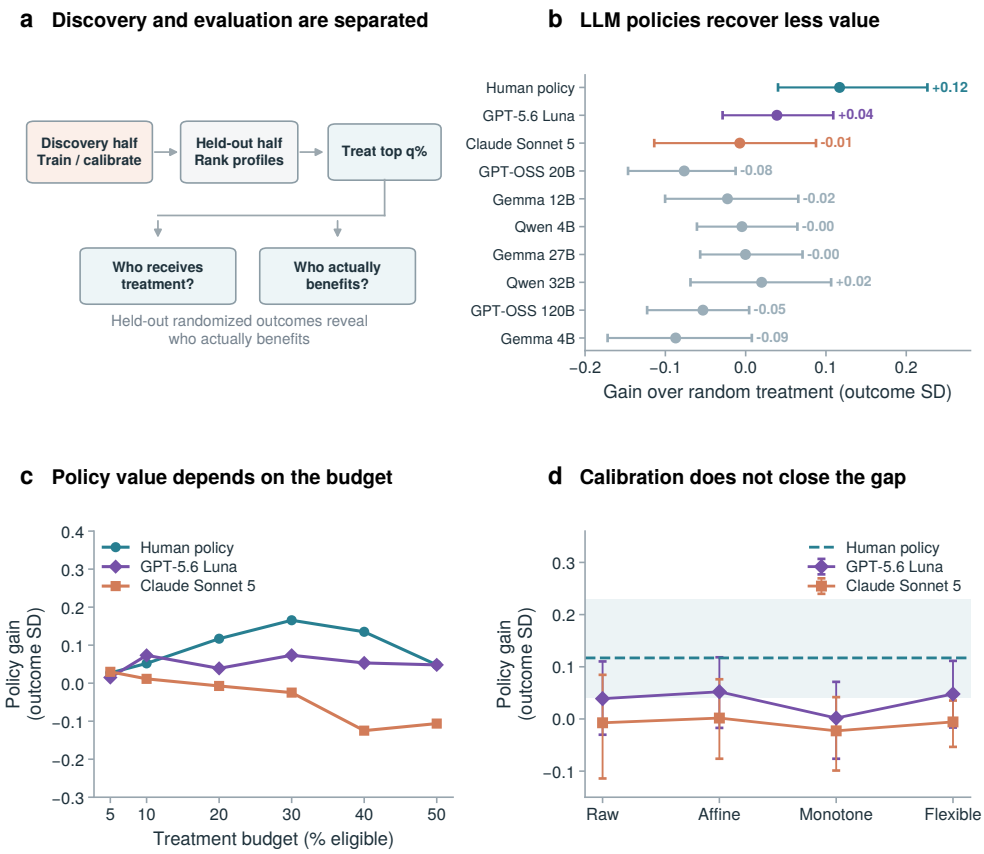

Figure 2: **LLM rankings provide little value for targeting.** **(a)** A discovery half selects experiments, trains the human policy, and fits calibrators; every policy is evaluated on the same held-out half. **(b)** Targeting value at a 20% treatment budget for nine LLMs and a policy trained on randomized human data. **(c)** Targeting value for the human policy, GPT-5.6 Luna, and Claude Sonnet 5 as the budget varies from 5% to 50%. **(d)** Targeting value at a 20% budget before and after calibration of the GPT-5.6 Luna and Claude Sonnet 5 scores.

$\pi_q(X) \in \{0, 1\}$ treat the top fraction q of a policy's ranking. We measure its targeting value by the population average prescriptive effect (PAPE; Imai & Li, 2023),

$$V(q) = \mathbb{E}\big[(\pi_q(X) - q)\{Y(1) - Y(0)\}\big], \tag{3}$$

with outcomes in human outcome SD units. $V(q)$ is the gain from treating the policy's top-ranked respondents rather than a random subset of the same size; it is zero for random allocation. We estimate it using only evaluation-half randomized human outcomes, weighting the 10 experiments equally. Because every policy is scored on the same respondents and outcomes, differences in $V(q)$ reflect only the quality of the rankings.

LLM rankings cluster near random. At a 20% budget, the human policy has the largest targeting value among the human and nine LLM policies (Figure 2b). The LLM policies cluster near zero, and larger or proprietary models do not consistently perform better. The three policies shown keep this ordering across budgets (Figure 2c). The human policy's advantage is largest at intermediate budgets, where the choice of whom to treat matters most. GPT-5.6 Luna remains modestly positive across most budgets, whereas Claude Sonnet 5 falls below random allocation at larger budgets, consistent with its inflated but misdirected heterogeneity in Section 3.2.

Calibration does not close the gap. A natural remedy is to calibrate model scores against human data. We fit affine, isotonic, and flexible calibrators on the discovery half and apply them unchanged to the evaluation half (Figure 2d). Calibration modestly improves GPT-5.6 Luna under some specifications but does not consistently improve Claude Sonnet 5, and every calibrated policy remains below the human policy. Calibration can remap model scores, but it cannot recover information about who responds that the scores do not contain.

5 Why Hallucinated Heterogeneity Arises and Goes Undetected

Next, we examine why hallucinated heterogeneity arises and why validation misses it.

5.1 Errors Cancel Between Arms

For each treatment–outcome contrast, let $\widehat{\mu}_H(z)$ and $\widehat{\mu}_M(z)$ denote the human and simulated mean outcomes in arm $z \in \{0, 1\}$, and let s_H be the human outcome standard deviation. The standardized error in arm z and the resulting error in the average effect are

$$e_z = \frac{\widehat{\mu}_M(z) - \widehat{\mu}_H(z)}{s_H}, \qquad \frac{\widehat{\tau}_M - \widehat{\tau}_H}{s_H} = e_1 - e_0. \tag{4}$$

When the two arm errors are similar, they cancel in the average effect even if the outcome levels themselves are far off.

Two examples illustrate the pattern. In the Bergan experiment (Bergan, 2012), Claude Sonnet 5 concentrates most responses on a few scale points and overestimates both arm means by about 1.3 scale points, yet its average effect of -0.23 nearly matches the human estimate of -0.24 (Figure 3a). In the Harbridge-Yong and Paris experiment (Harbridge-Yong & Paris, 2021), GPT-5.6 Luna places almost all responses on 1 and 4 and underestimates the control and treatment means by 2.0 and 1.9 scale points. Its average effect of -0.26 against the human -0.43 is off by far less than either arm mean (Figure 3b). This concentration is widespread: across the benchmark, models give the same response in all five generations for 12.7% to 77.3% of profile–arm prompts (Appendix Table 6).

The cancellation holds across the benchmark. Arm errors are strongly correlated across the 530 contrasts, with $r = 0.83$ for GPT-5.6 Luna and $r = 0.64$ for Claude Sonnet 5 (Figure 3c). For every model, the RMSE of the average effect is lower than the RMSE of either arm mean (Figure 3d). Across the nine models, the median RMSE is 0.60 human SD for control means, 0.63 for treatment means, and only 0.41 for the average effect. Agreement on average effects therefore says little about whether simulated responses, and hence the differences between them, are correct.

5.2 Predicted Heterogeneity Follows Familiar Group Divides

We next ask what drives the heterogeneity that models produce. Section 3.1 showed that models recover baseline gaps but not treatment interactions. We now break this down by the characteristic defining the groups: political affiliation, ideology, gender, race, age, and education. Comparing treatment effects across levels of such pretreatment characteristics is a standard approach to studying effect modification (Wang et al., 2007; Imai & Ratkovic, 2013).

Recovery of baseline gaps is strongest for political affiliation and ideology, where group labels correspond to familiar differences in opinion. Agreement remains positive but weaker for gender, race, age, and education (Figure 4a). Models therefore encode which groups tend to differ, especially along politically salient dimensions.

Predicted treatment interactions follow the same shape. Across characteristics, models' $|\Delta|$ peaks at political affiliation and ideology and drops sharply for gender, race, age, and education, most steeply for the open-weight models (Figure 4b). The raw human profile looks flatter, but its demographic variation is mostly sampling noise; once noise is removed, human heterogeneity also concentrates in political affiliation and ideology (Appendix C.3). Models thus place heterogeneity along the right divides but exaggerate it there: along political divides, Claude Sonnet 5's interactions are twice the noise-adjusted human magnitude and correlate with human ones at only $r = 0.20$.

Together, these results explain why hallucinated heterogeneity arises and goes undetected. Errors in outcome levels cancel in the average effect, so standard validation passes. And the heterogeneity

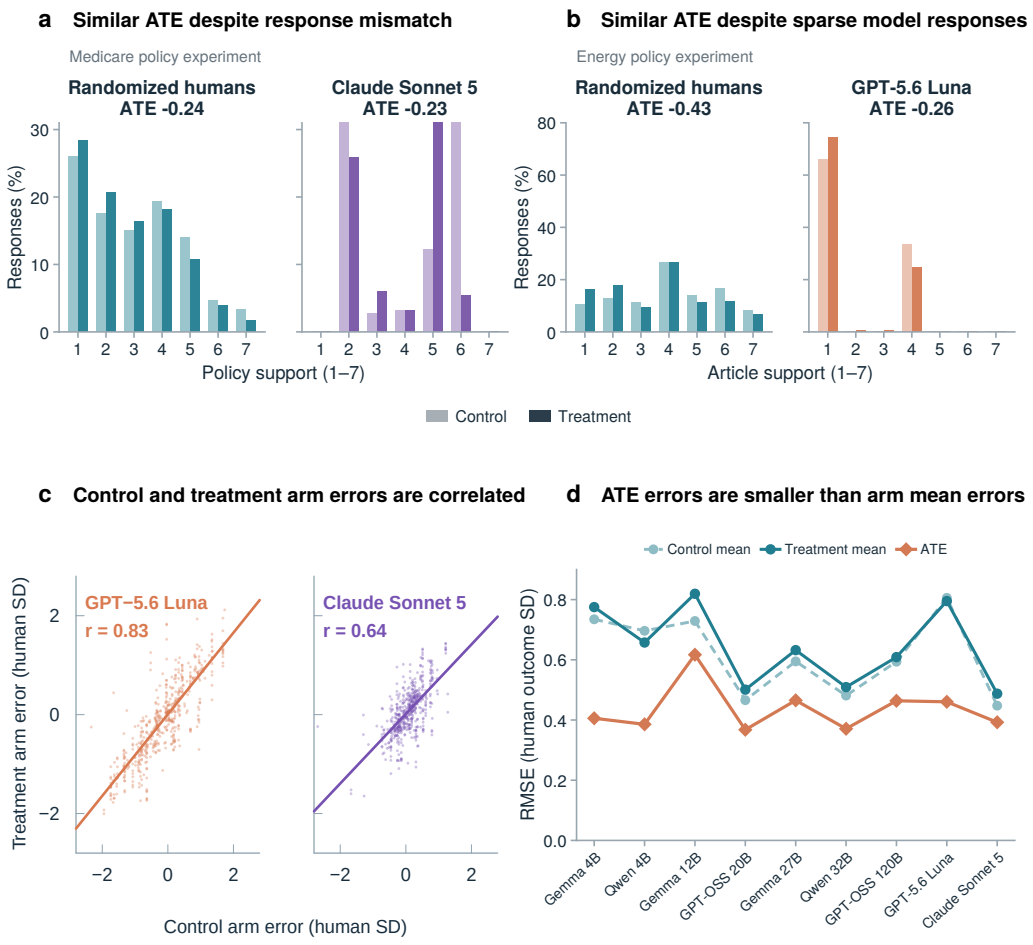

Figure 3: **Accurate average treatment effects can mask substantial response errors.** **(a)** Medicare policy experiment for randomized humans and Claude Sonnet 5. **(b)** Energy policy experiment for randomized humans and GPT-5.6 Luna. **(c)** Correlation of control and treatment arm errors. **(d)** RMSE of the mean outcomes in each arm and of average treatment effects across the same contrasts.

that models do produce looks plausible, because it falls on familiar group divides, especially partisan ones, even though its size and direction there do not match how groups respond.

6 Related Work

LLMs as simulators of human behavior and society. A growing literature uses LLMs to simulate human respondents, behavioral experiments, and synthetic societies (Argyle et al., 2023; Aher et al., 2023; Horton et al., 2023; Park et al., 2024; Piao et al., 2025). Yet simulation fidelity remains an open challenge. Studies of synthetic survey respondents document errors in response distributions and group representation (Bisbee et al., 2024; Dominguez-Olmedo et al., 2024; Boelaert et al., 2025). Research on demographic representation and persona prompting further finds limited persona effects and sensitivity to how identities are prompted (Hu & Collier, 2024; Lutz et al., 2025; Sen et al., 2025), while persona-based simulations can flatten human variation or exaggerate stereotypical differences (Cheng et al., 2023; Wang et al., 2025; Kang et al., 2025). Seemingly successful social simulations may also degrade under more realistic interaction conditions (Zhou et al., 2024). Recent work further documents systematic Sim2Real gaps between simulated and real users (Zhou et al., 2026b) and questions whether emergent behaviors in LLM societies remain robust under stricter validity criteria (Zhou et al., 2025). We extend this concern from behavioral fidelity to

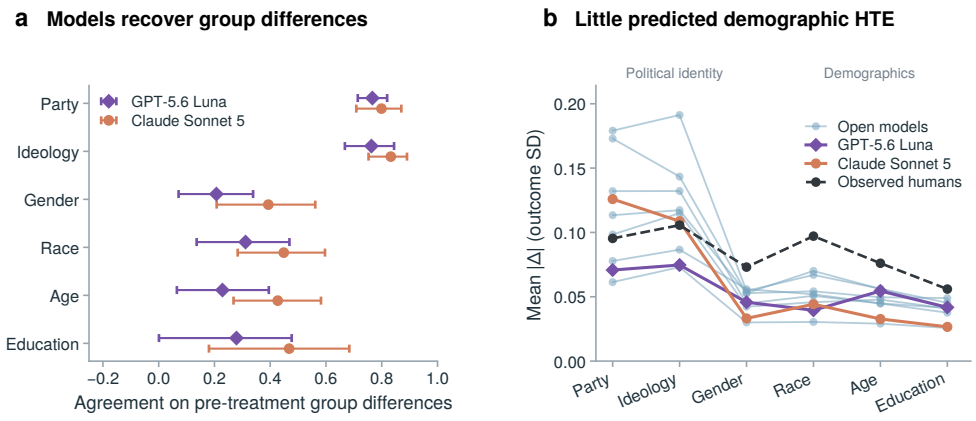

Figure 4: **Baseline structure and predicted heterogeneity vary across group types. (a)** Agreement with randomized human estimates of whether groups differ without the intervention, for GPT-5.6 Luna and Claude Sonnet 5. **(b)** Mean absolute treatment interaction $|\Delta|$ by characteristic for open-weight models, GPT-5.6 Luna, Claude Sonnet 5, and randomized humans. Human values include sampling noise; noise-adjusted magnitudes are in Appendix Figure 8b.

causal fidelity, asking whether LLM simulations preserve heterogeneity in responses to randomized interventions.

LLM simulations for intervention and policy analysis. Recent work asks whether synthetic respondents can predict experimental treatment effects. Most directly, Ashokkumar et al. (2026) show that LLM-simulated experiments can recover average treatment effects across a large collection of social-science studies. Other work combines LLM predictions with human experimental data rather than treating synthetic observations as direct substitutes (Broska et al., 2025; Ruan et al., 2025). These results establish the potential usefulness of LLMs for predicting average intervention effects, but many downstream applications—including targeting, subgroup analysis, and treatment allocation—depend on correctly recovering *who* responds to treatment.

Average and heterogeneous treatment effects. Causal-inference research distinguishes average treatment effects (ATEs) from conditional or heterogeneous treatment effects (HTEs), with extensive methods for estimating treatment-effect heterogeneity and individualized treatment rules (Wager & Athey, 2018; Künzel et al., 2019; Curth & van der Schaar, 2021; Chernozhukov et al., 2025; Imai & Li, 2023). Importantly, accurate ATEs do not imply accurate HTEs: errors in subgroup or individual effects can cancel when averaged over the population. Our work connects this distinction to LLM social simulation by testing whether the aggregate causal fidelity documented in prior work extends to subgroup effects, respondent-level rankings, and treatment-allocation decisions.

7 Discussion

Our results bound the causal fidelity of synthetic respondents: LLMs recover average treatment effects far better than they recover how those effects vary across individuals and groups. The models represent baseline differences between groups more successfully than trial-specific differences in treatment response, and their predicted heterogeneity provides limited value for targeting. Agreement among models and calibration with experimental data do not close this gap. Synthetic respondents therefore require direct validation against randomized human outcomes at the level of the subgroup analysis or decision rule for which they will be used.

Limitations. We evaluate off-the-shelf models with standard persona prompting and no fine-tuning. Our benchmark is also almost entirely American: 130 of the 138 experiments use U.S. samples, so the results may not extend to non-U.S. populations.

Reproducibility statement

All human-side data are public (the Code Ocean capsule of Ashokkumar et al., 2026; the replication archives of Tappin et al., 2023; Klein et al., 2018; Coppock, 2023; and the SocSci210 release of Zhou et al., 2026a). The benchmark and evaluation measures are described in Section 2.2; implementation details appear in Appendix A. Analysis scripts extend the original capsule and will be released with the paper, together with audited prompt files, analysis-ready model outputs, and provenance records for the raw generation archives.

Ethics statement

This work uses only publicly released, de-identified replication data from IRB-approved human-subjects studies; no new human data were collected. Our findings caution against deploying LLM subgroup simulations in microtargeting and policy contexts; we report them to prevent, not enable, harm from stereotype-driven segmentation. Simulated persona outputs analyzed here may encode demographic stereotypes; we report them in aggregate as measurements of model error.

AI use statement

Generative AI was used in two ways. First, AI-assisted tools were used to aid and polish the manuscript's writing; all AI-assisted text was reviewed and revised by the authors. Second, the LLMs evaluated in this work were used to generate the synthetic response datasets analyzed in the paper. The models, prompts, decoding settings, and generation procedure are described in the Appendix. The authors verified the reported results and take responsibility for the final content of this work.

References

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Anthropic. Claude sonnet 5 system card. System card, 2026. URL https://www.anthropic.com/claude-sonnet-5-system-card.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, 656:115–122, 2026. doi: 10.1038/s41586-026-10742-x.

Daniel E. Bergan. The flexible correction model and party labels. Time-sharing Experiments for the Social Sciences, 2012. URL https://www.tessexperiments.org/study/bergan208.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. Machine bias. how do generative language models answer opinion polls? *Sociological Methods & Research*, 54 (3):1156–1196, 2025. doi: 10.1177/00491241251330582.

David Broska, Michael Howes, and Austin van Loon. The mixed subjects design: Treating large language models as potentially informative observations. *Sociological Methods & Research*, 54 (3):1074–1109, 2025. doi: 10.1177/00491241251326865.

Zihan Chen, Di Zhu, and Lei Nico Zheng. When synthetic users fail: A cross-domain benchmark of LLM-simulated human survey responses, 2026. URL https://arxiv.org/abs/2607.26348.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. CoMPoST: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10853–10875, 2023. doi: 10.18653/v1/2023.emnlp-main.669.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Fisher–schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. *Econometrica*, 93(4):1121–1164, 2025. doi: 10.3982/ECTA19303.

Alexander Coppock. *Persuasion in Parallel: How Information Changes Minds about Politics*. University of Chicago Press, 2023. doi: 10.7208/chicago/9780226821832.001.0001.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1810–1818. PMLR, 2021.

Peng Ding. *A First Course in Causal Inference*. Chapman and Hall/CRC, 2024.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 45850–45878, 2024.

Gemma Team. Gemma 3. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Laurel Harbridge-Yong and Celia Paris. You can't always get what you want: How majority-party agenda setting and ignored alternatives shape public attitudes. *Legislative Studies Quarterly*, 46 (2):323–356, 2021. doi: 10.1111/lsq.12279.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from Homo Silicus? Working Paper 31122, National Bureau of Economic Research, 2023.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.554. URL https://aclanthology.org/2024.acl-long.554/.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Dirk Hovy, Nigel Collier, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors, 2025. URL https://arxiv.org/abs/2510.17516.

Kosuke Imai and Michael Lingzhi Li. Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*, 118(541):242–256, 2023. doi: 10.1080/01621459.2021.1923511.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013. doi: 10.1214/12-AOAS593.

Minwoo Kang, Suhong Moon, Seung Hyeong Lee, Ayush Raj, Joseph Suh, David M. Chan, and John Canny. Deep binding of language model virtual personas: a study on approximating political partisan misperceptions. In *Conference on Language Modeling (COLM)*, 2025. doi: 10.48550/arXiv.2504.11673.

Richard A. Klein et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018. doi: 10.1177/2515245918810225.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30096–30111. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1530. URL https://aclanthology.org/2025.emnlp-main.1530/.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Xiaomin Li, Yuexing Hao, Jianheng Hou, Jintao Huang, Qianfeng Wen, Shirley Huang, Yifan Liu, Xiaoyi Liu, Yilan Fan, Yijun Wang, Koutian Wu, Ruoqi Gao, Muhammad Ahmed Mohsin, Jing Tang, Brihi Joshi, Heming Liu, Zheyuan Deng, Zonglin Di, Sankalp Jajee, Jiuyao Lu, Zhiwei Zhang, Saksham Kapoor, Ishan Gupta, Yunhan Zhao, Chanwoo Park, Yucheng Lu, Bing Hu, Weihang Xiao, Aravind Mohan, Hanwen Xing, Runyu Zhang, Mihir Kulshreshtha, Yuanda Xu, Qianyu Zhu, Dianzhuo Wang, Yuxin Xiao, Bowen Jiang, Yongye Su, Wenhao Chai, Zuxin Liu, Lawrence Yunliang Chen, Xuandong Zhao, Ethan Ye, Shivam Patel, Jason Xie, Alex Martin Richmond, Weixiang Ding, Emre Okcular, Diya Mathew, Ziheng Wang, Rana M. Shahroz Khan, Zhejian Peng, Fang Wu, Fan Nie, Xinyang Han, Yubin Kim, Jiawei Zhang, Zhenting Qi, Huangyuan Su, Xu Pan, Abinitha Gourabathina, Hyewon Jeong, Hemanth Neelgund Ramesh, Kumail Alhamoud, Kimia Hamidieh, Zidi Xiong, Samuel Schmidgall, Pengrui Han, Yepeng Huang, Yongheng Wang, Bowen Yang, Alex Gu, Yuchu Wang, Akshay Paruchuri, Brenna Li, Hejie Cui, Jiayuan Ding, Chaosheng Dong, Jiahao Wang, Yixuan He, Chi Wang, Pamela Bhattacharya, Tianyi Peng, Paul Pu Liang, Mitchell Gordon, Yilun Du, Marinka Zitnik, James Zou, Prasanna Tambe, Philip Torr, Emily Fox, Asu Ozdaglar, and Dawn Song. Matraix: Simulating the world with 8.3 billion persona agents, 2026. URL https://arxiv.org/abs/2608.04205.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23212–23237. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1261. URL https://aclanthology.org/2025.findings-emnlp.1261/.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. Model card, 2025. URL https://openai.com/index/gpt-oss-model-card/.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. Model release documentation, 2026. URL https://openai.com/index/gpt-5-6/.

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

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society, 2025.

Qwen Team. Qwen3: Think deeper, act faster. Qwen technical release, 2025. URL https://qwenlm.github.io/blog/qwen3/.

Xinrui Ruan, Xinwei Ma, Yingfei Wang, Waverly Wei, and Jingshen Wang. Can language models boost the power of randomized experiments without statistical bias?, 2025.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, and Markus Strohmaier. Missing the margins: A systematic literature review on the demographic representativeness of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24263–24289. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1246. URL https://aclanthology.org/2025.findings-acl.1246/.

Ben M. Tappin, Chloe Wittenberg, Luke B. Hewitt, Adam J. Berinsky, and David G. Rand. Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences*, 120(25):e2216261120, 2023. doi: 10.1073/pnas.2216261120.

Lars Van Der Laan, Ernesto Ulloa-Pérez, Marco Carone, and Alex Luedtke. Causal isotonic calibration for heterogeneous treatment effects. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34831–34854. PMLR, 2023. URL https://proceedings.mlr.press/v202/van-der-laan23a.html.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7:400–411, 2025. doi: 10.1038/s42256-025-00986-z.

Rui Wang, Stephen W. Lagakos, James H. Ware, David J. Hunter, and Jeffrey M. Drazen. Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194, 2007. doi: 10.1056/NEJMsr077003.

Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. The PIMMUR principles: Ensuring validity in collective behavior of LLM societies, 2025. URL https://arxiv.org/abs/2509.18052.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21692–21714. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1208. URL https://aclanthology.org/2024.emnlp-main.1208/.

Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tongshuang Wu, Yiming Yang, and Maarten Sap. OdysSim: Building foundation models for human behavior simulation. *arXiv preprint arXiv:2606.14199*, 2026a. doi: 10.48550/arXiv.2606.14199.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, and Maarten Sap. Mind the Sim2Real gap in user simulation for agentic tasks. In *Conference on Language Modeling*, 2026b. doi: 10.48550/arXiv.2603.11245. URL https://arxiv.org/abs/2603.11245.

Supplement to

"Hallucinated Heterogeneity: LLM Social Simulations Exaggerate Who Responds to Interventions"

A Benchmark and LLM Simulation

Benchmark. Each subgroup comparison is identified by a unique experiment, outcome, treatment–control contrast, and prespecified pair of moderator levels (e.g., Democrat versus Republican or younger versus older respondents). We estimate treatment effects separately for the two groups and compare their difference, using identical group ordering and outcome scaling in the human and model data.

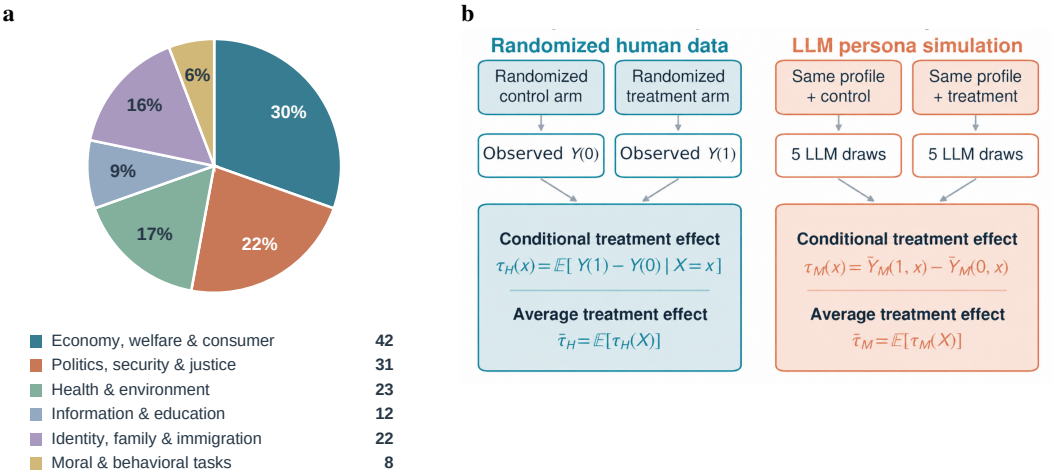

Figure 5: **Benchmark composition and simulation design.** **(a)** One primary topic label per randomized experiment. **(b)** Human experiments observe each respondent under one randomly assigned condition; matched LLM simulations generate responses for the same profile under both conditions.

Models and generation. We evaluate Gemma 3 4B, 12B, and 27B (Gemma Team, 2025); Qwen3 4B and 32B (Qwen Team, 2025); GPT-OSS 20B and 120B (OpenAI, 2025); GPT-5.6 Luna (OpenAI, 2026); and Claude Sonnet 5 (Anthropic, 2026). The seven open-weight checkpoints were run locally, while GPT-5.6 Luna and Claude Sonnet 5 were accessed through their hosted interfaces. Identical rendered profiles are generated once and retain the number of human respondents they represent, and the five generations for a profile–arm pair are averaged before those weights are applied, so repeated generations do not become additional respondents. Open-weight inference uses temperature 1 and top-$p = 1$ with independently advanced sampling streams.

Prompt format. Exact wording varies by source, but every prompt follows the same structure: a respondent profile, the source experiment's treatment or control material, the original outcome question and labeled response scale, and an instruction to return one legal response. Schematically:

> **System:** Simulate the response of the specified survey participant and return only the selected response number.
>
> **User:**
>
> ```
> [Respondent profile]
> [Source treatment or control material]
> [Source outcome question and response options]
> Reply with ONLY the single number you choose.
> ```

The same profile text and outcome request are used across arms; only the randomized experimental material changes. Decoding is restricted to the legal integer choices for each item, and malformed or out-of-range generations are regenerated rather than imputed.

B Model-Level Validation Metrics

B.1 Estimand identities and saturated decomposition

This subsection gives the algebra suppressed in Section 2. In the notation of Table 1, the group effect is $\tau(g) = \mathbb{E}[\delta \mid G = g]$. Let

$$\mu(z, g) = \mathbb{E}\{Y(z) \mid G = g\}$$

denote the mean potential outcome in group g under condition z. Then

$$\tau(g) = \mu(1, g) - \mu(0, g), \qquad B(z) = \mu(z, 1) - \mu(z, 0),$$

and the treatment interaction can be written equivalently as

$$\Delta = \tau(1) - \tau(0) = B(1) - B(0). \tag{5}$$

If $p = \Pr(G = 1)$, the population ATE is the group-share-weighted average $\tau = p\tau(1)+(1-p)\tau(0)$.

For binary z and g, the four group-by-condition means admit the saturated linear decomposition

$$\mu(z, g) = \alpha + \beta g + \gamma z + \eta z g. \tag{6}$$

This is an identity for the four cell means, not a constant-effect assumption. Substitution gives

$$B(0) = \beta, \qquad \tau(0) = \gamma, \qquad \tau(1) = \gamma + \eta, \qquad \Delta = \eta.$$

We apply the same definitions separately to randomized human and model responses; source superscripts are added only when the two estimates must be distinguished.

B.2 Model-level results

Table 2 reports model-level results for four of the validation targets in Table 1. The first three columns are Pearson correlations with randomized human estimates for average effects, baseline gaps, and treatment interactions, for which higher values are better. Outcome-level error is the pooled RMSE of standardized treatment and control arm means, for which lower values are better. Every row uses coverage common to all nine models; baseline correlations therefore use the subset of subgroup comparisons with arm-level estimates for all nine models.

Table 2: **Model-level results for the causal validation framework.** Level RMSE is measured in randomized human outcome SD units.

Model	ATE $r \uparrow$	Baseline $r \uparrow$	Interaction $r \uparrow$	Level RMSE $\downarrow$
Gemma 3 4B	0.323	0.544	0.084	0.755
Gemma 3 12B	0.566	0.644	0.152	0.776
Gemma 3 27B	0.563	0.708	0.196	0.614
Qwen3 4B	0.456	0.385	0.036	0.677
Qwen3 32B	0.637	0.660	0.163	0.495
GPT-OSS 20B	0.640	0.681	0.229	0.484
GPT-OSS 120B	0.624	0.695	0.077	0.601
GPT-5.6 Luna	0.582	0.663	−0.013	0.801
Claude Sonnet 5	0.739	0.720	0.143	0.468

C Additional Analyses

C.1 Treatment-effect rankings defined by GPT-5.6 Luna

The tertiles in Figure 1c are defined by Claude Sonnet 5's own predictions, so the gap they open could reflect that one model rather than LLM simulations in general. Figure 7 repeats the comparison

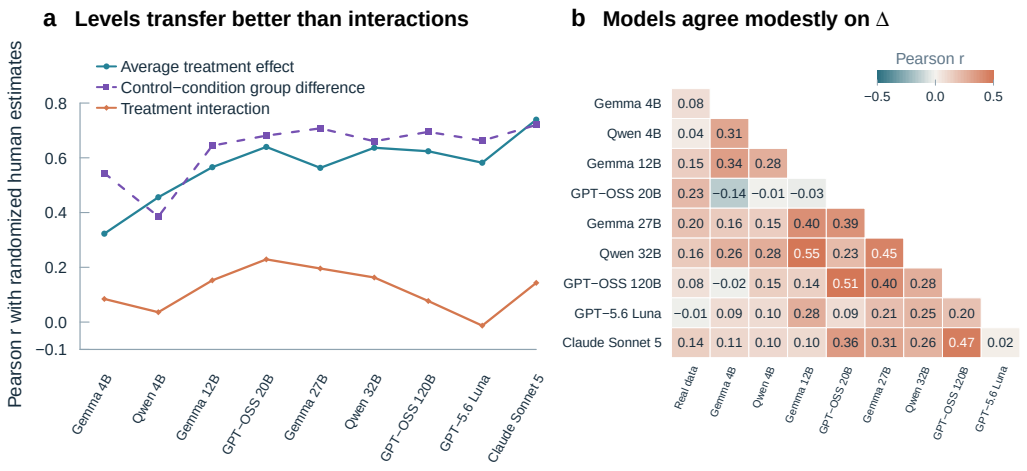

Figure 6: **Average effects and baseline gaps transfer better than treatment interactions. (a)** Correlations with randomized human estimates for average treatment effects, baseline gaps, and treatment interactions across nine models. **(b)** Pairwise correlations among models, and between each model and randomized humans, for the same treatment interactions Δ. Average effect and interaction correlations use all 530 contrasts and 2,906 subgroup comparisons; baseline gaps use the subset of comparisons with arm-level estimates for all nine models.

with tertiles defined by GPT-5.6 Luna. The pattern is the same: across all 530 contrasts the simulated and human gaps correlate at $r = -0.02$, and the simulated gaps are $1.6\times$ larger in root-mean-square magnitude. Which model defines the ranking does not change the conclusion.

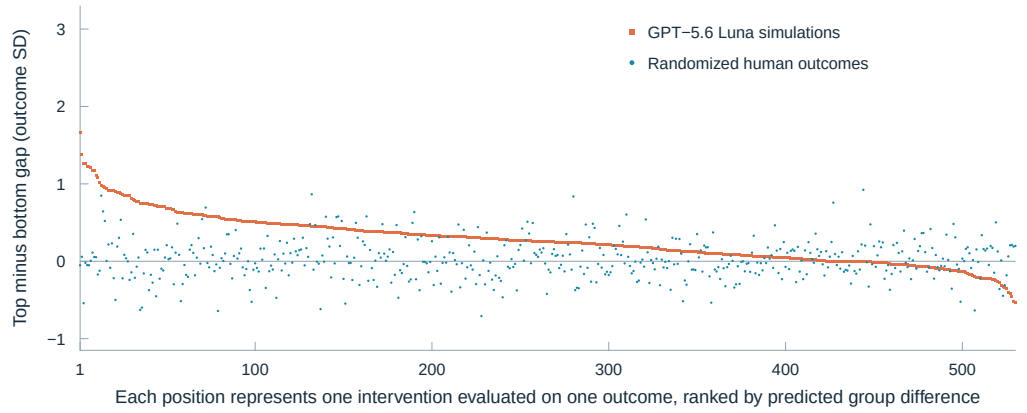

Figure 7: **GPT-5.6 Luna substantially overestimates treatment-effect heterogeneity.** Within every contrast, profiles are divided into empirical tertiles using CATE predicted by GPT-5.6 Luna from four of five generations; the simulated gap is estimated from the held-out generation, averaged over the five choices.

C.2 Heterogeneity inflation

For each model m, profiles are ranked by predicted CATE and divided into empirical tertiles within each contrast k. Let $\widehat{L}_{M,mk}$ be the model's ranked-group gap (Table 1), also known as the GATES contrast between the highest and lowest tertiles (Chernozhukov et al., 2025), and $\widehat{L}_{H,mk}$ the ran-

domized human gap for those same model-defined groups. We summarize their magnitudes by

$$\mathrm{RMS}_{M,m} = \sqrt{K^{-1}\sum_{k=1}^{K} \widehat{L}_{M,mk}^2}, \qquad \mathrm{RMS}_{H,m} = \sqrt{K^{-1}\sum_{k=1}^{K} \widehat{L}_{H,mk}^2},$$

and define the inflation ratio

$$I_{\mathrm{RMS},m} = \frac{\mathrm{RMS}_{M,m}}{\mathrm{RMS}_{H,m}}, \qquad K = 530.$$

Cross-fitting over generations. If the same five generations define the tertiles and estimate $\widehat{L}_{M,mk}$, selecting profiles with extreme estimated CATEs also selects extreme simulation errors, which inflates the model gap even when the simulator itself is less heterogeneous. The human gap is not affected, because the tertiles are defined without human outcomes. We therefore cross-fit over generations. For each held-out generation $g \in \{1, \ldots, 5\}$, the other four generations define the tertiles, generation g alone estimates the model gap, and the human gap is computed in the same fold-specific tertiles. Both gaps are averaged over the five folds. All results in the main text use these cross-fitted gaps.

Table 3 reports both versions. Cross-fitting lowers the ratio most for models whose generations vary most, such as GPT-5.6 Luna (3.83 to 1.64) and Qwen3 4B (4.33 to 1.43), and barely changes it for Claude Sonnet 5 (2.34 to 2.20), which gives the same response in all five generations for 77% of prompts (Table 6). The cross-fitted ratio exceeds one for every model, and its 95% interval excludes one for all but Gemma 3 4B. The remaining biases mostly work against finding inflation: ranking on four rather than five generations slightly weakens the separation between tertiles, and the human RMS still contains human sampling noise (Appendix C.3). The held-out generation's own noise raises the model RMS, which averaging over five folds reduces. Because each model defines different tertiles, the randomized human RMS can differ across rows.

Partial rankers. Three models' cross-fitted rankings partly track human gaps: GPT-OSS 20B ($r = 0.52$), Gemma 3 12B (0.44), and Gemma 3 27B (0.39). Their agreement comes from a minority of contrasts. In the quarter of contrasts where each model predicts its largest gaps, the human gap averages 0.21 SD for GPT-OSS 20B, 0.26 SD for Gemma 3 12B, and 0.09 SD for Gemma 3 27B, against about zero in the other contrasts, and across all contrasts the human gap is positive for only 50–55% of them. These models thus pick out a few interventions with large heterogeneity rather than ranking respondents reliably within interventions. They still overstate the gap by 1.5–2.4×, and none improves targeting: on the ten targeting experiments at a 20% budget, their targeting values are −0.08 SD for GPT-OSS 20B (95% interval −0.15 to −0.01), −0.02 SD for Gemma 3 12B, and 0.00 SD for Gemma 3 27B, against 0.12 SD for the human policy (Figure 2b).

Table 3: **Magnitude inflation in model-defined treatment-effect heterogeneity.** In-sample: the same five generations define the tertiles and estimate the model gap. Cross-fitted: four generations define the tertiles and the held-out generation estimates the model gap, averaged over the five folds. RMS values are in randomized human outcome SD units. Brackets are 95% intervals from 5,000 bootstrap resamples of experiments.

	In-sample		Cross-fitted			
Model	r	I_{RMS}	r	Model RMS	Human RMS	I_{RMS}
Gemma 3 4B	0.027	2.58	0.010	0.351	0.260	1.35 [0.89, 2.02]
Gemma 3 12B	0.389	3.46	0.444	0.881	0.365	2.41 [2.07, 3.18]
Gemma 3 27B	0.277	3.30	0.387	0.784	0.332	2.36 [1.98, 3.04]
Qwen3 4B	0.019	4.33	0.075	0.363	0.254	1.43 [1.21, 1.77]
Qwen3 32B	0.292	3.54	0.339	0.454	0.331	1.37 [1.15, 1.70]
GPT-OSS 20B	0.398	3.06	0.524	0.525	0.340	1.54 [1.36, 1.81]
GPT-OSS 120B	0.099	3.49	0.126	1.051	0.385	2.73 [2.32, 3.54]
GPT-5.6 Luna	0.010	3.83	−0.015	0.413	0.252	1.64 [1.42, 1.96]
Claude Sonnet 5	0.234	2.34	0.264	0.803	0.365	2.20 [1.84, 2.62]

C.3 HUMAN SAMPLING NOISE

Randomization identifies each human interaction, but it does not make its estimate precise. Subgroup comparisons retain group-by-arm cells with as few as 10 respondents, and an interaction combines four estimated cell means, so it is much noisier than an average effect. Noise in the human benchmark lowers every model–human correlation, and it lowers interaction correlations more than average-effect correlations.

Measurement model. For subgroup comparison i, write the human estimate as $\widehat{\Delta}_{H,i} = \Delta_{H,i} + \epsilon_i$, where ϵ_i has mean zero and variance σ_i^2 and is independent of the true interaction and of the model prediction M_i. Then

$$\mathrm{Corr}(M, \widehat{\Delta}_H) = \mathrm{Corr}(M, \Delta_H)\sqrt{\lambda}, \qquad \lambda = \frac{\mathrm{Var}(\Delta_H)}{\mathrm{Var}(\Delta_H) + \mathrm{Var}(\epsilon)},$$

so we estimate the errors-in-variables correlation by $r_{\mathrm{EIV}} = r/\sqrt{\widehat{\lambda}}$, with $\widehat{\lambda} = 1 - \overline{\sigma^2}/\widehat{\mathrm{Var}}(\widehat{\Delta}_H)$. We estimate σ_i^2 separately for every comparison from 100 random splits of respondents into halves. The two half-sample estimates differ with variance $4\sigma_i^2$, so $\widehat{\sigma}_i^2$ is the average of $(\widehat{\Delta}_{H,i}^{(1)} - \widehat{\Delta}_{H,i}^{(2)})^2/4$. Where an analytic respondent-clustered variance is available (257 comparisons), the two agree closely (mean ratio 1.05). Average effects are corrected the same way, using half-sample variances for the Nature experiments and analytic variances elsewhere. Intervals and two-sided p-values come from 5,000 bootstrap resamples of experiments, applied jointly to average effects and interactions, with λ re-estimated in every resample. Model-side simulation noise is not removed, so r_{EIV} remains conservative for the models.

Average effects versus interactions. Sampling noise accounts for 76% of the variance of the human interaction estimates ($\widehat{\lambda} = 0.24$, 95% interval 0.03–0.38) but only 14% for average effects ($\widehat{\lambda} = 0.86$). Table 4 reports both correlations before and after correction. Every observed gap is significant. After correction, every gap remains positive, and the 95% interval excludes zero for Claude Sonnet 5 (0.80 versus 0.29, $p = 0.04$), GPT-5.6 Luna, Qwen3 4B, and Qwen3 32B (Figure 8a). Because $\widehat{\lambda}$ is itself imprecise, the corrected interaction correlations have wide intervals; they bound how far noise can explain the gap rather than pin down the models' true accuracy. Recovery also differs across models. GPT-OSS 20B reaches a corrected interaction correlation of 0.47 (0.17 to 0.80), so interaction recovery is weaker than average-effect recovery for every model but not negligible for all. In 1.7% of resamples the estimated interaction reliability is not positive, so the corrected correlation is undefined; these resamples are omitted.

Table 4: **Average-effect and interaction correlations before and after correcting for human sampling noise.** Gap is the average-effect minus the interaction correlation. Brackets are 95% intervals and p is two-sided, both from 5,000 bootstrap resamples of experiments.

	Observed				Noise-corrected			
Model	ATE r	Δ r	Gap	p	ATE r	Δ r	Gap	p
Gemma 3 4B	0.32	0.08	0.24	0.016	0.35	0.17 [−0.03, 0.55]	0.17 [−0.23, 0.42]	0.293
Gemma 3 12B	0.57	0.15	0.41	< 0.001	0.61	0.31 [0.02, 0.63]	0.30 [−0.01, 0.61]	0.058
Gemma 3 27B	0.56	0.20	0.37	< 0.001	0.61	0.40 [0.03, 0.84]	0.20 [−0.29, 0.60]	0.385
Qwen3 4B	0.46	0.04	0.42	< 0.001	0.49	0.07 [−0.13, 0.29]	0.42 [0.17, 0.68]	0.010
Qwen3 32B	0.64	0.16	0.47	< 0.001	0.69	0.33 [−0.01, 0.65]	0.35 [0.02, 0.70]	0.042
GPT-OSS 20B	0.64	0.23	0.41	< 0.001	0.69	0.47 [0.17, 0.80]	0.22 [−0.14, 0.55]	0.164
GPT-OSS 120B	0.62	0.08	0.55	< 0.001	0.67	0.16 [−0.29, 0.68]	0.51 [−0.04, 0.98]	0.069
GPT-5.6 Luna	0.58	−0.01	0.60	< 0.001	0.63	−0.03 [−0.23, 0.22]	0.65 [0.39, 0.89]	0.002
Claude Sonnet 5	0.74	0.14	0.60	< 0.001	0.80	0.29 [−0.13, 0.76]	0.50 [0.03, 0.93]	0.039

Robustness. Table 5 repeats the comparison under three changes. First, we raise the minimum group-by-arm cell size and recompute the average-effect correlation on the contrasts that still contribute a subgroup comparison, so both correlations describe the same benchmark composition. Even at the baseline threshold, 57 of the 530 contrasts contribute no eligible comparison, so these matched

values differ slightly from Table 4. Second, we weight every experiment equally. Third, we leave out one data source at a time. The corrected gap is positive for eight or nine models in every specification, but the number of models whose interval excludes zero ranges from one to six.

Composition matters most for GPT-5.6 Luna. Its matched average-effect correlation is 0.51 at the baseline threshold, against 0.63 on all 530 contrasts, and falls to 0.34 at the highest threshold. Its gap nevertheless remains positive at every threshold and excludes zero at all but the highest. Claude Sonnet 5's matched average-effect correlation stays near 0.78. Its gap is not significant at minimum cells of 30 and 50, where its corrected interaction correlation rises to 0.56–0.59, and it is significant again at 100 and 200. Equal-experiment weighting lowers the human reliability to 0.09, leaving corrected interaction correlations near zero for both hosted models. Without the Nature experiments, reliability falls to 0.04 and 38% of resamples are undefined, so the correction is uninformative there. Dropping any other source leaves the results close to the main analysis. Heterogeneity recovery is therefore consistently weaker than average-effect recovery, but the size of the corrected gap depends on which comparisons are included.

Table 5: **Robustness of the noise-corrected comparison.** ATE and Δ are correlations corrected for errors in variables; "Gap" is their difference with a 95% interval from 5,000 bootstrap resamples of experiments. In the upper panel, average effects use only the contrasts that retain at least one subgroup comparison. $\widehat{\lambda}$: share of human interaction variance that is signal. Last column: models (of nine) with a positive gap / with an interval excluding zero.

				Claude Sonnet 5			GPT-5.6 Luna			
Specification	Contr.	Comp.	$\widehat{\lambda}$	ATE	Δ	Gap	ATE	Δ	Gap	Models
Minimum group-by-arm cell size (matched contrasts)										
10	473	2,906	0.24	0.79	0.29	0.50 [0.01, 0.93]	0.51	−0.03	0.54 [0.26, 0.81]	9 / 3
20	452	2,510	0.25	0.79	0.29	0.50 [0.04, 1.08]	0.51	0.05	0.46 [0.23, 0.72]	9 / 3
30	416	2,087	0.38	0.77	0.59	0.18 [−0.08, 0.78]	0.49	0.13	0.37 [0.15, 0.61]	8 / 2
50	353	1,590	0.33	0.78	0.56	0.23 [−0.06, 0.82]	0.49	0.11	0.39 [0.17, 0.64]	9 / 1
100	208	806	0.35	0.78	0.26	0.52 [0.18, 0.88]	0.40	0.07	0.33 [0.14, 0.58]	9 / 5
200	90	253	0.48	0.80	0.09	0.70 [0.16, 1.15]	0.34	0.05	0.29 [−0.02, 0.64]	9 / 6
Weighting and sources (minimum cell 10)										
Equal experiment weights	530	2,906	0.09	0.80	0.01	0.79 [0.41, 1.25]	0.67	−0.08	0.74 [0.40, 1.21]	8 / 2
Without Nature experiments	85	318	0.04	0.76	−0.09	0.85 [0.11, 1.70]	0.57	−0.27	0.83 [0.17, 1.80]	8 / 2
Without Political advertisements	482	2,670	0.42	0.80	0.37	0.43 [0.03, 0.91]	0.60	0.02	0.57 [0.29, 0.79]	9 / 5
Without SocSci210	511	2,845	0.24	0.80	0.30	0.50 [0.02, 0.96]	0.67	−0.02	0.69 [0.43, 0.93]	9 / 3
Without Persuasion in Parallel	520	2,896	0.24	0.79	0.30	0.48 [0.01, 0.92]	0.64	−0.03	0.67 [0.39, 0.92]	9 / 3
Without Many Labs 2	522	2,895	0.23	0.78	0.30	0.49 [0.00, 0.93]	0.57	−0.02	0.60 [0.32, 0.86]	9 / 3

Magnitude of human heterogeneity. Sampling noise also inflates absolute human interactions: under a normal approximation, even if every true interaction were zero, $\mathbb{E}|\widehat{\Delta}_{H,i}| = \sigma_i\sqrt{2/\pi}$. For demographic characteristics the observed human mean $|\widehat{\Delta}|$ of 0.075 SD is almost exactly the 0.071 SD expected from noise alone. Because $\mathbb{E}[\widehat{\Delta}_H^2] - \mathbb{E}[\sigma^2] = \mathbb{E}[\Delta_H^2]$, we report noise-adjusted magnitudes on the root-mean-square scale (Figure 8b). After adjustment, signal accounts for 43% of the mean-square human interaction for political affiliation and ideology but only 4% for gender, race, age, and education, giving noise-adjusted magnitudes of 0.11 and 0.02 SD. Both intervals are wide, and the political interval reaches zero. Human heterogeneity is thus concentrated along the same political divides as model heterogeneity, where models overstate its size: Claude Sonnet 5's root-mean-square interaction there is 0.22 SD. Within political divides, its interactions correlate with human estimates at $r = 0.20$ ($r_{\mathrm{EIV}} = 0.31$, 95% interval -0.16 to 0.89).

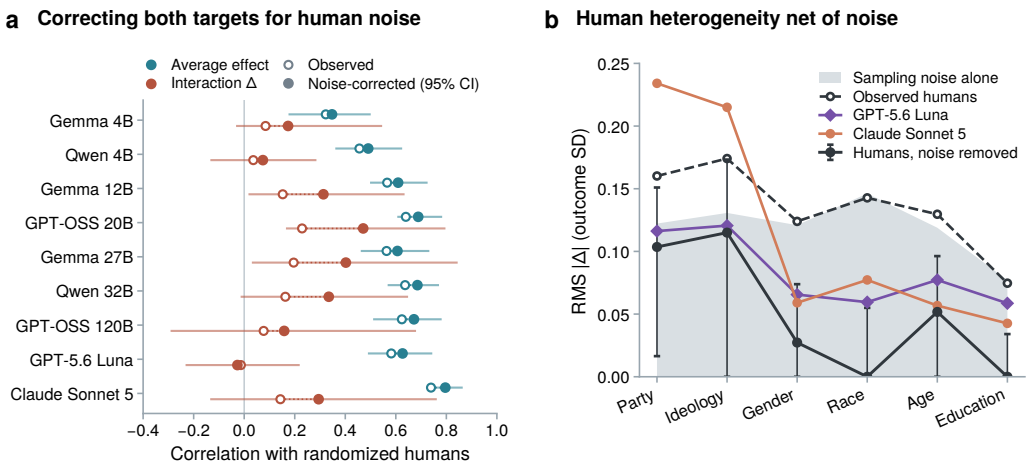

Figure 8: **Human sampling noise and treatment effect heterogeneity. (a)** Correlations with randomized human estimates before (open) and after (filled) correcting for human sampling noise, with 95% intervals for the corrected values. **(b)** Root-mean-square interaction by characteristic for randomized humans before and after removing sampling noise, the level produced by sampling noise alone, and the two hosted models. Human intervals are 95% bootstrap intervals.

C.4 Simulation repeatability

Each profile–arm prompt is simulated five times under the same prompt and decoding configuration. To assess simulation repeatability, we classify a prompt as exactly repeated only when all five parsed numeric responses agree. Table 6 reports this rate over the common set of 149,209 prompts.

Table 6: **Exact agreement across five repeated simulation draws.** Entries are the percentage of profile–arm prompts for which all five numeric responses are identical; all models are evaluated on the same prompt set.

Model	Five identical responses (%)
Gemma 3 4B	32.8
Gemma 3 12B	23.8
Gemma 3 27B	24.1
Qwen3 4B	19.7
Qwen3 32B	12.7
GPT-OSS 20B	18.7
GPT-OSS 120B	38.5
GPT-5.6 Luna	51.5
Claude Sonnet 5	77.3

Exact repetition varies substantially across models. Claude Sonnet 5 gives the same response in all five generations for 77.3% of prompts, compared with 51.5% for GPT-5.6 Luna and 12.7% for Qwen3 32B. Because the criterion requires all five responses to match, it is a stringent measure of concentration rather than a measure of predictive accuracy.

D Targeting Analysis

D.1 Selection of experiments

Using randomized human data alone, we follow generic machine-learning inference for HTE in randomized experiments (Chernozhukov et al., 2025). Before any model fitting, respondents are deterministically assigned in approximately equal proportions to a discovery half and an evaluation

half. The selection universe contains all 530 contrasts in 138 experiments. For each contrast k, an ExtraTrees transformed-outcome learner uses all model-visible pretreatment characteristics (80 trees, maximum depth 6, minimum leaf size 20, and 80% of features per split). Within the discovery half, five-fold respondent-level cross-fitting gives every respondent a CATE score from a learner not trained on that respondent's outcome. The evaluation half is not used at this stage.

Let S_{ik} denote the centered, unit-standard-deviation out-of-fold CATE score and Γ_{ik} the Horvitz–Thompson transformed randomized outcome. We estimate the continuous BLP

$$\Gamma_{ik} = \alpha_k + \beta_k S_{ik} + \varepsilon_{ik}$$

using the discovery-half out-of-fold predictions and compute an HC1 standard error and the one-sided p-value for $H_0 : \beta_k \leq 0$. When experiment e contains K_e contrasts, we take its smallest contrast-level p-value and form the within-experiment Bonferroni value

$$p_e = \min\left\{1, K_e \min_{k \in e} p_k\right\}.$$

Experiments are ranked by p_e, and the contrast attaining the minimum is retained from each experiment. The 10 highest-ranked experiments all satisfy $p_e < 0.05$ in the discovery half (Table 7).

Table 7: **Experiments selected for the targeting analysis.** Each row is one experiment, ranked by p_e in the discovery half. BLP slopes are in randomized human outcome SD units per one SD of the cross-fitted CATE score. p_e adjusts for the number of contrasts within the experiment.

Rank	BLP slope	HC1 SE	p_e
1	1.206	0.240	8.1×10^{-7}
2	0.839	0.181	1.7×10^{-5}
3	1.122	0.343	0.0050
4	2.051	0.798	0.0058
5	0.963	0.305	0.0058
6	0.679	0.246	0.0088
7	1.526	0.438	0.0088
8	1.701	0.750	0.0129
9	0.141	0.070	0.0220
10	0.162	0.081	0.0233

D.2 Policy evaluation and calibration

For Figure 2, each experiment receives equal weight. After selection, the human CATE learner is refit on the entire discovery half and used to score only evaluation-half respondents. Each LLM policy scores the same evaluation respondents without using human outcomes. Within that common held-out cohort, each policy treats the top q fraction of its ranking, and targeting value is computed only from evaluation-half randomized outcomes.

We calibrate the GPT-5.6 Luna and Claude Sonnet 5 scores separately within each selected contrast, treating the discovery-half Horvitz–Thompson transformed outcome Γ_{ik} as the observed pseudo-outcome for the latent treatment effect. This follows the regression view of causal calibration, in which a raw CATE score is mapped to the average randomized treatment effect among respondents receiving that score (Van Der Laan et al., 2023). The affine calibrator regresses Γ_{ik} on an intercept and the raw LLM score by ordinary least squares. The isotonic calibrator minimizes squared error subject to the fitted map being nondecreasing, yielding a monotone step function. The flexible calibrator fits a one-dimensional ExtraTrees regression (Geurts et al., 2006) with 100 trees, maximum depth 4, and minimum leaf size 20, without imposing monotonicity. Each fitted map is applied without refitting to evaluation-half LLM scores.