Reason-Mediated Behavioral Models for Auditing LLM Social Simulators

Anonymous authors
Paper under double-blind review

Abstract

Large language models are increasingly used as social simulators, including as synthetic survey respondents. Most evaluations ask whether simulated outcomes resemble human outcomes. We argue that this is necessary but too weak: a simulator can match the final answer while using the wrong behavioral mechanism. We study this problem through a 94-person sunscreen concept test in which each respondent evaluated three product concepts and wrote open-ended rationales. We map those rationales into signed reason states Z, where positive signs support adoption and negative signs block it. This gives a practical audit: holding respondent descriptors D, category context K, and concept treatment X fixed, do human-stated reasons help predict behavior Y, and can an LLM simulate the same reason state without seeing the human rationale or outcome? Human-stated reasons substantially improve held-out prediction of purchase intent. LLM-simulated reasons are more brittle: they often sound plausible, but frequently echo the concept board rather than recover the respondent's acceptance or rejection path. The paper contributes an evaluation framework for social simulators. Reason states do not identify natural causal effects by themselves, but they provide an interpretable test of whether a simulator's stated reasons align with human evidence.

1 Introduction

LLM-based social simulation is moving from demonstration to evaluation. Recent systems simulate individual agents, synthetic samples, and multi-agent societies (Argyle et al., 2023; Park et al., 2023; Anthis et al., 2025). The central question is no longer only whether these simulations are believable. It is whether they are faithful enough to support scientific or practical inference.

A common validation strategy compares simulated outcomes with human outcomes. In a concept test, for example, one might ask whether a model predicts the same winning product or a similar purchase-intent distribution. This is necessary, but not sufficient. Outcome agreement alone cannot tell us whether the model recovered the behavioral path that produced the outcome. A sunscreen concept can fail because it is too expensive, because its proof is not trusted, because the texture feels wrong for daily use, or because the respondent does not identify with the positioning. These mechanisms imply different interventions, even when they produce the same rating.

This paper studies a stricter object: the reason path between a stimulus and a response. We represent this path with a signed reason state Z. A positive sign means that a reason supports adoption; a negative sign means that it blocks adoption; zero means that the reason is inactive. The goal is not to infer a natural causal effect from open-ended text alone. Rather, we ask whether a fixed, human-readable reason codebook can support a reason-mediated behavioral model for auditing social simulators.

The resulting test is simple. Respondent descriptors D include demographics, language, and personality items. Category context K contains sunscreen-specific prior state: use frequency, brands owned, white-cast concern, and trust in SPF claims. Concept treatment X contains the price, claims, authority cue, origin, and proof points in the concept board. Human

rationales are mapped into Z. A readout $f_\theta(D, K, X, Z)$ is trained to predict behavior Y. We then replace human Z with LLM-simulated $\widehat{Z}$ and ask whether that reason state helps the same readout. This separates two questions that are often conflated: can the model predict an answer, and can it recover the reason state that makes the answer intelligible?

We evaluate this protocol on a small sunscreen concept test with 94 human respondents and three concepts. Each respondent rated purchase intent, believability, and differentiation, wrote rationales, ranked concepts, and made a final pick. The study is intentionally narrow. Its purpose is to test the evaluation idea, not to make population claims about sunscreen buyers.

The paper makes three contributions.

1. We define a reason-mediated behavioral model for concept-test simulation: D, K, X shape Z, and Z helps predict Y.
2. We show that human rationale-derived reason states improve held-out prediction of purchase intent.
3. We show a negative result for current LLM reason simulation: LLM-simulated reasons can be fluent and plausible while failing to match the human reason path.

2 Related Work

LLM social simulation. Prior work has shown that LLM agents can generate believable individual behavior, memories, plans, and social interactions (Park et al., 2023); synthetic samples can approximate some survey patterns (Argyle et al., 2023); and agent societies can be used to explore group dynamics. At the same time, recent critiques emphasize that plausible simulation is not the same as valid social evidence. LLMs may flatten identity groups when used as participant replacements (Wang et al., 2025), and apparent emergent dynamics may be observationally compatible with data leakage or prompt artifacts (Barrie & Törnberg, 2025). These critiques motivate evaluations that inspect the mechanism of a simulated response, not only the surface form or final answer.

Causal structure and refutation. Causal inference papers often separate four tasks: modeling assumptions, identifying the estimand, estimating the effect, and refuting or stress-testing the result (Sharma & Kiciman, 2020; Pearl, 2009). They also show why naive outcome comparisons can be misleading when the process that generates exposure is entangled with the process that generates behavior (Sharma et al., 2015). Work on LLMs and causality asks whether language models can produce valid causal arguments (Kıcıman et al., 2024). Observed rationales are not randomized mediators, so they do not by themselves identify natural direct or indirect effects in the sense of mediation analysis (Imai et al., 2010). A weaker but useful target is refutation: define an interpretable interface, hold the readout fixed, and check whether the interface behaves coherently under prediction, signed perturbation, ablation, and simulator substitution.

Behavioral mediators. The reason-state codebook is motivated by behavioral theories in which beliefs, attitudes, norms, values, and perceived constraints shape intentions and actions (Fishbein & Ajzen, 1975; Ajzen, 1991; Schwartz, 1992). For concept tests, the useful coding level is lower than broad constructs such as attitude or value. A node such as price/value, safety trust, or sensory fit is specific enough to be coded from text and manipulated in a model, but general enough to transfer across related concept tests.

3 Problem Setup

Variables. The unit of analysis is a respondent-concept cell: respondent i evaluating concept c. We use five objects:

$$D_i, \quad K_i, \quad X_c, \quad M_{ic}, \quad Z_{ic}, \quad Y_{ic}.$$

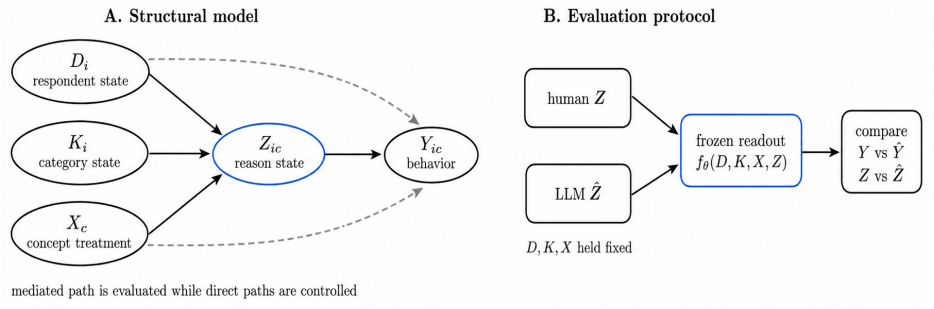


Figure 1: The reason-mediated behavioral model. We hold D, K, X fixed and test whether human or LLM-simulated reasons Z support the same behavioral readout. Bold arrows show the evaluated mediated path; dotted gray arrows show controlled direct paths.

D_i is the respondent state: demographics, language, and personality measures. K_i is the category state: sunscreen-use frequency, brands owned, white-cast concern, and trust in SPF claims. X_c is the concept treatment: price, claims, authority cue, positioning, and proof points. M_{ic} is the measured intermediate-rating vector, here believability and differentiation. Z_{ic} is the reason state expressed or predicted for respondent i on concept c. Y_{ic} is behavior, mainly purchase intent on a 1–5 scale.

Reason states. Z is a signed vector of intermediate behavioral reasons:

$$Z_{icj} \in \{-1, 0, +1\}.$$

For reason family j, $+1$ means the reason supports purchase, -1 means it blocks purchase, and 0 means it is inactive. The sign is coded from the rationale, not from the purchase-intent score. For example, price/value is positive when the respondent says the product is worth the price and negative when she says it is too expensive for daily use.

Core reason states. The main paper uses a core reason-state vector, denoted Z_{core}. This is the subset of reason families that are interpretable and not direct restatements of the outcome. It contains price/value, proof trust, natural orientation, premium or K-beauty aspiration, local fit, sensory/routine fit, and safety trust. Outcome-near labels, such as explicit purchase intention, are kept only for sensitivity checks because they can leak Y.

Reason node	Family	Sign meaning
price/value	Value and constraint	+1: worth the price; −1: too expensive or poor value
proof trust	Evidence and authority	+1: claims are trusted; −1: claims are doubted
natural orientation	Natural/cultural cue	+1: natural or traditional cue helps; −1: cue reduces trust
premium/K-beauty	Novelty and status	+1: premium or K-beauty appeal; −1: excessive or irrelevant
local fit	Local adaptation	+1: fits Indian skin/climate; −1: local fit is missing
sensory/routine fit	Daily-use fit	+1: texture/no-white-cast helps; −1: routine fit blocks use
safety trust	Safety and risk	+1: safety is reassured; −1: safety/regulatory anxiety blocks use

Table 1: Core Z codebook used in the main readout. The same node can be positive, negative, or inactive for a respondent-concept cell.

Reason-mediated behavioral model. We use reason-mediated behavioral model in a restricted, operational sense. The model represents behavior through an explicit reason state:

$$p(Y, Z \mid D, K, X) = p(Y \mid D, K, X, Z)\; p(Z \mid D, K, X). \tag{1}$$

This factorization should be read as an audit model rather than an identification claim for natural direct or indirect effects. Once a codebook and readout are fixed, we can ablate reason families, flip signs, and replace human reasons with LLM-simulated reasons. The question becomes whether the simulator places the right reason state in the middle, alongside getting the final answer right.

Why direct outcome simulation is insufficient. A direct synthetic respondent gives

$$D, K, X \longrightarrow \widehat{Y}.$$

Our protocol asks for the stronger decomposition

$$D, K, X \longrightarrow \widehat{Z} \longrightarrow \widehat{Y}.$$

The second test can fail even when the final answer is directionally correct. That failure is useful: it catches plausible rationales that do not play the same behavioral role as human-stated reasons.

4 Methodology

4.1 Concept Test

Study. The empirical setting is a sunscreen concept test with 94 usable human respondents and three product concepts, yielding 282 respondent-concept cells. The target sample was English-speaking Indian women aged 22–40, urban or semi-urban, with recent sunscreen purchase. Six responses were excluded for attention, rurality, or straight-line quality rules; the analytic sample had mean age 28.7 and city-tier counts of 51, 19, and 24 for Tier 1, Tier 2, and Tier 3. The three concepts varied on price, authority, cultural positioning, local fit, and sensory promise: SunVeda, an Ayurvedic/botanical concept at Rs. 499; Dr. Anita, a dermatologist/clinical Indian concept at Rs. 699; and Hae, a Korean prestige hybrid concept at Rs. 1,299. Appendix B gives the full distribution and concept text.

Measured outcomes. For each concept, respondents provided purchase intent, believability, differentiation, and an open-ended rationale. They then ranked the concepts and made a final pick. The main target is purchase intent $Y \in \{1, \ldots, 5\}$. We also report top-two-box intent, $T = \mathbb{1}[Y \geq 4]$, and final-pick checks.

Questionnaire design. The survey first collected screening and respondent descriptors, including age, city tier, education, income proxies, languages, and BFI-10 personality items. It then asked category questions about sunscreen-use frequency, brands owned, white-cast concern, and SPF-claim trust, before showing the three concept boards in randomized order. For each board, respondents answered purchase intent, believability, differentiation, and an open-ended "why" question. Thus D comes from respondent descriptors; K from sunscreen-category questions; X from the concept board; M from believability and differentiation; Z from the open-ended rationale; and Y from purchase intent, ranking, and final pick. Appendix B gives the full mapping and an example concept board.

4.2 Reason-State Construction

Codebook. The reason codebook maps open-ended rationales into signed reason states. It was built from the sunscreen concept-test task but uses families that are meant to be reusable across concept tests: value, proof trust, safety, local fit, novelty/status, routine fit, and category-specific authority cues. The leaves are domain-specific; the parent families are the transferable layer.

Codebook construction. The codebook was constructed before the predictive readout was evaluated. We first collected all open-ended rationales and listed recurring supports and objections without using purchase intent as a label. Surface variants were then merged into parent families, such as price/value, proof trust, local fit, safety trust, and sensory/routine fit. Outcome-near labels, including explicit intention or generic liking, were removed or kept only for sensitivity checks because they can leak Y. Each retained family was assigned a signed coding rule: $+1$ for support, -1 for objection, and 0 for inactive. The resulting main-paper codebook is Table 1; the appendix gives the machine-readable node names.

Coding rule. The extractor is allowed to use the respondent's rationale and the concept text for disambiguation, but not the rating, rank, final pick, or any gold reason label. It also cannot mark a reason only because the concept contains that attribute. For example, the Hae concept contains a K-beauty cue, but premium/K-beauty is coded positive only if the respondent uses that cue as a reason.

Extractor. The final human Z file was produced by a small-extractor, larger-verifier pipeline. A smaller LLM extractor (gpt-5.4-mini) coded the respondent-concept rationales into JSON over the fixed codebook, using the concept text only for disambiguation. Lexical and embedding extractors, including an Ollama nomic-embed-text run, were used as audit signals rather than as the final source of Z. We then selected verifier rows where extraction was most likely to be brittle: zero-Z rows, warning cases, embedding–LLM disagreements, dense reason rows, and substantive hard cases such as Hae price rejection and Ayurveda skepticism. A larger verifier (gpt-5.5) recoded those rows; verifier outputs overrode the smaller-extractor outputs for the selected rows, while all other rows remained from gpt-5.4-mini. The adjudicated raw nodes were then compressed into the seven core Z_{core} families used in the main readout. No extractor saw purchase intent, believability, differentiation, ranking, final choice, or any gold reason label.

4.3 Human Reason Model

Readout. The human reason model is a supervised readout

$$f_\theta(D, K, X, Z) \to Y.$$

The primary readout is ridge regression with five-fold GroupKFold cross-validation by respondent, so all rows from a respondent remain in the same fold. Numeric variables are imputed and scaled; categorical variables are one-hot encoded; predictions are clipped to the 1–5 scale. This is the main model for purchase intent. We separately use balanced logistic regression to test whether the same features predict the respondent's final chosen concept. Random forests and gradient boosting are nonlinear checks: they test whether the main pattern depends on a linear readout.

What this tests. The human reason model asks whether human rationale-derived reason states have behavioral content after controlling for respondent descriptors, category context, and concept features. If Z_{core} improves held-out prediction, then it is more than decorative explanation.

4.4 LLM Reason Simulation Protocols

Design. We evaluate three pre-specified LLM simulation protocols. They answer different questions. Survey simulation asks the LLM to behave like a respondent and produce answer text, which is then scored into a purchase-intent distribution. Reason-state simulation asks the LLM to predict $\widehat{Z}$ directly from D, K, X and the codebook. Behavior readout asks whether a supplied reason state, human or simulated, predicts Y under the same human reason model. The comparison is therefore not one generic "LLM simulation" result; it separates final-answer simulation from reason-state simulation.

Persona matching. The persona-based survey simulations use a Nemotron-Personas-India pool (NVIDIA, 2025). The matching code keeps female, urban personas aged 22–40 and

assigns one persona to each human respondent using seeded stratified matching by city tier and age band. The primary bucket is (city tier, age band); fallbacks use same city tier or same age band. This is a prior over respondent background for persona-conditioned simulations. It should not be confused with observed D: in the human study, D comes from survey fields; in persona-conditioned LLM simulations, part of the respondent description comes from the matched Nemotron persona. The full simulation run includes demographic-only and persona-conditioned arms, so the persona prior can be ablated for native survey simulation. The main reason-state evaluation uses only arms that produce text from which a comparable $\widehat{Z}$ can be extracted.

Survey simulation: context-conditioned response. The context-conditioned protocol uses Gemma 4 31B IT for high-volume synthetic respondent generation. It combines the matched persona with a stratum-level sunscreen context paragraph generated from non-outcome category variables only: sunscreen-use frequency, brands owned, white-cast concern, SPF-claim trust, languages, and personality summaries. Product outcomes and rationales are excluded. The model writes a short survey-style response. That response is scored into a purchase-intent probability mass function by semantic similarity rating (SSR) (Maier et al., 2025), and the same generated text is mapped into $\widehat{Z}$ for the $\widehat{Z} \to Y$ reason-path test. SSR mechanics are in Appendix F.

Reason-state simulation: codebook-conditioned no-leak generator. The codebook-conditioned protocol asks directly for $\widehat{Z}$ using only observed D, K, X and the codebook. This constrained $D, K, X \to \widehat{Z}$ generation used the OpenAI verifier/generator pipeline rather than the Gemma survey-simulation model, because it is lower-volume and requires structured sparse JSON rather than respondent-style text. It does not use a Nemotron persona prior, and it does not see human open text, ratings, ranks, final pick, or gold reason labels. This is the strictest generation setting: the model must infer a sparse reason state from observed context and concept treatment alone. We therefore treat the results as protocol comparisons under a shared human reason model, not as a head-to-head model-family benchmark.

4.5 Voice-Pool Survey Simulation

Motivation. Single-prompt synthetic respondents tend to compress toward moderate answers. That is a problem for concept tests because many business decisions depend on tails: strong rejection due to price or safety doubt, and strong adoption due to proof, fit, or aspiration. The voice-pool simulation is a capacity test for this missing tail behavior. For the same respondent-concept cell, it samples three response modes: skeptic, balanced, and enthusiast. Each mode is prompted with the matched persona, the stratum context, the target concept, and the respondent's own open-ended answers on the other two concepts. The target concept's human rationale and outcome are never shown.

Aggregation. Each voice produces two samples. All six samples are scored with SSR. The primary aggregate averages the six PMFs:

$$\hat{p}(Y \mid D, K, X) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \hat{p}_v(Y \mid D, K, X), \quad \mathcal{V} = \{\text{skeptic}, \text{balanced}, \text{enthusiast}\}.$$

We also evaluate selector checks. Oracle selection estimates how much useful signal exists in the voice pool. Learned selectors ask whether that signal can be recovered from deployable features.

4.6 Evaluation

Human reason model evaluation. The first evaluation asks whether human-stated reasons help predict human behavior. We compare D, K, X against $D, K, X + Z_{\text{core}}$, and also compare $D, K, X + M$ against $D, K, X + M + Z_{\text{core}}$. This tests whether reason states add signal beyond observed covariates and measured intermediate ratings.

Model	MAE	Spearman	Top2 AUC	Final-pick AUC
D, K, X	0.863	0.245	0.628	0.602
$D, K, X + Z_{\text{core}}$	0.625	0.678	0.882	0.726
$D, K, X + M$	0.717	0.538	0.762	0.703
$D, K, X + M + Z_{\text{core}}$	<u>0.564</u>	<u>0.749</u>	<u>0.902</u>	<u>0.762</u>
$D, K, X + M + Z_{\text{all}}$	0.535	0.772	0.917	0.779

Table 2: Main readout results. Best is bold; second-best is underlined. Z_{all} includes outcome-near labels and is only a sensitivity check. The main claim uses Z_{core}.

Reason-path checks. For each core reason node j, we set that node to $+1$ and to -1 inside the learned readout:

$$\Delta_j = \mathbb{E}\left[f_\theta(D, K, X, do(Z_j = +1))\right] - \mathbb{E}\left[f_\theta(D, K, X, do(Z_j = -1))\right]. \tag{2}$$

Because the codebook defines $+1$ as support and -1 as opposition, Δ_j should be positive. This is an internal consistency check for the emulator, not proof of a natural causal effect.

LLM simulation evaluation. The second evaluation asks whether LLM-simulated reason states can replace human-stated reasons. We evaluate $\widehat{Z}$ directly against human Z and indirectly by passing $\widehat{Z}$ through the human reason model. The no-leak reason generator receives only D, K, X and the codebook. It does not see human open text, purchase intent, believability, differentiation, rank, final choice, or gold reason labels.

Controls for $\widehat{Z}$. We use two negative controls. In the no-reason control, all reason states are set to zero. In the mismatched-reason control, human Z vectors are shuffled across rows; this preserves the overall frequency of reasons but breaks the respondent-concept match. A useful LLM-simulated $\widehat{Z}$ should improve over both controls.

Native PMF versus $\widehat{Z}$ readout. For survey simulations there are two different predictions. The native PMF is the simulator's own purchase-intent distribution, obtained from its generated text through SSR. The $\widehat{Z}$ readout first extracts signed reason states from the generated text and then predicts purchase intent with the human reason model $f_\theta(D, K, X, \widehat{Z})$. The native PMF tests outcome fit; the $\widehat{Z}$ readout tests whether the LLM-simulated reasons work in the human reason model.

Metrics. For purchase intent, lower MAE, $\frac{1}{n}\sum_r |Y_r - \widehat{Y}_r|$, and lower RMSE are better. Top-two-box intent is $T_r = \mathbb{1}[Y_r \geq 4]$; Top2 AUC is the ROC-AUC of $\widehat{Y}$ for predicting T. Final-pick AUC is the ROC-AUC for whether a concept was the respondent's final choice. For reasons, we use active-set overlap, $\text{IoU} = |A_{\text{human}} \cap A_{\text{model}}| / |A_{\text{human}} \cup A_{\text{model}}|$.

5 Results

5.1 Human Reason Model Evaluation

Human rationale-derived Z_{core} substantially improves held-out prediction (Table 2). The baseline model using only D, K, X has MAE 0.863. Adding Z_{core} lowers MAE to 0.625. Adding Z_{core} on top of believability and differentiation also helps, lowering MAE from 0.717 to 0.564. Respondent-level bootstrap intervals support the same lift; Appendix C reports the intervals.

The signed reason states also behave coherently inside the human reason model (Figure 2). When a core reason is set from blocker ($Z = -1$) to supporter ($Z = +1$), the predicted purchase intent increases. This does not identify a natural mediator effect, but it checks that the signed codebook and learned readout are aligned.

Reason states do not determine behavior by themselves. Two rows can share the same active reasons and still differ because the concept and respondent context differ. Appendix C shows

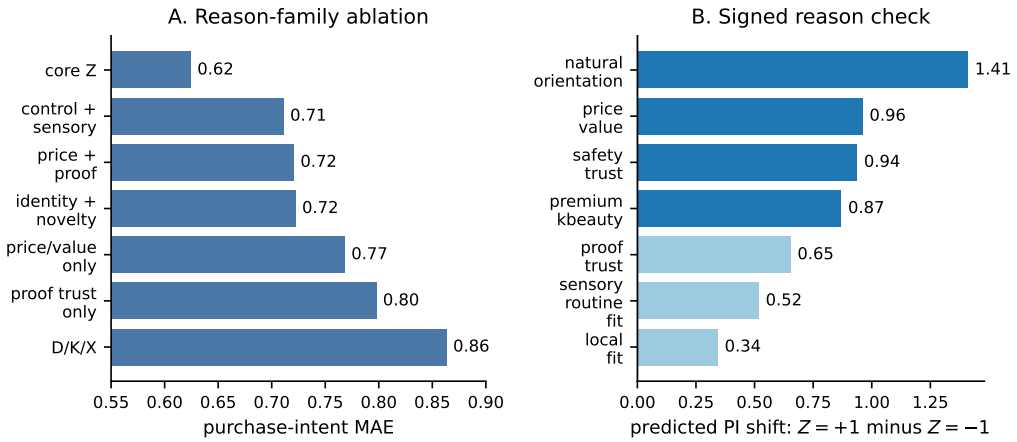


Figure 2: Reason-state checks for the human reason model. Left: ablation compares readouts that use different subsets of reason families. Right: each core reason node is set from blocker ($Z = -1$) to supporter ($Z = +1$); positive values mean the predicted purchase intent moves in the codebook-consistent direction.

that the same reason family can have different strength across concepts. This is important: Z is not merely a disguised label for Y.

5.2 LLM Simulation Evaluation

The main negative result is that LLM-simulated reason states do not yet substitute for human rationale-derived reasons. Figure 3 separates outcome simulation from reason-state simulation. The human-reasons bar uses the observed rationale-derived Z. The no-reason control sets all Z values to zero. The mismatched-reasons control shuffles human Z across respondent-concept rows. Survey-simulated and voice-pool reasons are $\widehat{Z}$ states extracted from LLM survey text and decoded through the human reason model. The voice-pool native PMF is different: it is the voice-pool simulator's direct purchase-intent prediction through SSR, without passing through $\widehat{Z}$.

Human reasons perform best. LLM-simulated reasons often hurt when decoded through the same human reason model, and the direct no-leak reason generator is worse than both no-reason and mismatched-reason controls on the hard sample. The common failure mode is concept echoing. The LLM reads the concept board and reproduces the intended selling points, even when a respondent rejects those same points. For example, one human rationale expressed skepticism that plant-based ingredients could provide adequate protection. The human code marks proof trust and natural orientation as blockers. An LLM-simulated reason may instead mark natural orientation and local fit as supports because the concept presents those cues positively.

The voice-pool results show a capacity gap. Oracle voice selection is much better than uniform aggregation, suggesting that useful modes exist in the pool. However, learned selectors using only deployable features remain close to uniform aggregation. The bottleneck is therefore not only generation; it is also selecting the appropriate response mode for a given respondent-concept cell.

6 Discussion and Limitations

Outcome matching is an incomplete validation target for social simulation. A simulator can match a winner or an aggregate distribution while missing the reason path that would make the result actionable. The sunscreen concept test shows that human rationale-derived reasons improve held-out prediction, move predictions in the expected direction under signed

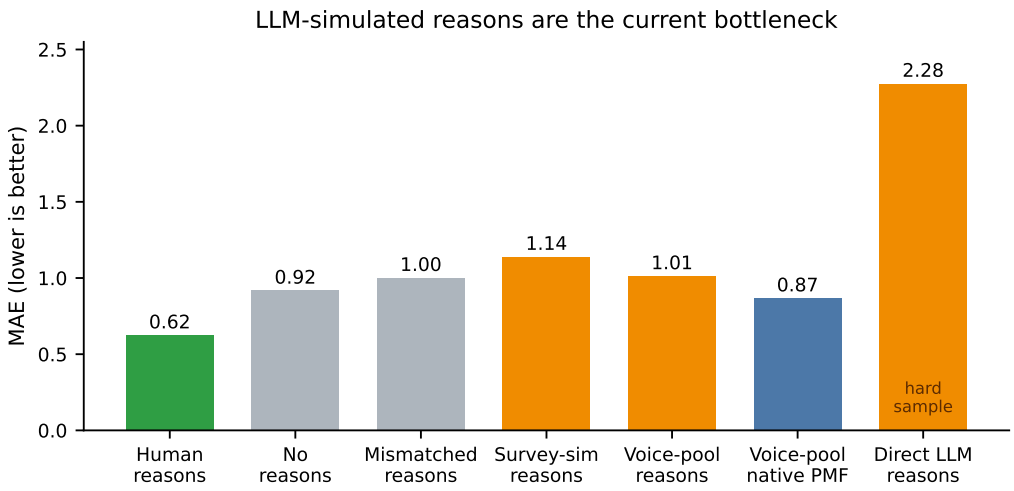


Figure 3: LLM reason simulation bottleneck. Green: human reason states passed through the human reason model. Gray: no-reason and mismatched-reason controls. Orange: LLM-simulated $\widehat{Z}$ decoded through the same human reason model. Blue: the voice-pool simulator's native PMF, which predicts Y directly and does not test the $\widehat{Z} \to Y$ path. The direct LLM reason bar is evaluated on 30 hard rows.

perturbations, and expose concept-level heterogeneity. LLM-simulated reasons do not yet pass the same test: they often echo the concept board rather than recover the respondent's acceptance or rejection path.

The causal claim is intentionally limited. We do not estimate natural direct or indirect effects, and the study is observational with respect to stated reasons. The contribution is an evaluation object: a fixed reason codebook and a fixed human reason model create a manipulable interface for prediction, perturbation, and simulator substitution. The study is small and covers one category. Future work should repeat the protocol in other categories, pre-register the codebook, and compare stronger LLM reason generators, retrieval, and learned voice selectors.

7 Conclusion

Matching human answer distributions is not enough for social simulation. A simulator can predict the winner and still miss the mechanism. We propose a reason-mediated behavioral model: D, K, X shape Z, and Z helps predict Y. In one sunscreen concept test, human rationale-derived reasons improve prediction. Current LLM-simulated reasons do not recover the same reason path reliably. Reasons are therefore not merely explanatory text; they are part of what a simulator should be evaluated against.

References

Icek Ajzen. The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2):179–211, 1991.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. Position: Llm social simulations are a promising research method. In ICML 2025 Position Paper Track, 2025.

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. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 819–862, 2023.

Christopher Barrie and Petter Törnberg. Emergent llm behaviors are observationally equivalent to data leakage. arXiv preprint arXiv:2505.23796, 2025.

Martin Fishbein and Icek Ajzen. Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. Addison-Wesley, 1975.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. Psychological Methods, 15(4):309–334, 2010.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. Transactions on Machine Learning Research, 2024.

Benjamin F. Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C. Luhmann, Robbie Dow, Kli Pappas, and Thomas V. Wiecki. Llms reproduce human purchase intent via semantic similarity elicitation of likert ratings. arXiv preprint arXiv:2510.08338, 2025. URL https://arxiv.org/abs/2510.08338.

NVIDIA. Nemotron-personas-india. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/nvidia/Nemotron-Personas-India.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 2023.

Judea Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2 edition, 2009.

Shalom H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. Advances in Experimental Social Psychology, 25:1–65, 1992.

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. arXiv preprint arXiv:2011.04216, 2020.

Amit Sharma, Jake M. Hofman, and Duncan J. Watts. Estimating the causal impact of recommendation systems from observational data. arXiv preprint arXiv:1510.05569, 2015.

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.

A Reproducibility and Ethics

Reproducibility. All results were computed from the local sunscreen concept-test workspace: human survey rows, fixed reason-code files, processed feature matrices, LLM-simulated reason checks, and figure scripts. This draft is a new LaTeX/PDF copy, not an overwrite of the previous paper.

Ethics. Synthetic respondents can mislead users if they are treated as replacements for human evidence. This paper argues for a stricter test. The study is small, and the results should not be used as population claims without more validation.

B Concept-test Questionnaire and Stimuli

Target and sample. The concept test targeted Indian women aged 22–40 who had purchased sunscreen in the past six months. The fielding brief specified English-speaking, urban or semi-urban respondents with an NCCS A/B skew. After exclusions, the analytic file contains 94 usable respondents. Six responses were excluded for attention, rurality, or straight-line quality rules.

Quantity	Observed distribution
Respondents	94 usable; 6 excluded before analysis
Age	range 22–40; mean 28.7; median 28
City tier	Tier 1: 51; Tier 2: 19; Tier 3: 24
Concept order	A,B,C: 23; A,C,B: 11; B,A,C: 20; B,C,A: 12; C,A,B: 19; C,B,A: 9
Sunscreen-use frequency code	1: 58; 2: 21; 3: 5; 4: 6; 5: 4
White-cast concern	ratings 4–5: 80 / 94
SPF-claim trust	ratings 1–5: 2, 14, 46, 22, 10
Final pick code	A: 32; B: 39; C: 19; none: 4

Table 3: Sample and design distribution. Code meanings are preserved from the survey export when the raw label map is not present in the analysis file.

Questionnaire fields. The file contains age, city tier, education, chief-earner education, household durables, income, languages, BFI-10 personality items, sunscreen-use frequency, sunscreen brands owned, white-cast concern, SPF-claim trust, purchase intent, believability, differentiation, open-ended rationales, rankings, and final pick.

Symbol	Survey source	Examples
D_i	Respondent descriptors	age, city tier, education, income proxy, languages, BFI-10 personality items
K_i	Sunscreen-category context	use frequency, brands owned, white-cast concern, SPF-claim trust
X_c	Concept board	price, origin, authority cue, claims, proof points, sensory promise
M_{ic}	Measured intermediate ratings	believability and differentiation for concept c
Z_{ic}	Open-ended rationale after concept c	signed reason states extracted from the respondent's text
Y_{ic}	Behavioral response	purchase intent; ranking and final pick as secondary outcomes

Table 4: How the questionnaire maps to the variables in the reason-mediated behavioral model.

Scale anchors. Purchase intent used a 1–5 scale: 1 definitely would not buy; 2 probably would not buy; 3 might or might not buy; 4 probably would buy; 5 definitely would buy.



Figure 4: Example concept board shown to respondents. The full study used three boards: SunVeda, Dr. Anita, and Hae.

Believability used 1 not at all believable to 5 extremely believable. Differentiation used 1 not at all different to 5 extremely different.

Concept	Price / positioning	Stimulus summary
A: SunVeda	Rs. 499 / 50 ml; Ayurvedic-botanical	"Daily sun protection, rooted in Ayurveda." Daily sunscreen for Indian skin and Indian sun; mineral and botanical UV filters with turmeric, kumkumadi, and manjishtha. Benefits included SPF 50 PA++++, tinted no-white-cast formula, lightweight non-greasy daily use, and avoidance of oxybenzone, octinoxate, and parabens. Proof included Kerala formulation, Indian-skin testing, AYUSH certification, and reef-friendly filters.
B: Dr. Anita	Rs. 699 / 50 ml; dermatologist-clinical Indian	"Dr. Anita Invisible SPF 50. Formulated for Indian skin. Zero white cast." Dermatologist-formulated sunscreen for Indian skin tones, heat, humidity, and pollution. Benefits included SPF 50 PA++++, invisible finish, sweat/humidity resistance, matte finish, non-comedogenic use, and face/neck daily use. Proof included Indian dermatologist formulation, a 12-week test on 142 Indian women, 91% no visible white cast, and BIS-tested SPF/PA ratings.
C: Hae	Rs. 1,299 / 50 ml; Korean prestige hybrid	"Hae. K-beauty sun fluid for the Indian sun." Korean-formulated, India-available lightweight sun fluid with next-generation Korean UV filters, barely-there feel, and subtle tone-up tint. Benefits included SPF 50+ PA++++, watery texture, primer use, vegan, fragrance-free, and alcohol-free claims. Proof included Seoul formulation/manufacture, Indian distribution, non-US-approved filter caveat, Seoul clinical testing, and Mumbai compatibility check.

Table 5: Concept stimuli used in the three-arm sunscreen concept test.

C Experiment Details

Rows and artifacts. The unit is a respondent-concept row. The concept test contains 94 respondents and three concepts, giving 282 rows. The artifact set includes the human survey matrix, fixed reason-code files, processed feature matrices, synthetic response outputs, voice-pool outputs, a hard no-leak staged-generator sample, and LLM-readout checks.

Human-reason-model experiments. We train purchase-intent readouts using five-fold GroupKFold by respondent. The headline model is ridge regression; final-pick checks use balanced logistic regression. Robustness checks include respondent bootstrap intervals, concept-held-out splits, zero-Z controls, shuffled-Z controls, and reason-family ablations. We also force each core reason node to $-1, 0, +1$ and measure readout movement.

Reason-surface experiments. We group rows by signed reason signatures and inspect the mapping from reason states to purchase intent across concepts. This asks whether the same reason family has stable direction while still allowing concept-level heterogeneity. It also checks that Z is not merely a deterministic rewrite of Y.

Reason-surface plot. Figure 5 reports the observed reason-to-behavior contrasts by concept.

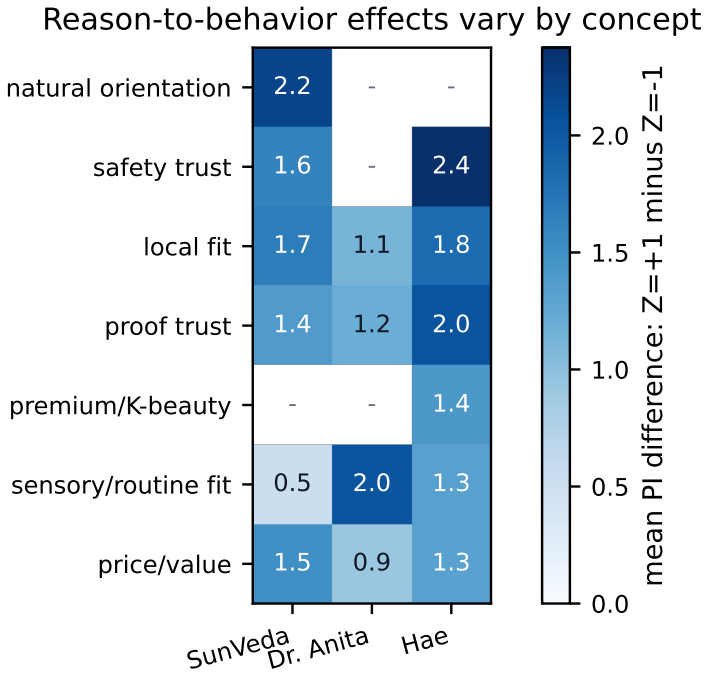


Figure 5: Observed reason-to-behavior associations by concept. Each cell is the mean purchase-intent difference between rows where a reason supports purchase ($Z = +1$) and rows where it blocks purchase ($Z = -1$). Blank cells mean that this small study did not contain both sides of the contrast.

LLM-simulated reason experiments. We replace human Z with $\widehat{Z}$ from the context-conditioned response simulation, the voice-pool simulation, and a staged no-leak LLM generator. Each $\widehat{Z}$ is decoded through the same human reason model. We report reason IoU, signed-node accuracy, active precision/recall, and human-reason-model MAE.

Voice-selection experiments. For the voice-pool simulation, we compare the current average over skeptic/balanced/enthusiast samples with row-level oracle selection, sample-level oracle

selection, and learned hard/soft selectors. The oracle results measure available capacity; learned selectors measure deployability.

Transition and warm-start experiments. We also train exploratory transition models from observed respondent/category state to reason state, $D, K, X \to Z$, and warm-start models that use the same respondent's other-concept reasons to predict the held-out concept. These are exploratory because rare reason nodes are hard to recover in a 94-person study.

Bootstrap intervals. For the clean reason readout, $D, K, X + Z_{\text{core}}$ improves over D, K, X by 0.239 MAE with 95% CI $[0.172, 0.299]$. Adding Z_{core} on top of M improves over $D, K, X + M$ by 0.154 MAE with 95% CI $[0.099, 0.209]$.

Hard-sample LLM reason metrics. The no-leak staged generator was evaluated on 30 difficult rows. Mean reason IoU was 0.203; exact active-set match was 0.000; signed-node accuracy was 0.448; active recall was 0.625; active precision was 0.210. Human-reason-model MAE was 1.278 for human Z, 2.276 for staged LLM $\widehat{Z}$, 1.760 for zero-Z, and 1.853 for shuffled-Z.

D Models and Readouts

Human reason model. The human reason model is the readout $f_\theta(D, K, X, Z) \to Y$. The primary model is ridge regression with $\alpha = 1.0$, median imputation for numeric variables, most-frequent imputation for categoricals, standard scaling for numeric variables, and one-hot encoding for categorical concept features. Predictions are clipped to 1–5.

Other readouts. Gradient boosting and random forest readouts were used as nonlinear robustness checks. They test whether the reason-state lift depends on using a linear readout.

Prompted LLM readout. The prompt-only LLM readout used gpt-5.4-mini on a stratified sample. One arm used reason states only; another used D, K, X, M, Z plus a verbalized structured-model prior. These readouts did not beat the structured readout.

LLM reason generator. The staged no-leak $\widehat{Z}$ generator used the OpenAI verifier/generator pipeline. The reported hard-sample run uses a GPT-5-family structured generator; the same family is used for larger-model verification in the human Z extraction pipeline. This choice is deliberate: $D, K, X \to \widehat{Z}$ generation is a smaller, structured JSON task, while the Gemma runs are high-volume respondent-style survey simulations. The generator saw only D, K, X and the reason codebook. It did not see open text, outcomes, or gold reason labels.

LLM survey-simulation models. Synthetic responses use Gemma 4 31B IT for response generation and Qwen3-Embedding-0.6B for semantic-similarity rating embeddings. These Gemma outputs are used in two ways: as native outcome predictions through SSR, and as generated survey text from which $\widehat{Z}$ can be extracted for the $\widehat{Z} \to Y$ test. Direct-Likert arms produce a single integer. SSR arms produce a short free-text answer, which is converted into a 1–5 purchase-intent probability mass function using six locked reference-statement sets. Appendix E gives the full arm table.

Persona matching. Nemotron-Personas-India rows are filtered to female, urban personas aged 22–40. Each human respondent is matched to one persona by seeded stratified matching on city tier and age band. The primary match bucket is exact city tier and exact age band; fallback buckets use same city tier or same age band. Assigned personas are removed from the pool, so the matching is without replacement.

Context-SSR details. The Context-SSR arm combines the matched Nemotron persona with a stratum-level sunscreen context block. The context block is generated from non-outcome survey fields: sunscreen-use frequency, brands owned, white-cast concern, SPF-claim trust,

languages, and personality summaries. It excludes purchase intent, believability, differentiation, rankings, final choice, and open-ended product rationales.

E Simulation Arm Taxonomy

Arm taxonomy. The simulation run contains eight arms. They form a controlled ladder from demographic-only prompting to persona, psychographic, category-context, and own-answer conditioning. The reason-path claims use the LLM reason simulation protocols described in Section 4; the full arm list is included for reproducibility.

Arm	Name used here	Conditioning
A1	Demo-DLR	demographics only; direct 1–5 rating
A2	Demo-SSR	demographics only; free text scored by SSR
N0	Persona-DLR	Nemotron persona; direct 1–5 rating
N1	Persona-SSR	Nemotron persona; free text scored by SSR
N2a	OCEAN-Numerical	persona plus numeric Big Five scores
N2b	OCEAN-Narrative	persona plus prose psychographic narrative
N3	Context-SSR	persona plus stratum-level sunscreen context
N4	Half-Q	persona plus own non-target survey answers

Table 6: Simulation arms. DLR is direct Likert rating. SSR is semantic similarity rating: the model writes text, and an embedding-based rater maps the text to a 1–5 probability mass function.

Generation scale. The bulk run covers 94 respondents, three concepts, eight arms, and three samples per arm-cell, giving $94 \times 3 \times 8 \times 3 = 6{,}768$ generated responses. Response generation uses Gemma 4 31B IT through a local OpenAI-compatible endpoint. SSR embeddings use Qwen3-Embedding-0.6B.

F Semantic Similarity Rating

Text-to-PMF scoring. For SSR arms, the model output is not parsed as a number. It is embedded and compared with six locked sets of reference statements covering the 1–5 purchase-intent scale. The final probability mass function is the average across reference sets:

$$\hat{p}(Y = k \mid r) = \frac{1}{6} \sum_{s=1}^{6} \hat{p}_s(Y = k \mid \mathsf{embed}(r)), \quad k \in \{1, \ldots, 5\}.$$

Mean purchase intent is then $\sum_{k=1}^{5} k\hat{p}(Y = k)$, and top-two-box mass is $\hat{p}(Y = 4) + \hat{p}(Y = 5)$. In the reason-path experiments, the native PMF is used as an outcome prediction; the generated text is separately mapped into $\widehat{Z}$ for the human-reason-model test.

G Reason Codebook

Human reason extraction prompt excerpt.

> You extract respondent-stated reason states for a sunscreen concept test. Return only JSON. Do not infer hidden psychology from the rating, rank, or final pick; those are not provided. Use the concept context only to disambiguate words like price, Korean, dermatologist, SPF, or Indian skin. Classify only reasons that are supported by the respondent's own open-ended text. Do not mark a node just because the concept has that attribute.

No-leak reason generation prompt excerpt.

> Task: predict reason states Z_hat for one respondent-concept cell using only respondent state D, category prior K, and concept treatment X. You do not see

Core node	General family	Positive / negative interpretation
cz_price_value	Value/control	worth the price / too expensive or poor value
cz_proof_trust	Proof trust	evidence is believed / claims are doubted
cz_natural_orientation	Natural/cultural fit	natural or traditional cue helps / cue reduces trust
cz_premium_kbeauty	Novelty/status	premium or K-beauty appeal / feels excessive or irrelevant
cz_local_fit	Local fit	Indian skin/climate fit helps / local fit is missing
cz_sensory_routine_fit	Routine fit	texture/no-white-cast helps daily use / routine fit blocks adoption
cz_safety_trust	Safety trust	safety is reassured / safety or regulatory anxiety blocks adoption

Table 7: Core reason codebook used for Z_{core}.

human open text, gold reason labels, ratings, ranking, or final choice. Use a two-stage reasoning process internally: candidate reason discovery, then calibrated compression into a sparse signed Z_hat vector. Most rows should have 0–2 active reason families; use 3 only for a clear mixed tradeoff. Do not activate a generic concept attribute unless D/K suggests the respondent will use it as a decision reason.

H Voice-pool Aggregation

Algorithm. The voice-pool estimator keeps several possible response modes alive before averaging. In this version the three modes are skeptic, balanced, and enthusiast. Each voice produces a candidate reason state and a purchase-intent distribution. The selector weights them:

$$\hat{p}(Y \mid D, K, X) = \sum_{v \in \mathcal{V}} \pi_v(D, K, X)\, \hat{p}_v(Y \mid D, K, X).$$

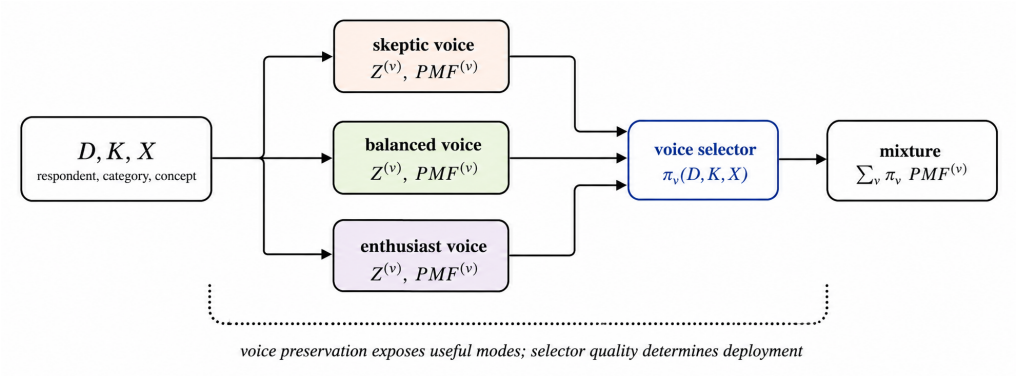


Figure 6: Voice-pool aggregation keeps skeptic, balanced, and enthusiast modes separate before mixing their predictions.

Selector tests. For each respondent-concept cell, uniform aggregation was compared with a row-level oracle voice, a sample-level oracle, and learned hard/soft selectors. The large oracle gap means the pool contains useful modes. The weak learned selector means selecting the right mode is still unsolved.

I Additional Result Tables

Readout	N	MAE	RMSE	Top2 AUC
human Z, human reason model	282	0.625	0.837	0.882
zero-Z control	282	0.918	1.103	0.566
shuffled-Z control	282	1.001	1.232	0.518
survey-simulated $\widehat{Z}$, human reason model	282	1.140	1.437	0.511
voice-pool $\widehat{Z}$, human reason model	282	1.013	1.285	0.618
voice-pool native PMF	282	<u>0.867</u>	<u>1.035</u>	<u>0.652</u>

Table 8: LLM-simulated reason and native-PMF comparisons. Best is bold; second-best is underlined.

Source	Current MAE	Oracle voice MAE
generic	0.867	<u>0.485</u>
context	<u>0.930</u>	0.517
demographic-context	0.935	0.476
persona	0.932	0.495

Table 9: Voice-pool capacity check. The oracle column is not deployable; it estimates possible gain if voice selection were solved.