Can Finetuing LLMs on Small Human Samples Increase Heterogeneity, Alignment, and Belief-Action Coherence?

Anonymous ACL submission

January 5, 2026

Abstract

There is ongoing debate about whether large language models (LLMs) can serve as substitutes for human participants in survey and experimental research. While recent work in fields such as marketing and psychology has explored the potential of LLM-based simulation, a growing body of evidence cautions against this practice: LLMs often fail to align with real human behavior, exhibiting limited diversity, systematic misalignment for minority subgroups, insufficient within-group variance, and discrepancies between stated beliefs and actions. This study examines an important and distinct question in this domain: whether finetuning on a small subset of human survey data, such as that obtainable from a pilot study, can mitigate these issues and yield realistic simulated outcomes. Using a behavioral experiment on information disclosure, we compare human and LLM-generated responses across multiple dimensions, including distributional divergence, subgroup alignment, belief–action coherence, and the recovery of regression coefficients. We find that finetuning on small human samples substantially improves heterogeneity, alignment, and belief–action coherence relative to the base model. However, even the best-performing finetuned models fail to reproduce the regression coefficients of the original study, suggesting that LLM-generated data remain unsuitable for replacing human participants in formal inferential analyses.

1 Introduction

With the rapid advancement of reasoning and language capabilities in large language models (LLMs), researchers have increasingly explored their use as human surrogates to simulate survey and experimental responses. One of the earliest works in this direction is by Argyle et al. (2023), who used GPT-3 to model public opinion patterns in political science. Since then, the use of LLMs as proxies for human participants has expanded across a range of disciplines, including marketing, economics, and psychology (e.g. Brand et al., 2023; Aher et al., 2023), as well as in natural language processing (NLP) tasks traditionally performed by humans, such as text quality evaluation and data annotation (Chiang and Lee, 2023; He et al., 2024). In parallel, another line of work has focused on finetuning open-source LLMs with human experimental data to better align model predictions with real behavioral distributions. For example, Centaur (Binz et al., 2025) is an LLM finetuned on a large-scale psychology dataset to more accurately mimic differences in human behavior, while SubPOP (Suh et al., 2025) leverages real survey data to model subpopulation opinion distributions. Together, these developments have spurred growing interest in whether LLMs can serve not only as reasoning systems but also as credible simulators of human heterogeneity and social behavior.

On the other hand, a growing body of research urges caution in using LLMs as substitutes for human participants. The central critique is that LLM-generated responses lack heterogeneity, tending to reproduce human averages while underrepresenting the diversity found in real populations (Bisbee et al., 2024; Wang et al., 2025a; Wu et al., 2025). This issue is most pronounced under default prompting conditions, where LLMs do not behave like an "average" respondent but instead reflect the perspectives of the WEIRD demographic (Western, Educated, Industrialized, Rich, and Democratic) that dominates their training data (Liu et al., 2022). To address this limitation, researchers have experimented with persona and context-based prompting, embedding demographic cues such as gender, race, or age to encourage more varied responses. While these methods introduce some between-group differences, they remain imperfect. Studies have also documented the phenomenon of identity flattening, where LLMs exhibit variation across demographic subgroups but not within them, portraying each subgroup as internally homogeneous (Lin, 2025; Wang et al., 2025a). This behavior often

mirrors out-group stereotyping, where the model reproduces what outsiders might expect a given group to believe, rather than authentic in-group diversity, reflecting biases inherent in online text corpora. Even finetuned models designed specifically for psychological realism face similar challenges. For example, Centaur (Binz et al., 2025) has been found to deviate systematically from human ratings of moral judgment (Schröder et al., 2025) and behavioral study data (Gao et al., 2025), suggesting limited generalizability to out-of-distribution behavioral contexts.

The aim of this study is not to enter the discourse of whether off-the-shelf or pre-finetuned LLMs can fully substitute for human respondents, but rather to examine whether finetuning on a small subset of human data, such as that obtainable from a pilot study, can mitigate the limitations documented in prior work. For reasons detailed below, we base our simulations on the behavioral experiment of Hunt et al. (2025), which investigates how information disclosure shapes attackers' judgments and decisions in a security scenario. Our findings show that LLMs finetuned on small fractions of the human dataset can substantially improve heterogeneity, distributional alignment, and value–action coherence relative to the base LLM. However, despite these improvements, none of the finetuned models reliably reproduce the original regression coefficients or underlying hypothesis tests, indicating that replacing human participants with LLMs remains risky and may introduce non-negligible bias into conclusions. Hence, we caution against the naive use of base models for experimental design and simulation, even when augmented with sociodemographic prompting, they tend to produce overly tight variance. Such compression can yield overconfident power analyses and lead to underpowered studies. In contrast, we show that finetuning on as few as 30 human observations substantially improves distributional alignment, producing variation that more closely resembles human data and yielding more realistic simulated responses. This makes finetuned models better suited for certain stages of survey design and pretesting than base LLMs. In sum, we contribute to the literature by showing that finetuning on a small sample of human data (30 observations) can substantially mitigate several issues identified in prior work, including the lack of heterogeneity, distributional misalignment, and the value–action gap. Both the original and simulated data will be released for future research upon acceptance.

2 Related Literature

2.1 Limitations of LLMs in Replicating Human Behavior

A growing body of research shows that LLMs often fail to capture the full heterogeneity of human responses when used to simulate survey participants (Wang et al., 2025a; Hewitt et al., 2024; Dominguez-Olmedo et al., 2024; Wu et al., 2025; Bisbee et al., 2024; Giorgi et al., 2024). Bisbee et al. (2024) examine this limitation within political science using simulations based on the American National Election Study. They find that while LLMs can accurately reproduce mean responses, they are systematically biased, overconfident, and display artificially low variance. Wang et al. (2025a) extend this line of research across multiple policy domains, such as health care, gun regulation, immigration, and criminal justice, using prompts that represent sixteen demographic identities. Their results reveal a pronounced lack of within-group variance: responses generated for each demographic subgroup are strikingly homogeneous. The authors describe this as identity flattening, whereby the model constructs one-dimensional, stereotyped "personalities" for each demographic. Gao et al. (2025) examine the alignment between human participants and LLM-simulated responses through controlled experiments. They report that this lack of heterogeneity and behavioral misalignment persists even after applying various mitigation strategies, including few-shot prompting, role or persona assignment, multilingual prompting, and retrieval-augmented generation (RAG) (Lewis et al., 2020). Shen et al. (2025) draw on the concept of the value–action gap from the social science literature (Blake, 1999; Godin et al., 2005; Chung and Leung, 2007), which refers to the discrepancy between individuals' stated values and their actual behaviors. They apply this framework to examine the gap between an LLM's expressed inclinations and its subsequent decisions. Collectively, these studies demonstrate that while LLMs can replicate average human judgments with reasonable accuracy, they systematically underrepresent the variance and contextual diversity that characterize real human populations.

2.2 Attempts to Enhance the Behavioral Realism of LLMs

Aiming to generate more diverse and contextually grounded outputs, researchers have increasingly incorporated persona or role-based prompting when simulating human responses (Argyle et al., 2023; Dillion et al., 2023). On standard reasoning and coding benchmarks, persona prompts (e.g., "act as a math teacher/engineer") and more advanced steering frameworks have shown modest but consistent improvements in performance, suggesting that explicit role framing can increase reasoning and improve alignment with task goals (Wang et al., 2025b; Kim et al., 2024; Kong et al., 2024). However, when applied to survey or opinion-elicitation settings, often referred to as sociodemographic prompting, these techniques produce mixed or negative results. Studies find that large language models tend to misrepresent marginalized groups and align more closely with white or majority identities than with minority participants, generating stereotyped or homogenized responses that fail to capture genuine within-group heterogeneity (Lutz et al., 2025; Sun et al., 2025; Wang et al., 2025a).

2.3 Gaps in Existing Work

Despite substantial work documenting the limitations of LLMs as simulated survey respondents, very little research has explored whether these issues can be mitigated through targeted finetuning on small samples of human data. Existing studies typically evaluate base models or rely on prompt engineering, or persona conditioning, but do not examine whether a small, realistic amount of human data such as that obtainable from a pilot study can meaningfully improve heterogeneity, alignment, and behavioral coherence. This gap leaves open the critical question of whether modest finetuning can address the flaws identified in earlier research, or whether these limitations are inherent to current LLM architectures.

3 Data

This study builds on a large behavioral experiment by Hunt et al. (2025), which examined the effects of information disclosure on attackers' judgments and decisions in a security setting. We selected this experiment for use in the current research for three main reasons. First, the design of the experiment, which elicited both beliefs and decisions from the participants, enables us to assess the coherence between LLMs' stated beliefs and their subsequent actions. Second, we obtained the original data directly from the authors of the experiment, with confirmation that the data had not yet been publicly released or fed into any LLM prior to the analysis presented herein. Ensuring that the model had not previously seen the dataset is critical, as prior work demonstrates that LLMs can memorize training data, including personally identifiable information, and reproduce it verbatim with only one related document in the training corpus (Zhou et al., 2024; Carlini et al., 2021, 2022). Such data contamination undermines the validity of LLM evaluations by producing spuriously high or unreliable performance estimates (Sainz et al., 2023; Li et al., 2024; Ranaldi et al., 2024). Third, the experiments in Hunt et al. (2025) focused on information disclosure in a high-stakes, strategic environment (i.e., security and defense), a domain that has not been studied in the context of LLMs and behavioral realism.

The sample we analyze consists of behavioral data from 929 U.S. adults [1] who played the role of attackers deciding whether to attack either of two targets that were defended by a centralized defender. In the experiment, the defender released one of four disclosures related to the deployment of new security measures to protect her assets at the targets. Participants (i.e., attackers) were also informed of the defender's relative valuations of both targets, where the differences between her valuations of each target could be large or small. Appendix B outlines the manipulations in more detail. This 2 x 4 design created a total of eight treatment groups (i.e., two levels of the valuation factor, and four levels of the disclosure factor). After receiving the experimental manipulations, participants then reported their beliefs regarding where the technology was deployed – expressed as probabilities that the new security measures were deployed at Target 1, Target 2, or both, and then made a final decision on whether to attack Target 1, attack Target 2, or not attack. Participants were incentivized with base pay for completing the study and potential bonus payments based on the outcomes of their attacks. The full experimental script can be found in Appendix A.

[1] The original dataset in Hunt et al. (2025) had 975 responses, but we exclude data points with missing demographic information for both finetuning and inference.

4 Methods

Our goal is to explore how well various LLM-based approaches can predict various results derived from this experimental dataset, and in particular how an approach finetuned on a subset of this data (e.g. that would come from a pilot study) can improve these predictions. We consider three different approaches, described below. All simulations were conducted using GPT-4.1 (Achiam et al., 2023) with the default temperature setting of 1.0. The finetuned models use the same base architecture and temperature configuration to ensure comparability. We mimic most prior work discussed above and first ran simulations using an unfinetuned GPT model. We incorporated participant demographics as role prompts to approximate more realistic responses. For finetuning, we used two sampling strategies and three dataset sizes, resulting in six models in total. The first strategy uses random sampling, and the second strategy, which we refer to as balanced sampling, draws equal proportions [2] of responses from each ethnic subgroup in order to address the systematic misalignment for minority groups noted in prior research. The three dataset sizes correspond to the number of treatment groups included in finetuning. This design allows us to test whether finetuning on only a subset of treatment groups enables the model to generalize to the remaining groups. Specifically, we sample 25% of the data from the first treatment group (1a1b), from the first four treatment groups, and from all eight treatment groups. For the multi group settings, we draw 25% from each included group. We name each model using its sampling strategy and dataset size. For example, the model finetuned on 25% of the first treatment group using random sampling is labeled Random (1a1b), with parallel naming for the remaining random and balanced models.

As outcomes, we examine how finetuning affects heterogeneity with just 25 % of the data from a single treatment group (the 1a1b treatment group; see Appendix B). To quantify distributional alignment, we use the Jensen–Shannon (JS) distance (Lin, 2002), a symmetric and bounded measure of distributional divergence that has also been applied by Gao et al. (2025) to compare LLM and human response distributions. For belief–action gap, we measure the conditional proportion of LLMs choosing to attack the target that they think has the highest probability that security is in place and compare it to humans. Lastly, we compute the accuracy of regression coefficient estimates relative to the regression analysis performed on the human data.

[2] We use a round-robin sampling strategy in which we randomly draw one sample from each ethnicity group in sequence until reaching 25% of the dataset. For ethnicity categories with smaller sample sizes (e.g., Other), we include all available observations and then continue the round-robin procedure with the remaining groups to reach the target proportion.

5 Results

In the following subsections, we examine how finetuning affects heterogeneity, distributional alignment, and the belief–action gap, as well as the accuracy of regression coefficient estimates relative to the regression analysis performed on the human data.

5.1 Heterogeneity

We find that even a small number of finetuned samples can drastically improve diversity and mitigate homogeneity issues relative to the unfinetuned baseline. Figure 1 shows the distribution of beliefs regarding the likelihood that the defender's security measures were deployed at Target 1 and both Targets. The analysis of Target 2's beliefs and the results from all models are presented in Appendix C, and support findings discussed here. In analyzing Target 1 (left panel) in Figure 1, the unfinetuned baseline exhibits a highly concentrated distribution, with most responses clustered in the 30–40 % and 40–50 % bins. In contrast, human responses follow a more bell-shaped pattern, with probability mass spread across multiple bins. This supports prior findings that base LLMs lack response heterogeneity. However, when we finetune the model on 25 % of the first treatment group using the random sampling strategy, the distribution becomes substantially more dispersed, showing probability mass across all bins and modest peaks around 50–70 %, closely resembling human patterns.

As shown in Figure 1 (right panel), the human distribution for the belief that technology is deployed at both targets is highly right-skewed, with most probability mass clustered near 0–10%. The finetuned models exhibit similar patterns as the human data and capture heterogeneity, but similar to the analysis of Target 1, the unfinetuned baseline suffers from homogeneity, with the probability mass across all simulation in the 30–50% bins. This suggests that the unfinetuned baseline does not place proper weight on the contextual variables provided in the study (e.g., valuations, disclosure

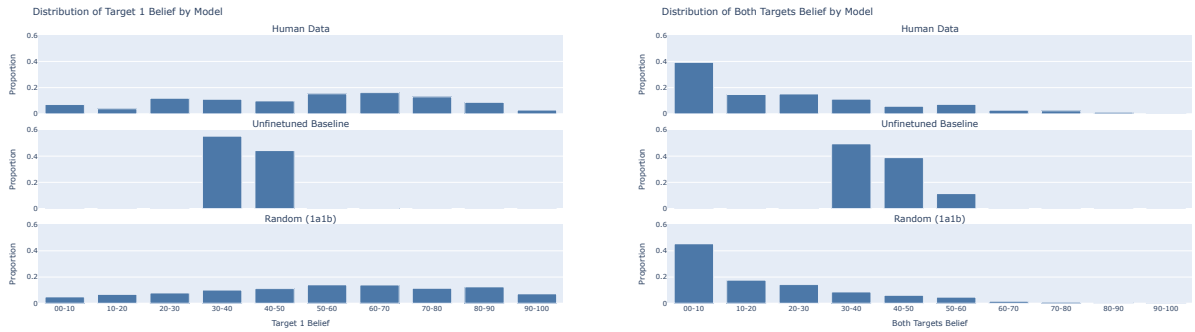

Figure 1: Distributions of beliefs that security is deployed at Target 1 (left) and at both targets (right) across human data and LLM simulation conditions. Each panel shows the empirical belief distribution for human participants alongside models finetuned under different sampling strategies.

Table 1: Number of unique belief structures across human data and simulation conditions.

Model	Num. Unique Beliefs
Human Data	340
Balanced (1a1b)	244
Random (1a1b)	374
Balanced (All Groups)	338
Balanced (4 Groups)	278
Random (All Groups)	308
Random (4 Groups)	291
Unfinetuned Baseline	19

type) and instead relies on a strong, low-variance prior across the three belief options. To further investigate this, instead of analyzing only one belief at a time, we also consider the entire belief *structure*, given that each participant in the experiment, and each LLM simulation, recorded these three beliefs as part of a single response. When analyzing the belief structures elicited by humans and LLM, we find that the unfinetuned LLM suffers from extreme homogeneity. As shown in Table 1, the unfinetuned model produces only 19 unique belief structure across all 929 simulations, compared to 340 in the human data. All finetuned models produce at least 200 unique combinations, indicating that even limited finetuning substantially increases the diversity of belief patterns expressed by the model.

5.2 Alignment

Given that finetuning on data subsets increases response heterogeneity, we next examine whether this added variation reflects meaningful alignment with human data or merely introduces random noise. Our results show that all finetuned models substantially reduce the Jensen–Shannon (JS) distance between LLM-simulated data and human responses. Figure 2 reports the JS distance by ethnicity and model. The ethnicity categories correspond to the default classifications used by Prolific, the participant recruitment platform through which the original data were collected. As expected, the unfinetuned baseline exhibits the largest JS distances, consistent with the patterns observed in the previous section. Across all finetuned variants, the JS distance decreases by at least half, indicating improved alignment with human response distributions.

However, the impact of balanced sampling is mixed. Surprisingly, incorporating more demographically diverse data does not necessarily reduce JS distance for minority groups, even though such sampling increases representation. Notably, the lowest JS distance is consistently observed in the White participant category, even for the balanced model, which aligns with prior findings that LLMs most closely reflect the WEIRD population. This pattern emphasizes the persistence of demographic bias in model alignment and highlights the need for more sophisticated sampling or finetuning strategies to achieve equitable representational fidelity. Indeed, our ablation study (see Appendix E) shows that adding demographic information does not meaningfully increase alignment in our context.

Overall, since misalignment decreases across every subgroup, we infer that finetuning enhances overall alignment between LLM and human responses. Consistent reductions in JS distance at the group level, as presented in Appendix D, further corroborate this improvement.

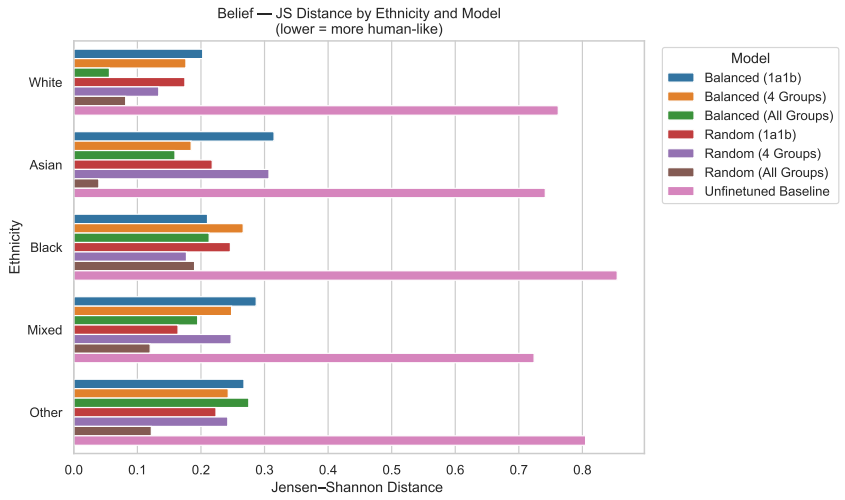

Figure 2: Jensen–Shannon distance between LLM-simulated and human belief distributions, disaggregated by ethnicity and finetuning strategy.

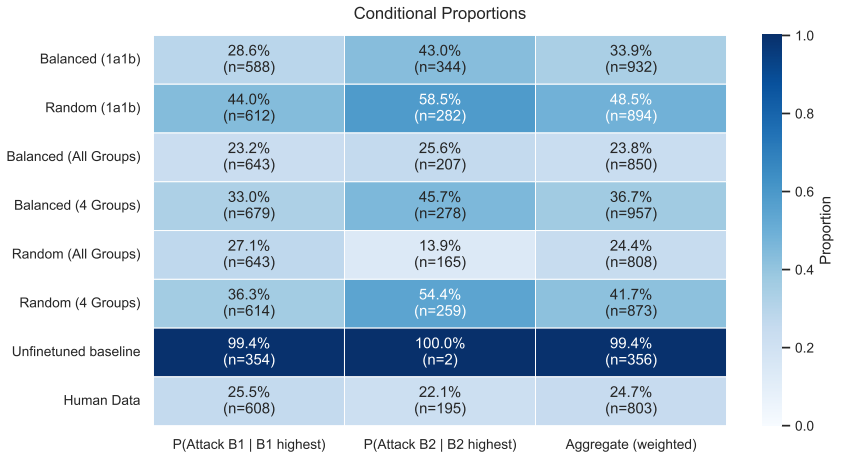

Figure 3: Conditional proportions underlying the value–action gap. Each cell reports the proportion of cases in which the model (or humans) chooses to attack a target given that it assigns the highest belief to that target being secured, along with the corresponding sample size. Columns show: P(Attack B1 | B1 highest), P(Attack B2 | B2 highest), and the aggregate weighted proportion.

5.3 Value Action Gap

While the original value-action gap paper (Shen et al., 2025)'s approach is theoretically grounded and conceptually novel, the authors do not benchmark their findings against the human value–action gap, which is essential for assessing how closely LLMs approximate human decision-making. In our experimental setting, we define a value–action gap as occurring when a participant, human or LLM, assigns the highest belief probability to a particular target being secured yet still chooses to attack that same target. In reality, if an attacker believes a target it highly secure, it's unlikely they would attack that target as opposed to selecting a less secure target or selecting not to attack. The unfinetuned model exhibits the largest value–action gap, almost always choosing to attack the target it assigns the highest probability of being secured. In contrast, all finetuned models display substantially smaller gaps. As shown in Figure 3, the value–action gap for the unfinetuned baseline is close to 100%, whereas even models finetuned on only 25% of a single treatment group reduce the gap by more than half. Unlike the alignment results in the previous section, where the relationship between the number of included treatment groups and JS distance was ambiguous, the value–action gap consistently shrinks as more treatment groups are incorporated into finetuning. Models trained on 25% of all eight treatment groups attain discrepancy rates that closely match human participants

(23.8% and 24.4% vs. 24.7%).

Moreover, the gap tends to be smaller for the balanced subgroup finetuning conditions than for the random sampling conditions, a pattern that did not emerge for distributional alignment. This suggests that while matching full belief distributions may require more sophisticated steering, reducing the value–action gap is comparatively easier to influence through targeted finetuning.

5.4 Coefficient Estimation

Given that finetuned models can increase heterogeneity, improve alignment with human data, and reduce the value–action gap, the natural next step becomes whether these simulated responses can recover the same hypothesis tests and coefficient estimates as the original human experiment. To assess this, we replicate the six regression models reported in the original study. These models fall into three conceptual categories: the first two examine whether beliefs predict an attacker's decision to abort, the next two predict which target attackers choose, and the final two assess target choice using the normative expected value difference.

Overall, all models produce inaccurate coefficient estimates, with the unfinetuned model and the models finetuned on only a single treatment group performing the worst. As shown in Table 2, there are 22 total coefficients across the six replicated regression models, 13 of which are statistically significant in the human data. The unfinetuned model yields only 14 coefficients because it never chooses the "abort" decision, making it impossible to estimate Models 1 and 2.

To evaluate how well each simulated dataset recovers the human regression results, we compute F1(Sign), which jointly accounts for false positives and false negatives while also requiring that true positives have the correct sign. Across all models, the F1(Sign) values remain suboptimal. Even the best-performing finetuned model achieves an F1(Sign) of only ≈ 0.769, and its false positive rate (FPR) remains high at ≈ 0.333.

These results indicate that, although finetuning on relatively small subsets of data substantially improves heterogeneity, alignment, and belief–action coherence, the resulting simulated datasets still introduce non-negligible bias when used for formal statistical inference.

6 Conclusion

This study examines whether finetuning on small samples of human participant data can mitigate well-documented limitations of using base LLMs to simulate survey responses (Gao et al., 2025; Wang et al., 2025a; Shen et al., 2025; Bisbee et al., 2024), using a behavioral experiment. Our analysis focuses on four key constructs: heterogeneity, distributional alignment, the value–action gap, and regression coefficient recovery. We find that finetuning on relatively small subsets of data substantially increases heterogeneity, improves alignment, and enhances belief–action coherence compared to the unfinetuned model, as demonstrated through both visual inspection and formal divergence metrics. Notably, heterogeneity can be substantially improved using as little as 25% of a single treatment group, approximately 30 observations. Alignment tends to strengthen as the number of treatment groups used for finetuning increases. Unexpectedly, however, sampling equal proportions from ethnicity subgroups does not consistently enhance alignment for minority groups relative to random sampling, and misalignment remains lowest for White participants even under balanced sampling. While unexpected, this pattern aligns with prior work showing that LLMs are more closely aligned with the WEIRD population (Liu et al., 2022). In contrast, when examining the value–action gap, balanced sampling performs consistently better than random sampling, suggesting that belief–action coherence may be easier to steer than subgroup-level distributional alignment, which may require more advanced finetuning strategies.

Despite successfully mitigating several undesirable properties of the base model, our findings show that none of the finetuned models accurately replicate the regression results from the original study. This limitation indicates that full replacement of human participants remains infeasible and may introduce significant threats to the validity of empirical conclusions. Instead, we argue that LLMs finetuned on small human samples, as few as 30 observations, are better suited for in silico prototyping than base LLMs. Such models can assist researchers in refining survey instruments and conducting power or sample size analyses, but should be viewed as complements to, rather than substitutes for, human subjects (Anthis et al., 2025; Dillion et al., 2023; Binz et al., 2025).

Table 2: Comparison of coefficient recovery metrics across simulation groups.Total Coef is the total number of coefficients estimated. Reference is the number significant in the human data. Sim is the number significant in the LLM-generated data. N Match counts coefficients for which both significance and sign are correctly recovered. FPR and FNR denote false-positive and false-negative rates, and F1 (Sign) is the F1 score incorporating both significance and sign.

Sim Group	Total Coef	Reference	Sim	N Match	FPR	FNR	F1 (Sign)
Balanced (1a1b)	22	13	10	5	0.5555	0.6154	0.4348
Random (1a1b)	22	13	3	0	0.1111	0.8462	0.0000
Balanced (All Groups)	22	13	15	9	0.6667	0.3077	0.6429
Balanced (4 Groups)	22	13	13	10	0.3333	0.2308	0.7692
Random (All Groups)	22	13	12	7	0.4444	0.3846	0.6087
Random (4 Groups)	22	13	13	9	0.4444	0.3077	0.6923
Unfinetuned Baseline	14	8	11	2	0.8333	0.2500	0.3636

7 Discussion

Several mechanisms may explain the misalignments between LLM-simulated and human responses. Wang et al. (2025a) argue that LLMs often engage in misportraying, rather than representing the authentic behaviors of a given demographic group, models reproduce what out-group members believe those in-group members would do. Since demographic identity is rarely encoded directly in the text on which LLMs are trained, explicit mentions of identity may instead express stereotypes by out-group authors. Liu et al. (2024) offer a related explanation, and it aligns with the sociolingustic theoretical model introduced by Zhou et al. (2025). They study LLM performance on datasets involving risky decision-making and find that LLMs align more closely with people's expectations of how others behave than with how individuals actually behave. Since people tend to assume that others are more rational than themselves, LLMs may encode these perceptions rather than genuine human behavior. These findings suggest that LLM decisions may be shaped less by representations of actual in-group behavior and more by culturally shared beliefs about what "typical" behavior should look like. Additional factors may also contribute to misalignment. The training data themselves may contain survey responses affected by social desirability bias (Crowne and Marlowe, 1960), meaning that models learn socially acceptable patterns of responding rather than authentic responses. More directly, reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) optimizes models to behave in ways that humans prefer LLMs to behave (Dahlgren Lindström et al., 2025), not necessarily in ways that mirror empirical human data.

A concurrent work by Krsteski et al. (2025) also examines finetuning on a small human survey sample, but in a different setting. They study panel surveys and use earlier waves as historical data to finetune or prompt LLMs and generate synthetic responses for a later target wave. A small subsample of human responses from that target wave is then used for statistical rectification of the model-based estimates, and their main focus is how to optimally allocate limited human responses between finetuning and post-hoc correction to reduce bias in LLM-based survey simulations. By contrast, our study does not focus on bias correction, rather, we evaluate whether finetuning on small samples can recover key behavioral constructs within a controlled experimental setting.

8 Limitations

This study has several limitations. First, we rely on a single case study from one behavioral domain, and our findings may not generalize to other experimental settings or survey contexts. Second, since finetuning was performed multiple times, all comparisons were conducted using a single base model. Future work could assess whether our results hold across different model families or architectures. Third, due to imbalance in the original dataset, our balanced finetuning sets could only include up to 25% of the data without serious violation of the equal-proportion constraint, which may limit the sample points available for minority subgroups. Finally, to remain faithful to the original experimental design, we did not systematically explore the sensitivity of LLMs to prompt phrasing, format, or ordering (Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024), factors that could meaningfully influence model behavior and should be examined in future research.

9 Ethics Statement

Although this study explores the simulation of human responses using LLMs, we do not advocate replacing human participants with LLMs for informing policy, healthcare, or other high-stakes decisions, particularly those affecting disadvantaged or marginalized populations, given the limitations identified in this work and in numerous prior studies. We adhere to the ACL Code of Ethics throughout the research process and comply with OpenAI's usage policies when using GPT models and fine-tuning tools. We ensured that the data from the original experiment contain no harmful or stereotypical content. We also confirmed with the original author that the study received IRB approval from their institution. All datasets and code will be released for academic use in accordance with ethical research guidelines.

Acknowledgments

TBD

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S Bernstein. 2025. Position: Llm social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

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

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, and 1 others. 2025. A foundation model to predict and capture human cognition. *Nature*, pages 1–8.

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

James Blake. 1999. Overcoming the 'value-action gap'in environmental policy: Tensions between national policy and local experience. *Local environment*, 4(3):257–278.

James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using llms for market research. *Harvard business school marketing unit working paper*, (23-062).

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Annual Meeting of the Association for Computational Linguistics*.

Shan-Shan Chung and Monica Miu-Yin Leung. 2007. The value-action gap in waste recycling: The case of undergraduates in hong kong. *Environmental Management*, 40(4):603–612.

Douglas P Crowne and David Marlowe. 1960. A new scale of social desirability independent of psychopathology. *Journal of consulting psychology*, 24(4):349.

Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2025. Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback: Ad lindström et al. *Ethics and Information Technology*, 27(2):28.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

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

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2025. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in llms using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188.

Gaston Godin, Mark Conner, and Paschal Sheeran. 2005. Bridging the intention–behaviour gap: The role of moral norm. *British journal of social psychology*, 44(4):497–512.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. *Preprint*.

Kyle Hunt, Şule Güney, Jun Zhuang, and Richard John. 2025. On the disclosure of defensive posture: Adversarial belief formation and target selection decisions. *Production and Operations Management*, 34(7):1995–2014.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks. *arXiv preprint arXiv:2408.08631*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113.

Stefan Krsteski, Giuseppe Russo, Serina Chang, Robert West, and Kristina Gligorić. 2025. Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. *arXiv preprint arXiv:2510.11408*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Zhicheng Lin. 2025. Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3):25152459251357566.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. 2024. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person (a): A systematic evaluation of sociodemographic persona prompting for large language models. *arXiv preprint arXiv:2507.16076*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Federico Ranaldi, Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. 2024. Investigating the impact of data contamination of large language models in text-to-sql translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13909–13920.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. Large language models do not simulate human psychology. *arXiv preprint arXiv:2508.06950*.

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the value-action gap: Do llms act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025a. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

Anyi Wang, Dong Shu, Yifan Wang, Yunpu Ma, and Mengnan Du. 2025b. Improving llm reasoning through interpretable role-playing steering. *arXiv preprint arXiv:2506.07335*.

Zengqing Wu, Run Peng, Takayuki Ito, and Chuan Xiao. 2025. Llm-based social simulations require a boundary. *arXiv preprint arXiv:2506.19806*.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. 2024. Quantifying and analyzing entity-level memorization in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19741–19749.

A Prompt

A.1 Prompt 1

{Background Info} Based on your preliminary surveillance, you have identified the following information: The value you place on the treasure at Beach 1: {Your Valuation 1} coins The value you place on the treasure at Beach 2: {Your Valuation 1} coins The value Queen Katrina places on the treasure at Beach 1: {Queen's Valuation 1} coins The value Queen Katrina places on the treasure at Beach 2: {Queen's Valuation 2} coins Therefore, if you successfully rob either beach, you would have {Reward Value} coins worth of treasure. There will a fixed cost for you to rob the beaches given that you will need to travel to Gold Island and execute the robbery. These costs are as follows: Cost to rob Beach 1: {Cost 1} coins Cost to rob Beach 2: {Cost 2} coins As you were watching the news today, you learned that Queen Katrina would be making a big announcement later this afternoon. You have decided to wait until you hear what she has to say before planning and executing your big robbery. Based on this initial information, which beach would you choose to rob? Output your choice (beach 1 or beach 2) inside square brackets (array 1) Final output should only contain array one.

A.2 Prompt 2

At this point, you will read Queen Katrina's announcement. Her announcement is in regards to new security measures that have been deployed to protect the treasure. You know that the new security measures could be deployed only at Beach 1, only at Beach 2, or at both beaches. If the new security measures are actually deployed at Beach 1, then the chance of a robbery being successful there is 30%. Similarly, if the new security measures are actually deployed at Beach 2, then the chance of a robbery being successful there is also 30%. The problem is that only Queen Katrina knows exactly where the new security measures are deployed. Furthermore, the information she releases could be totally true, totally false, or partially false (i.e., some parts of her announcement could be true, while other parts could be false). After reading her announcement, you will be asked a series of questions regarding your beliefs about the security measures, and your plans for the robbery. Queen Katrina's announcement: {Queen's Announcement} treasure. Reminder (data from previous page): The value Queen Katrina places on the treasure at Beach 1: {Queen's Valuation 1} coins The value Queen Katrina places on the treasure at Beach 2: {Queen's Valuation 2} coins What is the chance (in your belief) that the new security is deployed only at Beach 1, only at Beach 2, or both? Choose a value between 0–100% (note: the sum of all 3 prompts must equal 100%): Output your beliefs (3 percentages that add up to 100%) inside square brackets (array 2)

What influenced your responses to the question? Choose all options that apply: Option 1: The information that Queen Katrina released in her announcement Option 2: Queen Katrina's values of the treasure at the different beaches Output your answer inside square brackets (array 3)

Aside from the base pay for this study, you are now receiving more money which you can use to fund your robbery. In specific, 1 coin in the game is equivalent to 0.000001 USD (United States dollars). For example, it costs 150,000 coins to rob Beach 2 in this game, which is equivalent to $0.15. You now have $0.15 that you can use to fund your robbery. If you choose not to rob either beach, you keep the $0.15. If you choose to rob Beach 1, you will spend $0.10 of the $0.15, and you keep the remaining $0.05. If you succeed in robbing Beach 1, you will receive $0.80 in additional money. If you fail, you will receive $0.00 in additional money. If you choose to rob Beach 2, you will spend the total $0.15. If you succeed in robbing Beach 2, you will receive $0.80 in additional money. If you fail, you will receive $0.00 in additional money. Reminders The probability of successfully robbing a beach if the new security measures are deployed is 30%, and the probability of successfully robbing a beach if the new security measures are not deployed is 80The value you place on the treasure at Beach 1: {Your Valuation 1} coins The value you place on the treasure at Beach 2: {Your Valuation 2} coins Cost to rob Beach 1: {Cost 1} coins Cost to rob Beach 2: {Cost 2} coins The value Queen Katrina places on the treasure at Beach 1: {Queen's Valuation 1} coins The value Queen Katrina places on the treasure at Beach 2: {Queen's Valuation 2} coins What will be your final plan for the robbery? Choose one of the following responses: Option 1: Attempt to steal the treasure from Beach 1 Option 2: Attempt to steal the treasure from Beach 2 Option 3: Abort the heist altogether and do not attempt to steal treasure from either beach Output your decision (only 1, 2, or 3) inside square brackets (array

4) Briefly explain your response to the previous question (50 words maximum) Output your explanation inside square brackets (array 5) Final output should only contain array 2, 3, 4, 5, separated by a comma.

B Experimental Manipulations

Table 3: Manipulations from the experiments in Hunt et al. (2025)

Notation	Explanation
V1	Large difference in the defender's target valuations ($1.2 million vs. $400,000)
V2	Small difference in the defender's target valuations ($850,000 vs. $750,000)
A1	Information released regarding the deployment of the new defense measures; **no target-level information released**
A2	Information released regarding the deployment of the new defense measures *and* that the defense measures are deployed at **Target 1**
A3	Information released regarding the deployment of the new defense measures *and* that the defense measures are deployed at **Target 2**
A4	Information released regarding the deployment of the new defense measures *and* that the defense measures are deployed at **Targets 1 and 2**

C Belief

As shown in Figure 4, expanding finetuning to include randomly sampled data from four treatment groups ("Random (4 Groups)") and all eight groups ("Random (All Groups)") further improves diversity. The balanced variants, "Balanced (1a1b)," "Balanced (4 Groups)," and "Balanced (All Groups)," also show similarly enhanced diversity and alignment.

The distributions of the finetuned models for Targets 2 closely resemble the human data. As shown in Figure 5, they are right-skewed, with most probability mass concentrated in the lower ranges of the distribution. In contrast, the unfinetuned baseline shows a highly concentrated pattern, placing nearly all probability mass in the 20–40% range and assigning no probability beyond 40%. Compared to the belief distributions for Targets 1 and Both Targets, the peak of the unfinetuned model for Targets 2 is actually closer to the human data. Both the human data and the unfinetuned model place their highest probability mass in the 30–40% range. The human distribution is bimodal, with a secondary peak in the 0–10% bin. In contrast, most of the finetuned models exhibit a Gaussian-like, unimodal distribution centered around the middle bins. The exception is Random (1a1b), which remains noticeably right-skewed.

D Group-Level JS

At the group level, the unfinetuned model continues to show the largest divergence from the human data, whereas all finetuned models achieve at least roughly half the JS distance of the baseline. As shown in Figure 6, the reduction in JS distance is mildly correlated with the number of treatment groups included in the finetuning data: models finetuned on all eight groups perform slightly better than those trained on four groups, which in turn outperform the one-group models. The models finetuned on the first four groups exhibit somewhat lower divergence within those same groups—an expected pattern given their training data, but the differences are small. This suggests that the finetuned models retain a reasonable degree of generalizability beyond the specific treatment groups they were trained on.

E Ablation JS

Figure 7 presents an ablation study conducted on one of our finetuned models. Random – No Demographic refers to a variant trained without participants' background information. Random – No Decision and Random – No Belief refer to variants that exclude participants' reported final decisions and belief responses, respectively, during finetuning. Due to budget constraints, we perform this ablation on a single representative model (Random (4 Groups)) rather than across all finetuned models in the manuscript. As shown in Figure 7, the Jensen–Shannon (JS) distances for Random – No Demographic are similar to those of the original finetuned model across ethnic subgroups, suggesting that including demographic background information alone does not meaningfully reduce subgroup misalignment in belief distributions. In contrast, Random – No Belief exhibits larger JS distances than the other finetuned variants, indicating increased misalignment when decision information is excluded, nevertheless, its performance remains substantially better than that of the unfinetuned baseline. Finally, Random – No Decision performs comparably to the original model, implying that adding decision information does not indirectly improve belief alignment when belief data themselves are omitted during finetuning.

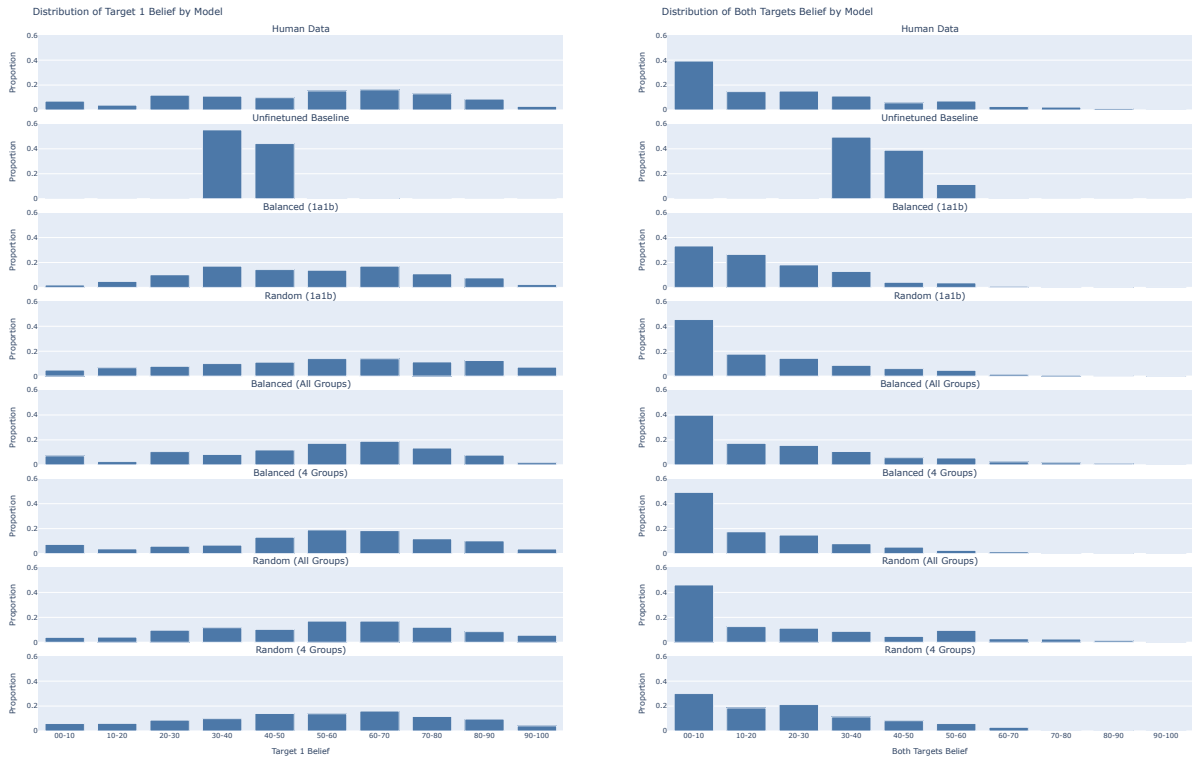

Figure 4: Distributions of beliefs that security is deployed at Target 1 (left) and at both targets (right) across human data and LLM simulation conditions. Each panel shows the empirical belief distribution for human participants alongside models finetuned under different sampling strategies.

Figure 5: Distribution of beliefs for Targets 2 across human data and LLM simulation conditions. Each panel shows the empirical belief distribution for human participants alongside models finetuned under different sampling strategies.

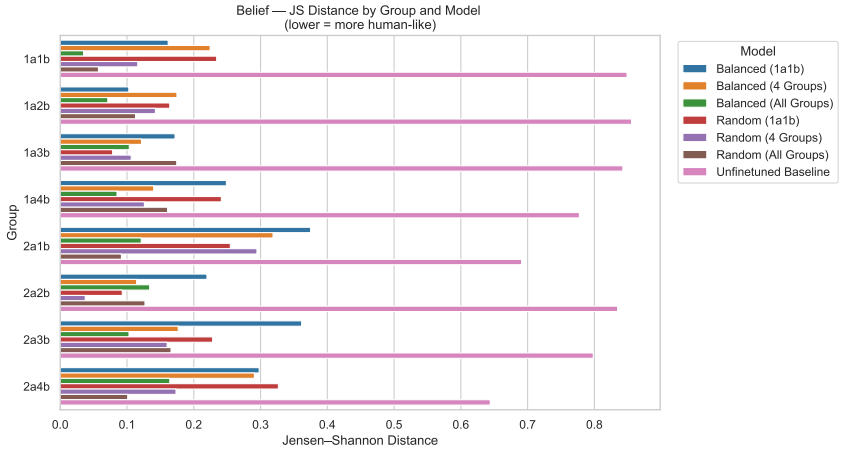

Figure 6: Jensen–Shannon (JS) distance between LLM-simulated and human belief distributions by treatment group and finetuning strategy.

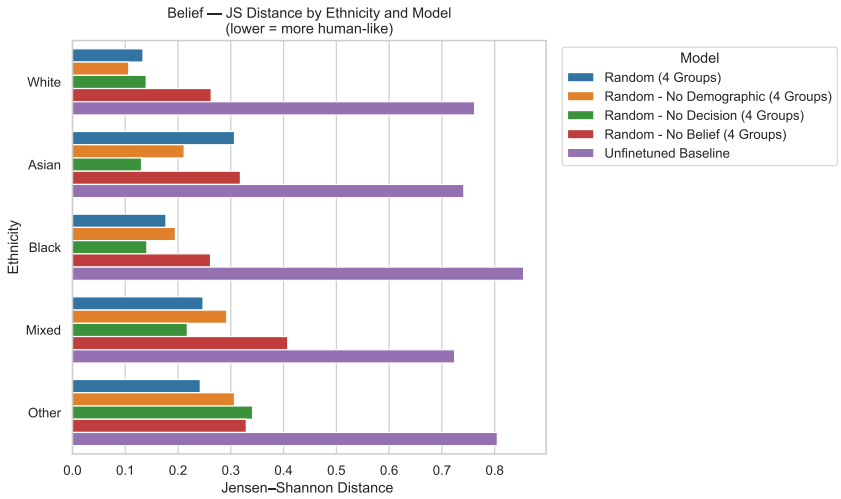

Figure 7: Ablation study of a finetuned LLM showing Jensen–Shannon (JS) distances between simulated and human belief distributions by ethnicity (lower values indicate closer alignment). “Random – No Demographic,” “Random – No Decision,” and “Random – No Belief” exclude demographic background, decision responses, and belief responses, respectively, during finetuning. The unfinetuned baseline is shown for comparison.